Nonconvex Recovery of Low-complexity Models

Qing Qu

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

COLUMBIA UNIVERSITY

2018

ABSTRACT

Nonconvex Recovery of Low-complexity Models

Qing Qu

Today we are living in the era of big data, there is a pressing need for efficient, scalable and robust optimization methods to analyze the data we create and collect. Although Convex methods offer tractable solutions with global optimality, heuristic nonconvex methods are often more attractive in practice due to their superior efficiency and scalability. Moreover, for better representations of the data, the mathematical model we are building today are much more complicated, which often results in highly nonlinear and nonconvex optimizations problems. Both of these challenges require us to go beyond convex optimization. While nonconvex optimization is extraordinarily successful in practice, unlike convex optimization, guaranteeing the correctness of nonconvex methods is notoriously difficult. In theory, even finding a local minimum of a general nonconvex function is NP-hard – nevermind the global minimum.

This thesis aims to bridge the gap between practice and theory of nonconvex optimization, by developing global optimality guarantees for nonconvex problems arising in real-world engineering applications, and provable, efficient nonconvex optimization algorithms. First, this thesis reveals that for certain nonconvex problems we can construct a model specialized initialization that is close to the optimal solution, so that simple and efficient methods provably converge to the global solution with linear rate. These problem include sparse basis learning and convolutional phase retrieval. In addition, the work has led to the discovery of a broader class of nonconvex problems – the so-called *ridable saddle* functions. Those problems possess characteristic structures, in which (i) all local minima are global, (ii) the energy landscape does not have any "flat" saddle points. More interestingly, when data are large and random, this thesis reveals that many problems in the real world are indeed *ridable saddle*, those problems include complete dictionary learning and generalized phase retrieval. For each of the aforementioned problems, the benign geometric structure allows us to obtain global recovery guarantees by using efficient optimization methods with arbitrary initialization.

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
$\|\cdot\|$	vector ℓ^2 norm or matrix operator norm
$\|\cdot\|_F$	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 $\{1, \dots, k\}$
$\mathrm{col}(\cdot)$, $\mathrm{row}(\cdot)$	the column and row spaces of a matrix
$\boldsymbol{P}_{\boldsymbol{v}}$, $\boldsymbol{P}_{\boldsymbol{v}^\perp}$	$\boldsymbol{P}_{\boldsymbol{v}} = \boldsymbol{v}\boldsymbol{v}^* / \|\boldsymbol{v}\|^2$, $\boldsymbol{P}_{\boldsymbol{v}} = \boldsymbol{I} - \boldsymbol{v}\boldsymbol{v}^* / \|\boldsymbol{v}\|^2$, the projections onto the span of the vector $\boldsymbol{v}$ and its orthogonal complement.
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 $\mathcal{L}$
$\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$	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 $W \sim \mathcal{N}(0,1)$
$\circledast$	circulant convolution
$\boldsymbol{C_a}$	the circulant matrix generated from a vector $\boldsymbol{a}$
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

First and foremost I want to sincerely thank my advisor Prof. John Wright. It is my honor to be his Ph.D student for a five year journey. His working ethics, long-term vision, and enthusiasm have reshaped my mindset beyond research. His advice on both research as well as on my career have been invaluable. His wealth of ideas, clarity of presentation, and inspiration help me to walk through the hard-time in this journey, making it stimulating and productive. I would like to thank my undergraduate advisor Prof. Yuantao Gu at Tsinghua University, who provide invaluable guidance that inspired me to become a researcher. I would also like to thank my master advisor Prof. Trac Tran at Johns Hopkins University, who introduced me into the field of compressed sensing and sparse representation, and provide unconditional support and freedom for pursuing my academic career.

Second, I am also grateful to have Prof. Xiaodong Wang, Prof. John Pasliey, Prof. Arian Maleki, Prof. Rene Vidal serving on my thesis committee, and offering numerous and invaluable feedbacks. I would especially like to thank Prof. Rene Vidal during my visit to Johns Hopkins University, and strong support for my job search. I would also like to thank Prof. Yuxin Chen (Princeton), Prof. Xin Chen (UIUC), Prof. Deliang Wang (OSU), Prof. Carlos Fernandez-Granda (NYU) during my visits.

The Microsoft Corporation selected me to be among its twelve fellows and paid the tuition for my last two years. I would like to express my sincere gratitude for the generous support. I would also like to thank Lin Xiao and Denny Zhou during my summer intern at Microsoft Research in 2016.

I also would like to thank my fellow group members and collaborators, Ju Sun, Yuqian Zhang, Henry Kuo, Yenson Lau, Cun Mu, Robert Colgan, Dar Gilboa, Sam Buchanan, Simon Zhai, and Tim Wang. In particular, I would like express my sincere gratitude towards my academic brother Ju Sun, his guidance is invaluable during my junior years and help me grow fast. I would also like to thank all my friends at Columbia: Shang Li, Xu Zhang, Le Zheng, Shan Zhong, Zhe Wang, Gaojianyong Wang, Yihong Li, Jiangfan Zhang, and Cong Han. It is such a memorable journey to walk through with these wonderful people.

Finally, last but by no means least, I want to express my heartfelt gratitude to my parents, and my wife. I would like to thank them for their unconditional support to pursue my academic dream and spiritual freedom.

Thanks for all your encouragement!

Qing Qu

Sept. 5, 2018

New York

To my family

Part I

Overview

Today we are living in an era of information explosion. As the sensors and sensing modalities proliferate, our world is inundated by unprecedented quantities of data, in the form of images and videos, gene expression data, web links, product rankings and more. For instance, cameras, hyperspectral sensors, etc., produce observations with millions, or even billions of dimensions (see Fig. 1 for an illustration). There is a pressing need for efficient, scalable and robust optimization methods to analyze the data we create and collect. Classical convex methods offer tractable solutions with global optimality (see Fig. 2). In contrast, for many applications heuristic nonconvex methods are more attractive due to their superior efficiency and scalability. Moreover, for better representations of the data, we are building increasingly complicated mathematical models, which often naturally results in highly nonlinear and nonconvex optimizations problems. Both of these challenges require us to go beyond convex optimization.

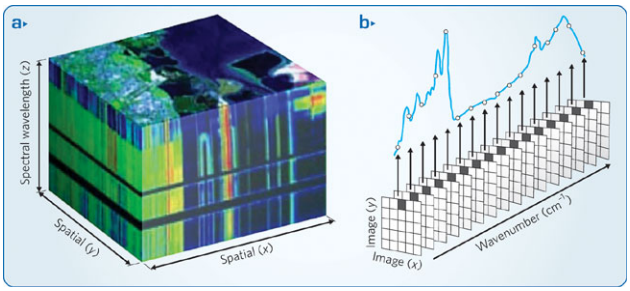


Figure 1: An illustration of high-dimensional data: hyperspectral data cube.

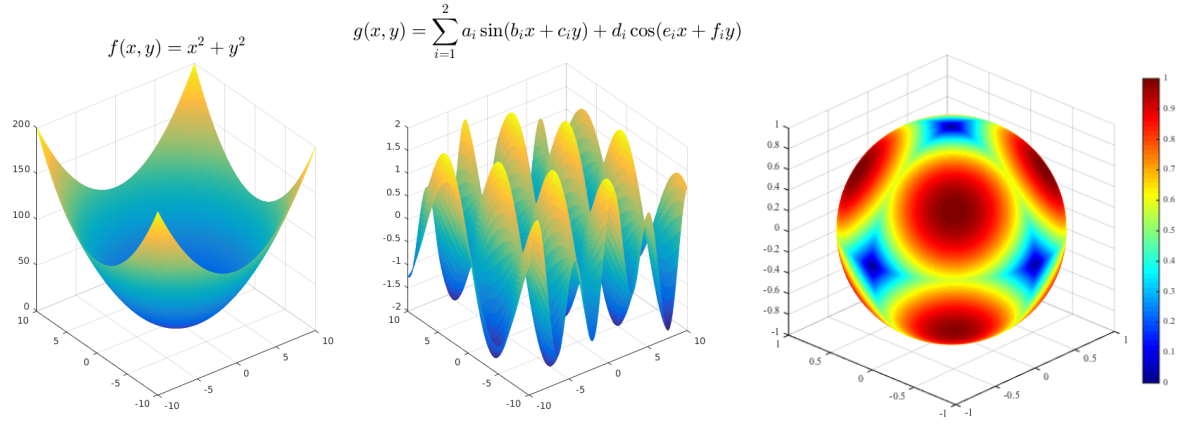


Figure 2: An illustration of function landscapes of convex problem (left), general nonconvex problems (middle), and "nice" nonconvex functions (right).

Recent advances in nonconvex optimization show that it has the potential to surmount both of these new challenges. Phase retrieval [CLS15b, CC15, SQW16], sparse coding [AAN13, SQW15a, AGMM15], deep neural networks [Kaw16], matrix and tensor factorization [TBSR15, ZL15], and dynamical systems [HMR16] are

some representative examples, where non-convex methods provide us dramatically more flexible, scalable, and effective computational tools. While nonconvex optimization is extraordinarily successful in practice, unlike convex optimization, guaranteeing the correctness of nonconvex methods is notoriously difficult. In theory, even finding a local minimum of a general nonconvex function is NP-hard [MK87] – nevermind the global minimum, which is the true object of our interest.

0.1 Contribution of the Thesis

The thesis bridges the gap between practice and theory of nonconvex optimization, by developing global optimality guarantees for several nonconvex problems arising in real-world engineering applications, and provable, efficient nonconvex optimization algorithms.

Finding a sparse vector in a subspace One of the fundamental nonconvex problems in signal processing and machine learning is the dictionary learning problem. Effective solutions to this problem play a crucial role in Effective solutions to this problem play a crucial role in many applications spanning low-level image processing to high-level visual recognition [MBP14]. The goal of complete dictionary learning, is to seek a compact representation

$$\underbrace{\boldsymbol{Y}}_{\text{Data}} = \underbrace{\boldsymbol{Q}}_{\text{Dictionary}} \quad \underbrace{\boldsymbol{X}}_{\text{Sparse Coefficients}}$$

of input data $\boldsymbol{Y} \in \mathbb{R}^{n\times p}$, where $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ is a complete dictionary (square and invertible), and $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ should be as sparse as possible. Because $\boldsymbol{Q}$ is complete, the row space of $\boldsymbol{Y}$ equals to the row space of $\boldsymbol{X}$, i.e., $\mathrm{row}(\boldsymbol{Y}) = \mathrm{row}(\boldsymbol{X})$: the rows of $\boldsymbol{X}$ are the sparsest vectors in the subspace $\mathcal{S} = \mathrm{row}(\boldsymbol{Y})$ [SWW12a]. Therefore, solving complete dictionary learning problem is equivalent to *finding the sparsest non-zero vector in a given subspace* $\mathcal{S}$. Mathematically, can we *globally* solve a nonconvex problem

$$\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_1, \quad \boldsymbol{x} \in \mathcal{S}, \quad \|\boldsymbol{x}\| = 1 \quad ? \tag{0.1.1}$$

Beyond dictionary learning, variants of the problem (0.1.1) have appeared in the context of applications to numerical linear algebra [CP86], graphical model learning [ZF13], nonrigid structure from motion [DLH12], spectral estimation and Prony's problem [BM05], sparse PCA [ZHT06], and blind source separation [ZP01].

For a simple and idealized *planted sparse subspace* $\mathcal{S}$, where there is only one sparse vector planted in an otherwise random subspace, the work in Part II shows that there exist efficient nonconvex methods that

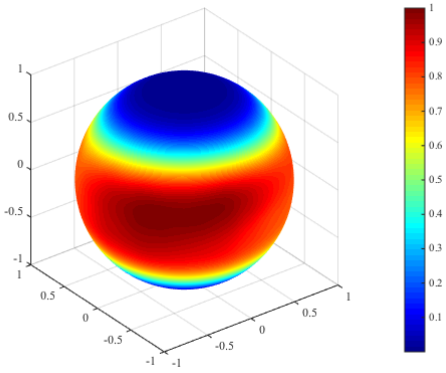


Figure 3: Function landscape of planted sparse vector model

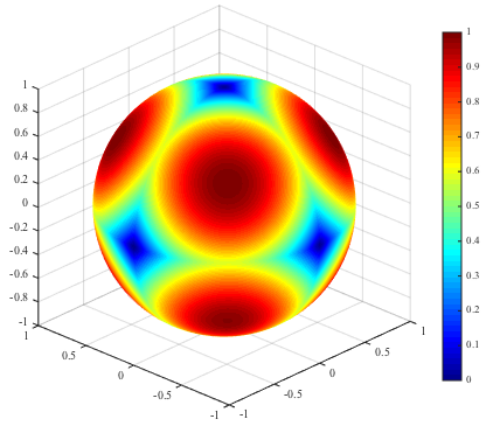


Figure 4: Function landscape of dictionary learning model

provably find the global solution with special initialization. For the dictionary learning setting, where the basis of $\mathcal{S} = \mathrm{row}(\boldsymbol{Y})$ are all sparse vectors, Part III shows that the problem (0.1.1) has no spurious local minima and all local minima correspond to the sparse basis.

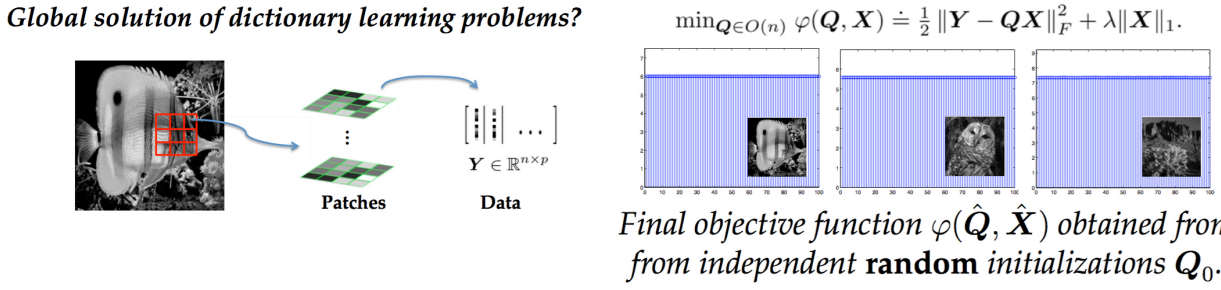


Figure 5: Evidence of global optima on real data problems: the final objective value does not depend on initialization.

Moreover, our results have provided new insights into the optimization landscape of the objective (0.1.1) (see Fig. 3 and Fig. 4): for the idealized *planted sparse* model, the planted sparse vector is a local minimum around which the function is local strongly convex (Fig. 3); for complete dictionary learning, the function landscape is highly symmetric and the sphere can be partitioned into *strong convexity*, *large gradient*, and *negative curvature* regions (see Fig. 4 and Fig. 6), and all *saddle points* can be escaped by using negative curvature information. The geometric analysis provides us new insights to design more efficient nonconvex optimization methods, and provides better explanations of algorithmic performance on real data. Fig. 5 shows evidence of global recovery in dictionary learning from natural image patches: over many random initializations, the algorithm yields the same final objective value, and the same dictionary, up to a scaled permutation of the columns.

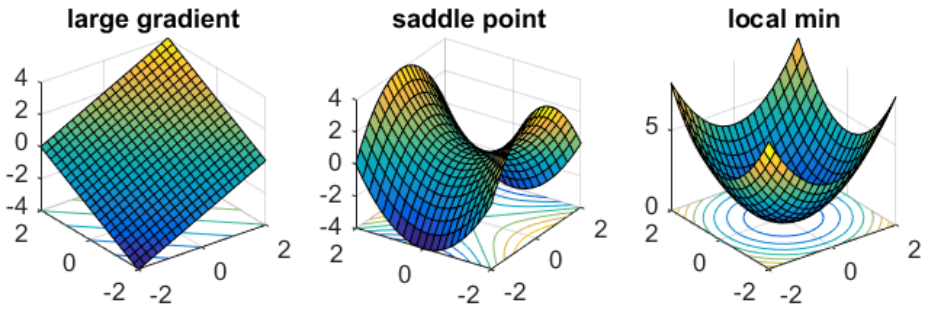


Figure 6: Regions of ridable saddle functions.

Ridable saddle functions The geometric result on complete dictionary learning in Part III leads to the discovery of a broader class of nonconvex functions, which we call *ridable saddle* functions[1] [SQW15d]. More concretely, a function $f(\boldsymbol{x}) : \mathcal{M} \mapsto \mathbb{R}$ is ridable saddle if all points $\boldsymbol{x}$ satisfy at least one of the following: (a) the gradient $\operatorname{grad} f(\boldsymbol{x})$ is large; (b) the Hessian $\operatorname{Hess} f(\boldsymbol{x})$ has a negative eigenvalue that is bounded away from 0; (c) $\boldsymbol{x}$ is near a local minimum, around which the function is strongly convex (see Fig. 6 for an illustration). In particular, a function $f(\boldsymbol{x}) : \mathcal{M} \mapsto \mathbb{R}$ is $(\alpha, \beta, \gamma, \delta)$-ridable saddle if it satisfies:

> **Definition 0.1 ($(\alpha, \beta, \gamma, \delta)$-ridable saddle function)** *A function $f : \mathcal{M} \mapsto \mathbb{R}$ is $(\alpha, \beta, \gamma, \delta)$ ridable saddle, i.e., any point $\boldsymbol{x} \in \mathcal{M}$ obeys **at least one of the following**: ($T_{\boldsymbol{x}}\mathcal{M}$ is the tangent space of $\mathcal{M}$ at point $\boldsymbol{x}$)*
>
> *1) **[Strong gradient]** $\|\operatorname{grad} f(\boldsymbol{x})\| \geq \beta$;*
>
> *2) **[Negative curvature]** There exists $\boldsymbol{v} \in T_{\boldsymbol{x}}\mathcal{M}$ with $\|\boldsymbol{v}\| = 1$ such that $\langle \operatorname{Hess} f(\boldsymbol{x})[\boldsymbol{v}], \boldsymbol{v}\rangle \leq -\alpha$;*
>
> *3) **[Strong convexity around minimizers]** There exists a local minimizer $\boldsymbol{x}_\star$ such that $\|\boldsymbol{x} - \boldsymbol{x}_\star\| \leq \delta$, and for all $\boldsymbol{y} \in \mathcal{M}$ that is in 2δ neighborhood of $\boldsymbol{x}_\star$, $\langle \operatorname{Hess} f(\boldsymbol{y})[\boldsymbol{v}], \boldsymbol{v}\rangle \geq \gamma$ for any $\boldsymbol{v} \in T_{\boldsymbol{y}}\mathcal{M}$ with $\|\boldsymbol{v}\| = 1$, i.e., the function f is γ-strongly convex in 2δ neighborhood of $\boldsymbol{x}_\star$.*

We remark in passing that requiring a function to be ridable may appear far too restrictive than it actually is. Indeed, one of the central results in Morse theory implies that a generic smooth function is ridable. This new geometric insight not only helps us design more efficient nonconvex optimization algorithms [NP06, GHJY15, LSJR16, JGN+17], but also shed light on solving other nonconvex problems in practice such as the generalized phase retrieval studied in this thesis.

Phase retrieval The phase retrieval problem tries to recover a signal $\boldsymbol{x} \in \mathbb{C}^n$ from nonlinear measurements of the form $\boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}|$, where $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ represents a linear map. Solving phase retrieval problem has

[1] It coincides with recent development in orthogonal tensor decomposition [GHJY15], where they discovered similar properties and call the function strict saddle.

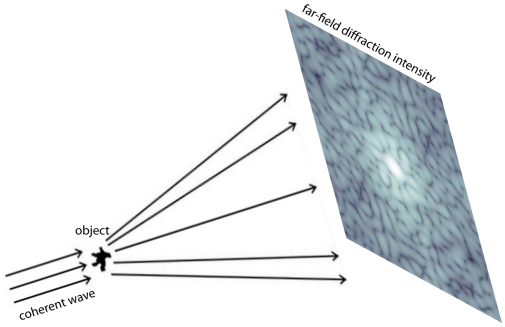


Figure 7: Phase retrieval for coded diffraction imaging, image courtesy of [SEC+15]

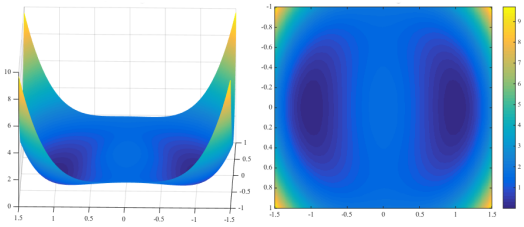


Figure 8: Function landscape of (0.1.2) in $\mathbb{R}^2$: only *global minima* and *saddle points*.

broad applications in X-ray crystallography, microscopy, astronomy, diffraction and array imaging[2], and optics [SEC+15] . To recover $\boldsymbol{x}$, it is natural to consider minimizing the following nonconvex Gaussian log-likelihood function [CL14]

$$\min_{\boldsymbol{z}\in\mathbb{C}^n} f(\boldsymbol{z}) \doteq \frac{1}{2m}\sum_{k=1}^{m}\left(y_k^2 - |\boldsymbol{a}_k^*\boldsymbol{z}|^2\right)^2. \tag{0.1.2}$$

When the sensing vectors $\boldsymbol{a}_k$ are independent and complex Gaussian, Part IV of this thesis reveals that $f(\boldsymbol{z})$ is a ridable-saddle function (see Fig. 8 for a visualization of 2D real case) with $m \sim \mathcal{O}(n\log^3 n)$ samples, which allows efficient, initialization free nonconvex optimization. In contrast, known convex methods require solving a huge semi-definite programming (SDP) problem [CSV13]. Again, our success derives from the benign geometric structure underlying the objective function (0.1.2): under natural data models, the function has a large sample limit, which (i) has no spurious local minima and (ii) can be optimized efficiently.

In real applications, the sensing matrix $\boldsymbol{A}$ is much more structured than the idealized i.i.d. Gaussian model. Motivated by applications such as *channel estimation* and *noncoherent optical communication*, Part V of this thesis studied a convolutional model, $\boldsymbol{y} = |\boldsymbol{a} \circledast \boldsymbol{x}|$. The measurements are generated by passing the signal $\boldsymbol{x}$ through a filter $\boldsymbol{a} \in \mathbb{C}^m$, where $\circledast$ denotes *cyclic convolution*. The convolutional structure also has huge benefits in computation by using fast Fourier transform. However, if we assume $\boldsymbol{a}$ is complex Gaussian, the statistical dependence across entries of $\boldsymbol{y}$ poses great challenges for analysis. By using tools of *decoupling theory* and *suprema of chaos processes of random circulant matrices*, the result in Part V shows that by optimizing a nonconvex objective using a simple gradient descent method, it recovers the true target $\boldsymbol{x}$ with $m \sim \mathcal{O}(n\text{poly}\log n)$ samples.

[2]See Fig. 7 for an example of diffraction imaging.

0.2 Organization

The rest of the thesis is organized as follows. In Part II and Part V, we show that for certain structured nonconvex problems (i.e., finding the sparsest vector in a subspace and convolutional phase retrieval), we can construct an model specialized initialization that is close to the optimal solution, so that simple and efficient methods provably converge to the global solution. In addition, the work in Part III and Part IV have led to the discovery of *ridable saddle* (or strict saddle [GHJY15]) functions – a broader class of nonconvex problems with benign geometric structure, that allows efficient and initialization free global optimization [SQW15d, GHJY15]. In Part III and Part IV, we show that when data are large and random, many problems in the real world are indeed *ridable saddle*, those problems include generalized phase retrieval (Part IV) and complete dictionary learning (Part III). Finally, in Part VI we conclude the thesis and discuss about future directions based on the current results. All the basic technical details are deferred to the appendices.

Part II

Finding a Sparse Vector in a Subspace

Is it possible to find the sparsest vector (direction) in a generic subspace $\mathcal{S} \subseteq \mathbb{R}^p$ with $\dim(\mathcal{S}) = n < p$? This problem can be considered a homogeneous variant of the sparse recovery problem, and finds connections to sparse dictionary learning, sparse PCA, and many other problems in signal processing and machine learning. In this paper, we focus on a *planted sparse model* for the subspace: the target sparse vector is embedded in an otherwise random subspace. Simple convex heuristics for this planted recovery problem provably break down when the fraction of nonzero entries in the target sparse vector substantially exceeds $O(1/\sqrt{n})$. In contrast, we exhibit a relatively simple nonconvex approach based on alternating directions, which provably succeeds even when the fraction of nonzero entries is $\Omega(1)$. To the best of our knowledge, this is the first practical algorithm to achieve linear scaling under the planted sparse model. Empirically, our proposed algorithm also succeeds in more challenging data models, e.g., sparse dictionary learning.

This part is based on our paper [QSW14]. The rest of Part II is organized as follows. In Chapter 2, we provide a nonconvex formulation and show its capability of recovering the sparse vector. Chapter 3 introduces the alternating direction algorithm. In Chapter 4, we present our main results and sketch the proof ideas. Experimental evaluation of our method is provided in Chapter 5. We conclude the paper by drawing connections to related work and discussing potential improvements in Chapter 6. The main proof details are retained to Chapter 7. Other basic auxiliary results are all deferred to Appendix A.

Chapter 1

Introduction

Suppose that a linear subspace $\mathcal{S}$ embedded in $\mathbb{R}^p$ contains a sparse vector $\boldsymbol{x}_0 \neq \boldsymbol{0}$. Given an arbitrary basis of $\mathcal{S}$, can we efficiently recover $\boldsymbol{x}_0$ (up to scaling)? Equivalently, provided a matrix $\boldsymbol{A} \in \mathbb{R}^{(p-n)\times p}$ with $\mathrm{Null}(\boldsymbol{A}) = \mathcal{S}$, [1] can we efficiently find a nonzero sparse vector $\boldsymbol{x}$ such that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$? In the language of sparse recovery, can we solve

$$\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_0 \quad \text{s.t.} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0},\ \boldsymbol{x} \neq \boldsymbol{0} \qquad ? \tag{1.0.1}$$

In contrast to the standard sparse recovery problem ($\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \boldsymbol{b} \neq \boldsymbol{0}$), for which convex relaxations perform nearly optimally for broad classes of designs $\boldsymbol{A}$ [CT05, Don06], the computational properties of problem (1.0.1) are not nearly as well understood. It has been known for several decades that the basic formulation

$$\min_{x} \|x\|_0, \quad \text{s.t.} \quad x \in \mathcal{S} \setminus \{0\}, \tag{1.0.2}$$

is NP-hard for an arbitrary subspace [McC83, CP86]. In this part of the thesis, we assume a specific random *planted sparse model* for the subspace $\mathcal{S}$: a target sparse vector is embedded in an otherwise random subspace. We will show that under the specific random model, problem (1.0.2) is tractable by an efficient algorithm based on nonconvex optimization.

[1] $\mathrm{Null}(\boldsymbol{A}) \doteq \{\boldsymbol{x} \in \mathbb{R}^p \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}\}$ denotes the null space of $\boldsymbol{A}$.

1.1 Motivation

The general version of Problem (1.0.2), in which $\mathcal{S}$ can be an arbitrary subspace, takes several forms in numerical computation and computer science, and underlies several important problems in modern signal processing and machine learning. Below we provide a sample of these applications.

Sparse Null Space and Matrix Sparsification: The *sparse null space* problem is finding the sparsest matrix $\boldsymbol{N}$ whose columns span the null space of a given matrix $\boldsymbol{A}$. The problem arises in the context of solving linear equality problems in constrained optimization [CP86], null space methods for quadratic programming [BHK+85], and solving underdetermined linear equations [GH87]. The *matrix sparsification* problem is of similar flavor, the task is finding the sparsest matrix $\boldsymbol{B}$ which is equivalent to a given full rank matrix $\boldsymbol{A}$ under elementary column operations. Sparsity helps simplify many fundamental matrix operations (see [DER86]), and the problem has applications in areas such as machine learning [SS00] and in discovering cycle bases of graphs [KMMP04]. [GN10] discusses connections between the two problems and also to other problems in complexity theory.

Sparse (Complete) Dictionary Learning: In dictionary learning, given a data matrix $\boldsymbol{Y}$, one seeks an approximation $\boldsymbol{Y} \approx \boldsymbol{A}\boldsymbol{X}$, such that $\boldsymbol{A}$ is a representation dictionary with certain desired structure and $\boldsymbol{X}$ collects the representation coefficients with maximal sparsity. Such compact representation naturally allows signal compression, and also facilitates efficient signal acquisition and classification (see relevant discussion in [MBP14]). When $\boldsymbol{A}$ is required to be complete (i.e., square and invertible), by linear algebra, we have[2] $\mathrm{row}(\boldsymbol{Y}) = \mathrm{row}(\boldsymbol{X})$ [SWW12b]. Then the problem reduces to finding sparsest vectors (directions) in the known subspace $\mathrm{row}(\boldsymbol{Y})$, i.e. (1.0.2). Insights into this problem have led to new theoretical developments on complete dictionary learning [SWW12b, HD13, SQW15a].

Sparse Principal Component Analysis (Sparse PCA): In geometric terms, Sparse PCA (see, e.g., [ZHT06, JL09, dEGJL07] for early developments and [KNV+15, MW15] for discussion of recent results) concerns stable estimation of a linear subspace spanned by a sparse basis, in the data-poor regime, i.e., when the available data are not numerous enough to allow one to decouple the subspace estimation and sparsification tasks. Formally, given a data matrix $\boldsymbol{Z} = \boldsymbol{U}_0\boldsymbol{X}_0 + \boldsymbol{E}$,[3] where $\boldsymbol{Z} \in \mathbb{R}^{p\times n}$ collects column-wise n data points, $\boldsymbol{U}_0 \in \mathbb{R}^{p\times r}$ is the sparse basis, and $\boldsymbol{E}$ is a noise matrix, one is asked to estimate $\boldsymbol{U}_0$ (up to sign, scale, and permutation). Such a factorization finds applications in gene expression, financial data analysis and pattern

[2] Here, $\mathrm{row}(\cdot)$ denotes the row space.

[3] Variants of multiple-component formulations often add an additional orthonormality constraint on $\boldsymbol{U}_0$ but involve a different notation of sparsity; see, e.g., [ZHT06, VCLR13, LV+15a, WLL14].

recognition [dEJL07]. When the subspace is known (say by the PCA estimator with enough data samples), the problem again reduces to instances of (1.0.2) and is already nontrivial[4]. The full geometric sparse PCA can be treated as finding sparse vectors in a subspace that is subject to perturbation.

In addition, variants and generalizations of the problem (1.0.2) have also been studied in applications regarding control and optimization [ZF13], nonrigid structure from motion [DLH12], spectral estimation and Prony's problem [BM05], outlier rejection in PCA [MR15], blind source separation [ZP01], graphical model learning [AHJK13], and sparse coding on manifolds [HXV13]; see also [NSU15] and the references therein.

1.2 Prior Arts

Despite these potential applications of problem (1.0.2), it is only very recently that efficient computational surrogates with nontrivial recovery guarantees have been discovered for some cases of practical interest. In the context of sparse dictionary learning, Spielman et al. [SWW12b] introduced a convex relaxation which replaces the nonconvex problem (1.0.2) with a sequence of linear programs:

$$\ell^1/\ell^\infty \text{ Relaxation:} \qquad \min_x \|x\|_1, \quad \text{s.t.} \quad x(i) = 1,\ x \in \mathcal{S},\ 1 \le i \le p. \tag{1.2.1}$$

They proved that when $\mathcal{S}$ is generated as a span of n random sparse vectors, with high probability (w.h.p.), the relaxation recovers these vectors, provided the probability of an entry being nonzero is at most $\theta \in O\left(1/\sqrt{n}\right)$. In the *planted sparse model*, in which $\mathcal{S}$ is formed as direct sum of a single sparse vector $\boldsymbol{x}_0$ and a "generic" subspace, Hand and Demanet proved that (1.2.1) also correctly recovers $\boldsymbol{x}_0$, provided the fraction of nonzeros in $\boldsymbol{x}_0$ scales as $\theta \in O\left(1/\sqrt{n}\right)$ [HD13]. One might imagine improving these results by tightening the analyses. Unfortunately, the results of [SWW12b, HD13] are essentially sharp: *when θ substantially exceeds $\Omega(1/\sqrt{n})$, in both models the relaxation* (1.2.1) *provably breaks down.* Moreover, the most natural semidefinite programming (SDP) relaxation of (1.0.1),

$$\min_{\boldsymbol{X}} \|\boldsymbol{X}\|_1, \quad \text{s.t.} \quad \left\langle \boldsymbol{A}^\top \boldsymbol{A}, \boldsymbol{X} \right\rangle = 0,\ \operatorname{trace}[\boldsymbol{X}] = 1,\ \boldsymbol{X} \succeq \boldsymbol{0}. \tag{1.2.2}$$

also breaks down at exactly the same threshold of $\theta \sim O(1/\sqrt{n})$.[5]

[4] [HD13] has also discussed this data-rich sparse PCA setting.

[5] This breakdown behavior is again in sharp contrast to the standard sparse approximation problem (with $\boldsymbol{b} \neq \boldsymbol{0}$), in which it is possible to handle very large fractions of nonzeros (say, $\theta = \Omega(1/\log n)$, or even $\theta = \Omega(1)$) using a very simple ℓ^1 relaxation [CT05, Don06]

Table 1.1: Comparison of existing methods for recovering a planted sparse vector in a subspace

Method	Recovery Condition	Time Complexity[6]
ℓ^1/ℓ^∞ Relaxation [HD13]	$\theta \in O(1/\sqrt{n})$	$O(n^3 p \log(1/\varepsilon))$
SDP Relaxation	$\theta \in O(1/\sqrt{n})$	$O\left(p^{3.5}\log\left(1/\varepsilon\right)\right)$
SOS Relaxation [BKS13b]	$p \geq \Omega(n^2), \theta \in O(1)$	$\sim O(p^7 \log(1/\varepsilon))$ [7]
Spectral Method [HSSS15]	$p \geq \Omega(n^2 \text{poly}\log(n)), \theta \in O(1)$	$O\left(np\log(1/\epsilon)\right)$
This work	$p \geq \Omega(n^4 \log n),\ \theta \in O(1)$	$O(n^5 p^2 \log n + n^3 p \log(1/\varepsilon))$

One might naturally conjecture that this $1/\sqrt{n}$ threshold is simply an intrinsic price we must pay for having an efficient algorithm, even in these random models. Some evidence towards this conjecture might be borrowed from the superficial similarity of (1.0.2)-(1.2.2) and *sparse PCA* [ZHT06]. In sparse PCA, there is a substantial gap between what can be achieved with currently available efficient algorithms and the information theoretic optimum [BR13, KNV+15]. Is this also the case for recovering a sparse vector in a subspace? *Is $\theta \in O\left(1/\sqrt{n}\right)$ simply the best we can do with efficient, guaranteed algorithms?*

Remarkably, this is not the case. Recently, Barak et al. introduced a new rounding technique for sum-of-squares relaxations, and showed that the sparse vector $\boldsymbol{x}_0$ in the planted sparse model can be recovered when $p \geq \Omega\left(n^2\right)$ and $\theta = \Omega(1)$ [BKS13b]. It is perhaps surprising that this is possible at all with a polynomial time algorithm. Unfortunately, the runtime of this approach is a high-degree polynomial in p (see Table 1.1); for machine learning problems in which p is often either the feature dimension or the sample size, this algorithm is mostly of theoretical interest only. However, it raises an interesting algorithmic question: *Is there a practical algorithm that provably recovers a sparse vector with $\theta \gg 1/\sqrt{n}$ portion of nonzeros from a generic subspace $\mathcal{S}$?*

1.3 Contributions and Recent Developments

In this thesis, we address the above problem under the planted sparse model. We allow $\boldsymbol{x}_0$ to have up to $\theta_0 p$ nonzero entries, where $\theta_0 \in (0,1)$ is a constant. We provide a relatively simple algorithm which, w.h.p., exactly recovers $\boldsymbol{x}_0$, provided that $p \geq \Omega\left(n^4 \log n\right)$. A comparison of our results with existing methods is shown in Table 1.1. After publication of this work, Hopkins et al. [HSSS15] proposed a different simple algorithm based on the spectral method. This algorithm guarantees recovery of the planted sparse vector

[6] All estimates here are based on the standard interior point methods for solving linear and semidefinite programs. Customized solvers may result in order-wise speedup for specific problems. ε is the desired numerical accuracy.

[7] Here our estimation is based on the degree-4 SOS hierarchy used in [BKS13b] to obtain an initial approximate recovery.

also up to linear sparsity, whenever $p \geq \Omega(n^2 \text{polylog}(n))$, and comes with better time complexity.[8]

Our algorithm is based on alternating directions, with two special twists. First, we introduce a special data driven initialization, which seems to be important for achieving $\theta = \Omega(1)$. Second, our theoretical results require a second, linear programming based rounding phase, which is similar to [SWW12b]. Our core algorithm has very simple iterations, of linear complexity in the size of the data, and hence should be scalable to moderate-to-large scale problems.

Besides enjoying the $\theta \sim \Omega(1)$ guarantee that is out of the reach of previous practical algorithms, our algorithm performs well in simulations – empirically succeeding with $p \geq \Omega(n \text{ polylog}(n))$. It also performs well empirically on more challenging data models, such as the complete dictionary learning model, in which the subspace of interest contains not one, but n random target sparse vectors. This is encouraging, as breaking the $O(1/\sqrt{n})$ sparsity barrier with a practical algorithm and optimal guarantee is an important problem in theoretical dictionary learning [AGM13, AAN13, AAJ+13, ABGM14, AGMM15]. In this regard, our recent work [SQW15a] presents an efficient algorithm based on Riemannian optimization that guarantees recovery up to linear sparsity. However, the result is based on different ideas: a different nonconvex formulation, optimization algorithm, and analysis methodology.

[8] Despite these improved guarantees in the planted sparse model, our method still produces more appealing results on real imagery data – see Section 5 for examples.

Chapter 2

Problem Formulation and Global Optimality

We study the problem of recovering a sparse vector $\boldsymbol{x}_0 \neq \boldsymbol{0}$ (up to scale), which is an element of a known subspace $\mathcal{S} \subset \mathbb{R}^p$ of dimension n, provided an arbitrary orthonormal basis $\boldsymbol{Y} \in \mathbb{R}^{p \times n}$ for $\mathcal{S}$. Our starting point is the nonconvex formulation (1.0.2). Both the objective and the constraint set are nonconvex, and hence it is not easy to optimize over. We relax (1.0.2) by replacing the ℓ^0 norm with the ℓ^1 norm. For the constraint $\boldsymbol{x} \neq \boldsymbol{0}$, since in most applications we only care about the solution up to scaling, it is natural to force $\boldsymbol{x}$ to live on the unit sphere $\mathbb{S}^{n-1}$, giving

$$\min_{\boldsymbol{x}} \ \|\boldsymbol{x}\|_1 , \quad \text{s.t.} \quad \boldsymbol{x} \in \mathcal{S}, \ \|\boldsymbol{x}\| = 1. \tag{2.0.1}$$

This formulation is still nonconvex, and for general nonconvex problems it is known to be NP-hard to find even a local minimizer [MK87]. Nevertheless, the geometry of the sphere is benign enough, such that for well-structured inputs it actually *will* be possible to give algorithms that find the global optimizer.

The formulation (2.0.1) can be contrasted with (1.2.1), in which effectively we optimize the ℓ^1 norm subject to the constraint $\|\boldsymbol{x}\|_\infty = 1$: because the set $\{\boldsymbol{x} : \|\boldsymbol{x}\|_\infty = 1\}$ is polyhedral, the ℓ^∞-constrained problem immediately yields a sequence of linear programs. This is very convenient for computation and analysis. However, it suffers from the aforementioned breakdown behavior around $\|\boldsymbol{x}_0\|_0 \sim p/\sqrt{n}$. In contrast, though the sphere $\|\boldsymbol{x}\| = 1$ is a more complicated geometric constraint, it will allow much larger number of nonzeros

in $\boldsymbol{x}_0$. Indeed, if we consider the global optimizer of a reformulation of (2.0.1):

$$\min_{\boldsymbol{q}\in\mathbb{R}^n} \|\boldsymbol{Y}\boldsymbol{q}\|_1, \quad \text{s.t.} \quad \|\boldsymbol{q}\| = 1, \tag{2.0.2}$$

where $\boldsymbol{Y}$ is any orthonormal basis for $\mathcal{S}$, the sufficient condition that guarantees exact recovery under the planted sparse model for the subspace is as follows:

Theorem 2.1 (ℓ^1/ℓ^2 recovery, planted sparse model) *There exists a constant $\theta_0 > 0$, such that if the subspace $\mathcal{S}$ follows the planted sparse model*

$$\mathcal{S} = \text{span}\,(\boldsymbol{x}_0, \boldsymbol{g}_1, \ldots, \boldsymbol{g}_{n-1}) \subset \mathbb{R}^p,$$

where $\boldsymbol{g}_i \sim_{i.i.d.} \mathcal{N}(\boldsymbol{0}, \frac{1}{p}\boldsymbol{I})$, and $\boldsymbol{x}_0 \sim_{i.i.d.} \frac{1}{\sqrt{\theta p}}\text{Ber}(\theta)$ are all jointly independent and $1/\sqrt{n} < \theta < \theta_0$, then the unique (up to sign) optimizer $\boldsymbol{q}^\star$ to (2.0.2), for any orthonormal basis $\boldsymbol{Y}$ of $\mathcal{S}$, produces $\boldsymbol{Y}\boldsymbol{q}^\star = \xi\boldsymbol{x}_0$ for some $\xi \neq 0$ with probability at least $1 - cp^{-2}$, provided $p \geq Cn$. Here c and C are positive constants.

Hence, *if* we could find the global optimizer of (2.0.2), we would be able to recover $\boldsymbol{x}_0$ whose number of nonzero entries is quite large – even linear in the dimension p ($\theta = \Omega(1)$). On the other hand, it is not obvious that this should be possible: (2.0.2) is nonconvex. In the next section, we will describe a simple heuristic algorithm for approximately solving a relaxed version of the ℓ^1/ℓ^2 problem (2.0.2). More surprisingly, we will then prove that for a class of random problem instances, this algorithm, plus an auxiliary rounding technique, actually recovers the global optimizer – the target sparse vector $\boldsymbol{x}_0$. The proof requires a detailed probabilistic analysis, which is sketched in Section 4.

Before continuing, it is worth noting that the formulation (2.0.1) is in no way novel – see, e.g., the work of [ZP01] in blind source separation for precedent. However, our algorithms and subsequent analysis are novel.

Chapter 3

Algorithm

To develop an algorithm for solving (2.0.2), it is useful to consider a slight relaxation of (2.0.2), in which we introduce an auxiliary variable $\boldsymbol{x} \approx \boldsymbol{Y}\boldsymbol{q}$:

$$\min_{q,x} f(\boldsymbol{q}, \boldsymbol{x}) \doteq \frac{1}{2} \|Yq - x\|^2 + \lambda \|x\|_1, \quad \text{s.t.} \quad \|q\| = 1. \tag{3.0.1}$$

Here, $\lambda > 0$ is a penalty parameter. It is not difficult to see that this problem is equivalent to minimizing the *Huber* M-estimator over $\boldsymbol{Y}\boldsymbol{q}$. This relaxation makes it possible to apply the alternating direction method to this problem. This method starts from some initial point $\boldsymbol{q}^{(0)}$, alternates between optimizing with respect to (w.r.t.) $\boldsymbol{x}$ and optimizing w.r.t. $\boldsymbol{q}$:

$$x^{(k+1)} = \arg\min_{\boldsymbol{x}} \frac{1}{2} \left\| Yq^{(k)} - x \right\|^2 + \lambda \|x\|_1, \tag{3.0.2}$$

$$q^{(k+1)} = \arg\min_{\boldsymbol{q}} \frac{1}{2} \left\| Yq - x^{(k+1)} \right\|^2 \text{ s.t. } \|\boldsymbol{q}\| = 1, \tag{3.0.3}$$

where $x^{(k)}$ and $q^{(k)}$ denote the values of x and q in the k-th iteration. Both (3.0.2) and (3.0.3) have simple closed form solutions:

$$x^{(k+1)} = S_\lambda[Yq^{(k)}], \qquad q^{(k+1)} = \frac{Y^\top x^{(k+1)}}{\left\| Y^\top x^{(k+1)} \right\|}, \tag{3.0.4}$$

where $S_\lambda[x] = \text{sign}(x) \max\{|x| - \lambda, 0\}$ is the soft-thresholding operator. The proposed ADM algorithm is summarized in Algorithm 0.

The algorithm is simple to state and easy to implement. However, if our goal is to recover the *sparsest* vector $\boldsymbol{x}_0$, some additional tricks are needed.

Algorithm 1 Nonconvex ADM for solving (3.0.1)

Input: A matrix $Y \in \mathbb{R}^{p\times n}$ with $\boldsymbol{Y}^\top \boldsymbol{Y} = \boldsymbol{I}$, initialization $q^{(0)}$, threshold parameter $\lambda > 0$.
Output: The recovered sparse vector $\hat{x}_0 = Y q^{(k)}$
for $k = 0, \ldots, O\left(n^4 \log n\right)$ **do**
$\boldsymbol{x}^{(k+1)} = S_\lambda[\boldsymbol{Y}\boldsymbol{q}^{(k)}]$,
$\boldsymbol{q}^{(k+1)} = \frac{\boldsymbol{Y}^\top \boldsymbol{x}^{(k+1)}}{\left\|\boldsymbol{Y}^\top \boldsymbol{x}^{(k+1)}\right\|}$,
end for

Initialization. Because the problem (2.0.2) is nonconvex, an arbitrary or random initialization may not produce a global minimizer.[1] In fact, good initializations are critical for the proposed ADM algorithm to succeed in the linear sparsity regime. For this purpose, we suggest using every normalized row of Y as initializations for q, and solving a sequence of p nonconvex programs (2.0.2) by the ADM algorithm.

To get an intuition of why our initialization works, recall the planted sparse model $\mathcal{S} = \mathrm{span}(\boldsymbol{x}_0, \boldsymbol{g}_1, \ldots, \boldsymbol{g}_{n-1})$ and suppose

$$\overline{\boldsymbol{Y}} \;=\; [\boldsymbol{x}_0 \mid \boldsymbol{g}_1 \mid \cdots \mid \boldsymbol{g}_{n-1}] \;\in\; \mathbb{R}^{p\times n}. \tag{3.0.5}$$

If we take a row $\overline{\boldsymbol{y}}^i$ of $\overline{\boldsymbol{Y}}$, in which $x_0(i)$ is nonzero, then $x_0(i) = \Theta\left(1/\sqrt{\theta p}\right)$. Meanwhile, the entries of $\boldsymbol{g}_1(i), \ldots \boldsymbol{g}_{n-1}(i)$ are all $\mathcal{N}(0, 1/p)$, and so their magnitude have size about $1/\sqrt{p}$. Hence, when θ is not too large, $x_0(i)$ will be somewhat bigger than most of the other entries in $\overline{\boldsymbol{y}}^i$. Put another way, $\overline{\boldsymbol{y}}^i$ *is biased towards the first standard basis vector* $\boldsymbol{e}_1$. Now, under our probabilistic model assumptions, $\overline{\boldsymbol{Y}}$ is very well conditioned: $\overline{\boldsymbol{Y}}^\top \overline{\boldsymbol{Y}} \approx \boldsymbol{I}$.[2] Using the Gram-Schmidt process[3], we can find an orthonormal basis $\boldsymbol{Y}$ for $\mathcal{S}$ via:

$$\overline{\boldsymbol{Y}} = \boldsymbol{Y}\boldsymbol{R}, \tag{3.0.6}$$

where $\boldsymbol{R}$ is upper triangular, and $\boldsymbol{R}$ is itself well-conditioned: $\boldsymbol{R} \approx \boldsymbol{I}$. Since the i-th row $\overline{\boldsymbol{y}}^i$ of $\overline{\boldsymbol{Y}}$ is biased in the direction of $\boldsymbol{e}_1$ and $\boldsymbol{R}$ is well-conditioned, the i-th row $\boldsymbol{y}^i$ of $\boldsymbol{Y}$ is also biased in the direction of $\boldsymbol{e}_1$. In other words, with this canonical orthobasis $\boldsymbol{Y}$ for the subspace, *the i-th row of $\boldsymbol{Y}$ is biased in the direction of the global optimizer*. The heuristic arguments are made rigorous in Appendix A.2 and Section 7.2.

What if we are handed some other basis $\widehat{\boldsymbol{Y}} = \boldsymbol{Y}\boldsymbol{U}$, where $\boldsymbol{U}$ is an arbitary orthogonal matrix? Suppose $\boldsymbol{q}_\star$ is a global optimizer to (2.0.2) with the input matrix $\boldsymbol{Y}$, then it is easy to check that, $\boldsymbol{U}^\top \boldsymbol{q}_\star$ is a global

[1] More precisely, in our models, random initialization *does* work, but only when the subspace dimension n is *extremely* low compared to the ambient dimension p.

[2] This is the common heuristic that "tall random matrices are well conditioned" [Ver10].

[3] ...QR decomposition in general with restriction that $R_{11} = 1$.

optimizer to (2.0.2) with the input matrix $\widehat{\boldsymbol{Y}}$. Because

$$\left\langle (\boldsymbol{Y}\boldsymbol{U})^\top \boldsymbol{e}_i, \boldsymbol{U}^\top \boldsymbol{q}_\star \right\rangle = \left\langle \boldsymbol{Y}^\top \boldsymbol{e}_i, \boldsymbol{q}_\star \right\rangle,$$

our initialization is *invariant* to any rotation of the orthobasis. Hence, *even if we are handed an arbitrary orthobasis for $\mathcal{S}$, the i-th row is still biased in the direction of the global optimizer*.

Rounding by linear programming (LP). Let $\overline{\boldsymbol{q}}$ denote the output of Algorithm **??**. As illustrated in Fig. 4.1, we will prove that with our particular initialization and an appropriate choice of λ, ADM algorithm uniformly moves towards the optimal over a large portion of the sphere, and its solution falls within a certain small radius of the globally optimal solution $\boldsymbol{q}_\star$ to (2.0.2). To exactly recover $\boldsymbol{q}_\star$, or equivalently to recover the exact sparse vector $\boldsymbol{x}_0 = \gamma \boldsymbol{Y}\boldsymbol{q}_\star$ for some $\gamma \neq 0$, we solve the linear program

$$\min_{\boldsymbol{q}} \|\boldsymbol{Y}\boldsymbol{q}\|_1 \quad \text{s.t.} \quad \langle \boldsymbol{r}, \boldsymbol{q} \rangle = 1 \tag{3.0.7}$$

with $\boldsymbol{r} = \overline{\boldsymbol{q}}$. Since the feasible set $\{\boldsymbol{q} \mid \langle \overline{\boldsymbol{q}}, \boldsymbol{q} \rangle = 1\}$ is essentially the tangent space of the sphere $\mathbb{S}^{n-1}$ at $\overline{\boldsymbol{q}}$, whenever $\overline{\boldsymbol{q}}$ is close enough to $\boldsymbol{q}_\star$, one should expect that the optimizer of (3.0.7) exactly recovers $\boldsymbol{q}_\star$ and hence $\boldsymbol{x}_0$ up to scale. We will prove that this is indeed true under appropriate conditions.

Chapter 4

Main Result and Sketch of Analysis

4.1 Main Results

In this section, we describe our main theoretical result, which shows that w.h.p. the algorithm described in the previous section succeeds.

Theorem 4.1 *Suppose that $\mathcal{S}$ obeys the planted sparse model, and let the columns of $\boldsymbol{Y}$ form an arbitrary orthonormal basis for the subspace $\mathcal{S}$. Let $\boldsymbol{y}^1, \ldots, \boldsymbol{y}^p \in \mathbb{R}^n$ denote the (transposes of) the rows of $\boldsymbol{Y}$. Apply Algorithm **??** with $\lambda = 1/\sqrt{p}$, using initializations $\boldsymbol{q}^{(0)} = \boldsymbol{y}^1 / \left\|\boldsymbol{y}^1\right\|, \ldots, \boldsymbol{y}^p / \left\|\boldsymbol{y}^p\right\|$, to produce outputs $\overline{\boldsymbol{q}}_1, \ldots, \overline{\boldsymbol{q}}_p$. Solve the linear program (3.0.7) with $\boldsymbol{r} = \overline{\boldsymbol{q}}_1, \ldots, \overline{\boldsymbol{q}}_p$, to produce $\widehat{\boldsymbol{q}}_1, \ldots, \widehat{\boldsymbol{q}}_p$. Set $i^\star \in \arg\min_i \|\boldsymbol{Y}\widehat{\boldsymbol{q}}_i\|_1$. Then*

$$\boldsymbol{Y}\widehat{\boldsymbol{q}}_{i^\star} = \gamma \boldsymbol{x}_0, \tag{4.1.1}$$

for some $\gamma \neq 0$ with probability at least $1 - cp^{-2}$, provided

$$p \geq Cn^4 \log n, \qquad \text{and} \qquad \frac{1}{\sqrt{n}} \leq \theta \leq \theta_0. \tag{4.1.2}$$

Here C, c and θ_0 are positive constants.

Remark 4.2 *We can see that the result in Theorem 4.1 is suboptimal in sample complexity compared to the global optimality result in Theorem 2.1 and Barak et al.'s result [BKS13b] (and the subsequent work [HSSS15]). For successful recovery, we require $p \geq \Omega\left(n^4 \log n\right)$, while the global optimality and Barak et al. demand $p \geq \Omega\left(n\right)$ and $p \geq \Omega\left(n^2\right)$, respectively. Aside from possible deficiencies in our current analysis, compared to Barak et al., we believe this is still the first practical and efficient method which is guaranteed to achieve $\theta \sim \Omega(1)$ rate. The*

lower bound on θ in Theorem 4.1 is mostly for convenience in the proof; in fact, the LP rounding stage of our algorithm already succeeds w.h.p. when $\theta \in O\left(1/\sqrt{n}\right)$.

4.2 A Sketch of Analysis

In this section, we briefly sketch the main ideas of proving our main result in Theorem 4.1, to show that the "initialization + ADM + LP rounding" pipeline recovers $\boldsymbol{x}_0$ under the stated technical conditions, as illustrated in Fig. 4.1. The proof of our main result requires rather detailed technical analysis of the iteration-by-iteration properties of Algorithm 0, most of which is deferred to the appendices.

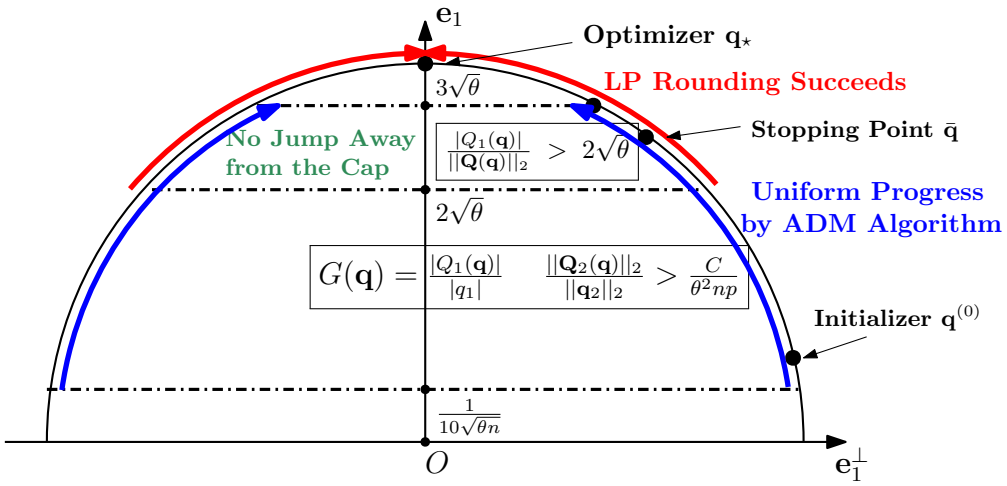


Figure 4.1: An illustration of the proof sketch for our ADM algorithm.

As noted in Section 3, the ADM algorithm is invariant to change of basis. So w.l.o.g., let us assume $\overline{\boldsymbol{Y}} = [\boldsymbol{x}_0 \mid \boldsymbol{g}_1 \mid \cdots \mid \boldsymbol{g}_{n-1}]$ and let $\boldsymbol{Y}$ to be its orthogonalization, i.e., [1]

$$\boldsymbol{Y} = \left[\frac{\boldsymbol{x}_0}{\|\boldsymbol{x}_0\|} \mid \mathcal{P}_{\boldsymbol{x}_0^\perp}\boldsymbol{G}\left(\boldsymbol{G}^\top \mathcal{P}_{\boldsymbol{x}_0^\perp}\boldsymbol{G}\right)^{-1/2}\right]. \tag{4.2.1}$$

When p is large, $\overline{\boldsymbol{Y}}$ is nearly orthogonal, and hence $\overline{\boldsymbol{Y}}$ is very close to $\boldsymbol{Y}$. Thus, in our proofs, whenever convenient, we make the arguments on $\overline{\boldsymbol{Y}}$ first and then "propagate" the quantitative results onto $\boldsymbol{Y}$ by perturbation arguments. With that noted, let $\boldsymbol{y}^1, \cdots, \boldsymbol{y}^p$ be the transpose of the rows of $\boldsymbol{Y}$, and note that these are all independent random vectors. To prove the result of Theorem 4.1, we need the following results. First, given the specified $\boldsymbol{Y}$, we show that our initialization is biased towards the global optimum:

[1] Note that with probability one, the inverse matrix square-root in $\boldsymbol{Y}$ is well defined. So $\boldsymbol{Y}$ is well defined w.h.p. (i.e., except for $\boldsymbol{x}_0 = 0$). See more quantitative characterization of $\boldsymbol{Y}$ in Appendix A.2.

Proposition 4.3 (Good initialization) *Suppose $\theta > 1/\sqrt{n}$ and $p \geq Cn$. It holds with probability at least $1 - cp^{-2}$ that at least one of our p initialization vectors suggested in Section 3, say $\boldsymbol{q}_i^{(0)} = \boldsymbol{y}^i / \left\|\boldsymbol{y}^i\right\|$, obeys*

$$\left| \left\langle \frac{\boldsymbol{y}^i}{\|\boldsymbol{y}^i\|}, \boldsymbol{e}_1 \right\rangle \right| \geq \frac{1}{10\sqrt{\theta n}}. \tag{4.2.2}$$

Here C, c are positive constants.

Proof See Section 7.2. ∎

Second, we define a vector-valued random process $\boldsymbol{Q}(\boldsymbol{q})$ on $\boldsymbol{q} \in \mathbb{S}^{n-1}$, via

$$\boldsymbol{Q}(\boldsymbol{q}) = \frac{1}{p} \sum_{i=1}^{p} \boldsymbol{y}^i S_\lambda \left[\boldsymbol{q}^\top \boldsymbol{y}^i \right], \tag{4.2.3}$$

so that based on (3.0.4), one step of the ADM algorithm takes the form:

$$\boldsymbol{q}^{(k+1)} = \frac{\boldsymbol{Q}\left(\boldsymbol{q}^{(k)}\right)}{\left\|\boldsymbol{Q}\left(\boldsymbol{q}^{(k)}\right)\right\|} \tag{4.2.4}$$

This is a very favorable form for analysis: the term in the numerator $\boldsymbol{Q}\left(\boldsymbol{q}^{(k)}\right)$ is a sum of p independent random vectors with $\boldsymbol{q}^{(k)}$ viewed as fixed. We study the behavior of the iteration (4.2.4) through the random process $\boldsymbol{Q}\left(\boldsymbol{q}^{(k)}\right)$. We want to show that w.h.p. the ADM iterate sequence $\boldsymbol{q}^{(k)}$ converges to some small neighborhood of $\pm\boldsymbol{e}_1$, so that the ADM algorithm plus the LP rounding (described in Section 3) successfully retrieves the sparse vector $\boldsymbol{x}_0/\|\boldsymbol{x}_0\| = \boldsymbol{Y}\boldsymbol{e}_1$. Thus, we hope that in general, $\boldsymbol{Q}(\boldsymbol{q})$ is more concentrated on the first coordinate than $\boldsymbol{q} \in \mathbb{S}^{n-1}$. Let us partition the vector $\boldsymbol{q}$ as $\boldsymbol{q} = [q_1; \boldsymbol{q}_2]$, with $q_1 \in \mathbb{R}$ and $\boldsymbol{q}_2 \in \mathbb{R}^{n-1}$; and correspondingly $\boldsymbol{Q}(\boldsymbol{q}) = [Q_1(\boldsymbol{q}); \boldsymbol{Q}_2(\boldsymbol{q})]$. The inner product of $\boldsymbol{Q}(\boldsymbol{q})/\left\|\boldsymbol{Q}(\boldsymbol{q})\right\|$ and $\boldsymbol{e}_1$ is strictly larger than the inner product of $\boldsymbol{q}$ and $\boldsymbol{e}_1$ if and only if

$$\frac{|Q_1(\boldsymbol{q})|}{|q_1|} > \frac{\|\boldsymbol{Q}_2(\boldsymbol{q})\|}{\|\boldsymbol{q}_2\|}.$$

In the following proposition, we show that w.h.p., this inequality holds uniformly over a significant portion of the sphere

$$\Gamma \doteq \left\{ \boldsymbol{q} \in \mathbb{S}^{n-1} \mid \frac{1}{10\sqrt{n\theta}} \leq |q_1| \leq 3\sqrt{\theta}, \|\boldsymbol{q}_2\| \geq \frac{1}{10} \right\}, \tag{4.2.5}$$

so the algorithm moves in the correct direction. Let us define the gap $G(\boldsymbol{q})$ between the two quantities $|Q_1(\boldsymbol{q})|\,/\,|q_1|$ and $\|\boldsymbol{Q}_2(\boldsymbol{q})\|\,/\,\|\boldsymbol{q}_2\|$ as

$$G(\boldsymbol{q}) \doteq \frac{|Q_1(\boldsymbol{q})|}{|q_1|} - \frac{\|\boldsymbol{Q}_2(\boldsymbol{q})\|}{\|\boldsymbol{q}_2\|}, \tag{4.2.6}$$

and we show that the following result is true:

> **Proposition 4.4 (Uniform lower bound for finite sample gap)** *There exists a constant $\theta_0 \in (0,1)$, such that when $p \geq Cn^4 \log n$, the estimate*
>
> $$\inf_{\boldsymbol{q}\in\Gamma} G(\boldsymbol{q}) \;\geq\; \frac{1}{10^4\theta^2 np}$$
>
> *holds with probability at least $1 - cp^{-2}$, provided $\theta \in (1/\sqrt{n}, \theta_0)$. Here C, c are positive constants.*

Proof See Section 7.3. ■

Next, we show that whenever $|q_1| \geq 3\sqrt{\theta}$, w.h.p. the iterates stay in a "safe region" with $|q_1| \geq 2\sqrt{\theta}$ which is enough for LP rounding (3.0.7) to succeed.

> **Proposition 4.5 (Safe region for rounding)** *There exists a constant $\theta_0 \in (0,1)$, such that when $p \geq Cn^4 \log n$, it holds with probability at least $1 - cp^{-2}$ that*
>
> $$\frac{|Q_1(\boldsymbol{q})|}{\|\boldsymbol{Q}(\boldsymbol{q})\|} \;\geq\; 2\sqrt{\theta}$$
>
> *for all $\boldsymbol{q} \in \mathbb{S}^{n-1}$ satisfying $|q_1| > 3\sqrt{\theta}$, provided $\theta \in (1/\sqrt{n}, \theta_0)$. Here C, c are positive constants.*

Proof See Section 7.4. ■

In addition, the following result shows that the number of iterations for the ADM algorithm to reach the safe region can be bounded grossly by $O(n^4 \log n)$ w.h.p..

> **Proposition 4.6 (Iteration complexity of reaching the safe region)** *There is a constant $\theta_0 \in (0,1)$, such that when $p \geq Cn^4 \log n$, it holds with probability at least $1 - cp^{-2}$ that the ADM algorithm in Algorithm 0, with any initialization $\boldsymbol{q}^{(0)} \in \mathbb{S}^{n-1}$ satisfying $\left|q_1^{(0)}\right| \geq \frac{1}{10\sqrt{\theta n}}$, will produce some iterate $\overline{\boldsymbol{q}}$ with $|\bar{q}_1| > 3\sqrt{\theta}$ at least once in at most $O(n^4 \log n)$ iterations, provided $\theta \in (1/\sqrt{n}, \theta_0)$. Here C, c are positive constants.*

Proof See Section 7.5. ■

Moreover, we show that the LP rounding (3.0.7) with input $\boldsymbol{r} = \overline{\boldsymbol{q}}$ exactly recovers the optimal solution w.h.p., whenever the ADM algorithm returns a solution $\overline{\boldsymbol{q}}$ with first coordinate $|\bar{q}_1| > 2\sqrt{\theta}$.

> **Proposition 4.7 (Success of rounding)** *There is a constant $\theta_0 \in (0,1)$, such that when $p \geq Cn$, the following holds with probability at least $1 - cp^{-2}$ provided $\theta \in (1/\sqrt{n}, \theta_0)$: Suppose the input basis is $\boldsymbol{Y}$ defined in (4.2.1)*

and the ADM algorithm produces an output $\overline{\boldsymbol{q}} \in \mathbb{S}^{n-1}$ with $|\overline{q}_1| > 2\sqrt{\theta}$. Then the rounding procedure with $\boldsymbol{r} = \overline{\boldsymbol{q}}$ returns the desired solution $\pm\boldsymbol{e}_1$. Here C, c are positive constants.

Proof See Section 7.6. ■

Finally, given $p \geq Cn^4 \log n$ for a sufficiently large constant C, we combine all the results above to complete the proof of Theorem 4.1.

Proof [Proof of Theorem 4.1]

W.l.o.g., let us again first consider $\overline{\boldsymbol{Y}}$ as defined in (3.0.5) and its orthogonalization $\boldsymbol{Y}$ in a "natural/canonical" form (4.2.1). We show that w.h.p. our algorithmic pipeline described in Section 3 exactly recovers the optimal solution up to scale, via the following argument:

1. **Good initializers**. Proposition 4.3 shows that w.h.p., at least one of the p initialization vectors, say $\boldsymbol{q}_i^{(0)} = \boldsymbol{y}^i / \left\|\boldsymbol{y}^i\right\|$, obeys

$$\left|\left\langle \boldsymbol{q}_i^{(0)}, \boldsymbol{e}_1 \right\rangle\right| \geq \frac{1}{10\sqrt{\theta n}},$$

   which implies that $\boldsymbol{q}_i^{(0)}$ is biased towards the global optimal solution.

2. **Uniform progress away from the equator**. By Proposition 4.4, for any $\theta \in (1/\sqrt{n}, \theta_0)$ with a constant $\theta_0 \in (0,1)$,

$$G(\boldsymbol{q}) = \frac{|Q_1(\boldsymbol{q})|}{|q_1|} - \frac{\|\boldsymbol{Q}_2(\boldsymbol{q})\|}{\|\boldsymbol{q}\|} \geq \frac{1}{10^4\theta^2 np} \tag{4.2.7}$$

   holds uniformly for all $\boldsymbol{q} \in \mathbb{S}^{n-1}$ in the region $\frac{1}{10\sqrt{\theta n}} \leq |q_1| \leq 3\sqrt{\theta}$ w.h.p.. This implies that with an input $\boldsymbol{q}^{(0)}$ such that $\left|q_1^{(0)}\right| \geq \frac{1}{10\sqrt{\theta n}}$, the ADM algorithm will eventually obtain a point $\boldsymbol{q}^{(k)}$ for which $\left|q^{(k)}\right| \geq 3\sqrt{\theta}$, if sufficiently many iterations are allowed.

3. **No jumps away from the caps**. Proposition 4.5 shows that for any $\theta \in (1/\sqrt{n}, \theta_0)$ with a constant $\theta_0 \in (0,1)$, w.h.p.,

$$\frac{Q_1(\boldsymbol{q})}{\|\boldsymbol{Q}(\boldsymbol{q})\|} \geq 2\sqrt{\theta}$$

   holds for all $\boldsymbol{q} \in \mathbb{S}^{n-1}$ with $|q_1| \geq 3\sqrt{\theta}$. This implies that once $|q_1^{(k)}| \geq 3\sqrt{\theta}$ for some iterate k, all the future iterates produced by the ADM algorithm stay in a "spherical cap" region around the optimum with $|q_1| \geq 2\sqrt{\theta}$.

4. **Location of stopping points**. As shown in Proposition 4.6, w.h.p., the strictly positive gap $G(\boldsymbol{q})$ in

(4.2.7) ensures that one needs to run at most $O\left(n^4 \log n\right)$ iterations to first encounter an iterate $\boldsymbol{q}^{(k)}$ such that $|q_1^{(k)}| \geq 3\sqrt{\theta}$. Hence, the steps above imply that, w.h.p., Algorithm 0 fed with the proposed initialization scheme successively produces iterates $\overline{\boldsymbol{q}} \in \mathbb{S}^{n-1}$ with its first coordinate $|\overline{q}_1| \geq 2\sqrt{\theta}$ after $O\left(n^4 \log n\right)$ steps.

5. **Rounding succeeds when** $|r_1| \geq 2\sqrt{\theta}$. Proposition 4.7 proves that w.h.p., the LP rounding (3.0.7) with an input $\boldsymbol{r} = \overline{\boldsymbol{q}}$ produces the solution $\pm \boldsymbol{x}_0$ up to scale.

Taken together, these claims imply that from at least one of the initializers $\boldsymbol{q}^{(0)}$, the ADM algorithm will produce an output $\overline{\boldsymbol{q}}$ which is accurate enough for LP rounding to exactly return $\boldsymbol{x}_0/\|\boldsymbol{x}_0\|_2$. On the other hand, our ℓ^1/ℓ^2 optimality theorem (Theorem 2.1) implies that $\pm \boldsymbol{x}_0$ are the unique vectors with the smallest ℓ^1 norm among all unit vectors in the subspace. Since w.h.p. $\boldsymbol{x}_0/\|\boldsymbol{x}_0\|_2$ is among the p unit vectors $\widehat{\boldsymbol{q}}_1, \ldots, \widehat{\boldsymbol{q}}_p$ our p row initializers finally produce, our minimal ℓ^1 norm selector will successfully locate $\boldsymbol{x}_0/\|\boldsymbol{x}_0\|_2$ vector.

For the general case when the input is an arbitrary orthonormal basis $\widehat{\boldsymbol{Y}} = \boldsymbol{Y}\boldsymbol{U}$ for some orthogonal matrix $\boldsymbol{U}$, the target solution is $\boldsymbol{U}^\top \boldsymbol{e}_1$. The following technical pieces are perfectly parallel to the argument above for $\boldsymbol{Y}$.

1. Discussion at the end of Section 7.2 implies that w.h.p., at least one row of $\widehat{\boldsymbol{Y}}$ provides an initial point $\boldsymbol{q}^{(0)}$ such that $\left|\left\langle \boldsymbol{q}^{(0)}, \boldsymbol{U}^\top \boldsymbol{e}_1 \right\rangle\right| \geq \frac{1}{10\sqrt{\theta n}}$.

2. Discussion following Proposition 4.4 in Section 7.3 indicates that for all $\boldsymbol{q}$ such that $\frac{1}{10\sqrt{\theta n}} \leq \left|\left\langle \boldsymbol{q}, \boldsymbol{U}^\top \boldsymbol{e}_1 \right\rangle\right| \leq 3\sqrt{\theta}$, there is a strictly positive gap, indicating steady progress towards a point $\boldsymbol{q}^{(k)}$ such that $\left|\left\langle \boldsymbol{q}^{(k)}, \boldsymbol{U}^\top \boldsymbol{e}_1 \right\rangle\right| \geq 3\sqrt{\theta}$.

3. Discussion at the end of Section 7.4 implies that once $\boldsymbol{q}$ satisfies $\left|\left\langle \boldsymbol{q}, \boldsymbol{U}^\top \boldsymbol{e}_1 \right\rangle\right|$, the next iterate will not move far away from the target:

$$\left|\left\langle \boldsymbol{Q}\left(\boldsymbol{q}; \widehat{\boldsymbol{Y}}\right) / \left\|\boldsymbol{Q}\left(\boldsymbol{q}; \widehat{\boldsymbol{Y}}\right)\right\|, \boldsymbol{U}^\top \boldsymbol{e}_1 \right\rangle\right| \ \geq \ 2\sqrt{\theta}.$$

4. Repeating the argument in Section 7.5 for general input $\widehat{\boldsymbol{Y}}$ shows it is enough to run the ADM algorithm $O\left(n^4 \log n\right)$ iterations to cross the range $\frac{1}{10\sqrt{\theta n}} \leq \left|\left\langle \boldsymbol{q}, \boldsymbol{U}^\top \boldsymbol{e}_1 \right\rangle\right| \leq 3\sqrt{\theta}$. So the argument above together dictates that with the proposed initialization, w.h.p., the ADM algorithm produces an output $\overline{\boldsymbol{q}}$ that satisfies $\left|\left\langle \overline{\boldsymbol{q}}, \boldsymbol{U}^\top \boldsymbol{e}_1 \right\rangle\right| \geq 2\sqrt{\theta}$, if we run at least $O\left(n^4 \log n\right)$ iterations.

5. Since the ADM returns $\overline{\boldsymbol{q}}$ satisfying $\left|\left\langle \overline{\boldsymbol{q}}, \boldsymbol{R}^\top \boldsymbol{e}_1 \right\rangle\right| \geq 2\sqrt{\theta}$, discussion at the end of Section 7.6 implies that we will obtain a solution $\boldsymbol{q}_\star = \pm \boldsymbol{U}^\top \boldsymbol{e}_1$ up to scale as the optimizer of the rounding program,

exactly the target solution.

Hence, we complete the proof. ■

Remark 4.8 *Under the planted sparse model, in practice the ADM algorithm with the proposed initialization converges to a global optimizer of* (3.0.1) *that correctly recovers* $\boldsymbol{x}_0$*. In fact, simple calculation shows such desired point for successful recovery is indeed the only critical point of* (3.0.1) *near the pole in Fig. 4.1. Unfortunately, using the current analytical framework, we did not succeed in proving such convergence in theory. Proposition 4.5 and 4.6 imply that after* $O(n^4 \log n)$ *iterations, however, the ADM sequence will stay in a small neighborhood of the target. Hence, we proposed to stop after* $O(n^4 \log n)$ *steps, and then round the output using the LP that provable recover the target, as implied by Proposition 4.5 and 4.7. So the LP rounding procedure is for the purpose of completing the theory, and seems not necessary in practice. We suspect alternative analytical strategies, such as the geometrical analysis that we will discuss in Section 6, can likely get around the artifact.*

Chapter 5

Numerical Results

5.1 Experimental Results

In this section, we show the performance of the proposed ADM algorithm on both synthetic and real datasets. On the synthetic dataset, we show the phase transition of our algorithm on both the planted sparse and the dictionary learning models; for the real dataset, we demonstrate how seeking sparse vectors can help discover interesting patterns on face images.

5.1.1 Phase Transition on Synthetic Data

For the planted sparse model, for each pair of (k, p), we generate the n dimensional subspace $\mathcal{S} \subset \mathbb{R}^p$ by direct sum of $\boldsymbol{x}_0$ and $\boldsymbol{G}$: $\boldsymbol{x}_0 \in \mathbb{R}^p$ is a k-sparse vector with uniformly random support and all nonzero entries equal to 1, and $\boldsymbol{G} \in \mathbb{R}^{p\times(n-1)}$ is an i.i.d. Gaussian matrix distributed by $\mathcal{N}(0, 1/p)$. So one basis $\boldsymbol{Y}$ of the subspace $\mathcal{S}$ can be constructed by $\boldsymbol{Y} = \texttt{GS}\left([\boldsymbol{x}_0, \boldsymbol{G}]\right)\boldsymbol{U}$, where $\texttt{GS}\left(\cdot\right)$ denotes the Gram-Schmidt orthonormalization operator and $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ is an arbitrary orthogonal matrix. For each p, we set the regularization parameter in (3.0.1) as $\lambda = 1/\sqrt{p}$, use all the normalized rows of $\boldsymbol{Y}$ as initializations of $\boldsymbol{q}$ for the proposed ADM algorithm, and run the alternating steps for 10^4 iterations. We determine the recovery to be successful whenever $\|\boldsymbol{x}_0/\left\|\boldsymbol{x}_0\right\| - \boldsymbol{Y}\boldsymbol{q}\| \leq 10^{-2}$ for at least one of the p trials (we set the tolerance relatively large as we have shown that LP rounding exactly recovers the solutions with approximate input). To determine the empirical recovery performance of our ADM algorithm, first we fix the relationship between n and p as $p = 5n\log n$, and plot out the phase transition between k and p. Next, we fix the sparsity level $\theta = 0.2$ (or

$k = 0.2p$), and plot out the phase transition between p and n. For each pair of (p, k) or (n, p), we repeat the simulation for 10 times. Fig. 5.1 shows both phase transition plots.

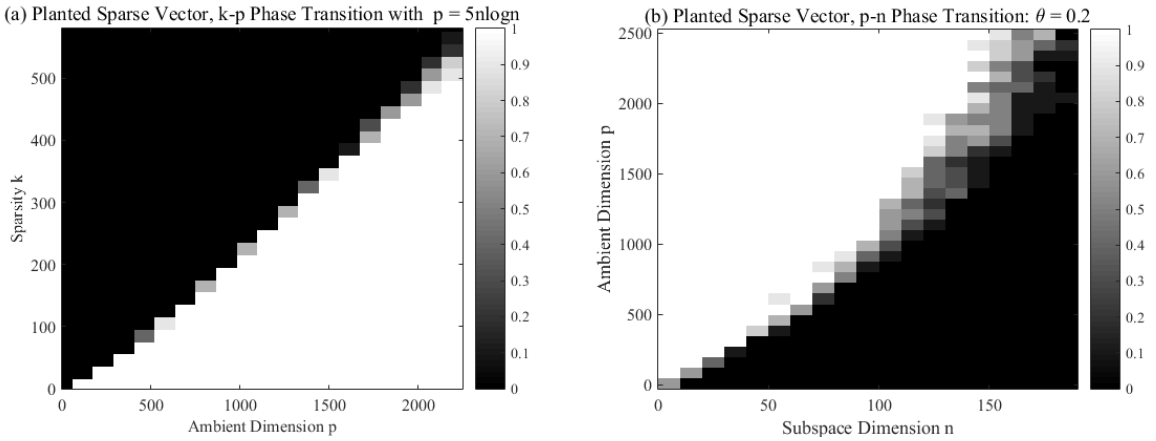


Figure 5.1: Phase transition for the planted sparse model using the ADM algorithm: (a) with fixed relationship between p and n: $p = 5n \log n$; (b) with fixed relationship between p and k: $k = 0.2p$. White indicates success and black indicates failure.

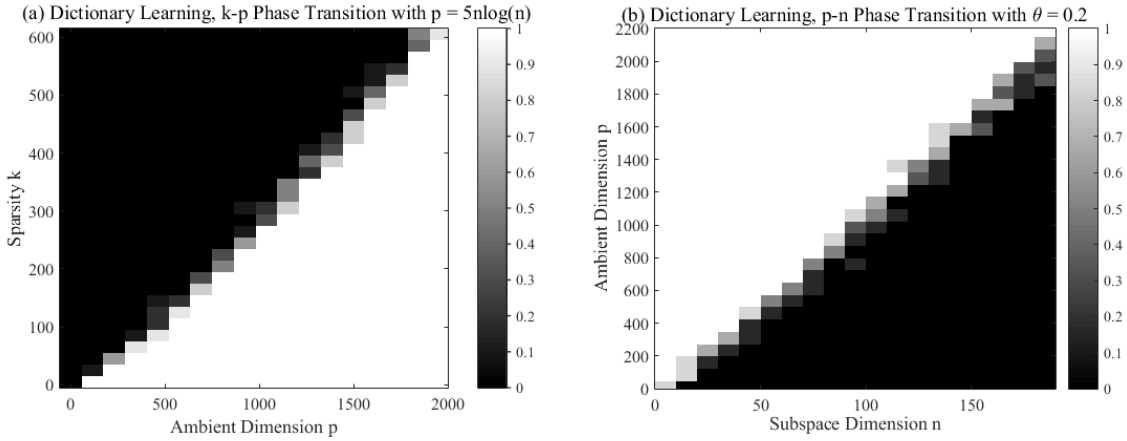


Figure 5.2: Phase transition for the dictionary learning model using the ADM algorithm: (a) with fixed relationship between p and n: $p = 5n \log n$; (b) with fixed relationship between p and k: $k = 0.2p$. White indicates success and black indicates failure.

We also experiment with the complete dictionary learning model as in [SWW12b] (see also [SQW15a]). Specifically, the observation is assumed to be $\boldsymbol{Y} = \boldsymbol{A}_0 \boldsymbol{X}_0$, where $\boldsymbol{A}_0$ is a square, invertible matrix, and $\boldsymbol{X}_0$ a $n \times p$ sparse matrix. Since $\boldsymbol{A}_0$ is invertible, the row space of $\boldsymbol{Y}$ is the same as that of $\boldsymbol{X}_0$. For each pair of (k, n), we generate $\boldsymbol{X}_0 = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]^\top$, where each vector $\boldsymbol{x}_i \in \mathbb{R}^p$ is k-sparse with every nonzero entry following i.i.d. Gaussian distribution, and construct the observation by $\boldsymbol{Y}^\top = \texttt{GS}\left(\boldsymbol{X}_0^\top\right) \boldsymbol{U}^\top$. We repeat the same experiment as for the planted sparse model described above. The only difference is that here we

determine the recovery to be successful as long as one sparse row of $\boldsymbol{X}_0$ is recovered by one of those p programs. Fig. 5.2 shows both phase transition plots.

Fig. 5.1(a) and Fig. 5.2(a) suggest our ADM algorithm could work into the linear sparsity regime for both models, provided $p \geq \Omega(n \log n)$. Moreover, for both models, the $\log n$ factor seems necessary for working into the linear sparsity regime, as suggested by Fig. 5.1(b) and Fig. 5.2(b): there are clear nonlinear transition boundaries between success and failure regions. For both models, $O(n \log n)$ sample requirement is near optimal: for the planted sparse model, obviously $p \geq \Omega(n)$ is necessary; for the complete dictionary learning model, [SWW12b] proved that $p \geq \Omega(n \log n)$ is required for exact recovery. For the planted sparse model, our result $p \geq \Omega(n^4 \log n)$ is far from this much lower empirical requirement. Fig 5.1(b) further suggests that alternative reformulation and algorithm are needed to solve (2.0.1) so that the optimal recovery guarantee as depicted in Theorem 2.1 can be obtained.

5.1.2 Exploratory Experiments on Faces

It is well known in computer vision that the collection of images of a convex object only subject to illumination changes can be well approximated by a low-dimensional subspaces in raw-pixel space [BJ03]. We will play with face subspaces here. First, we extract face images of one person (65 images) under different illumination conditions. Then we apply *robust principal component analysis* [CLMW11a] to the data and get a low dimensional subspace of dimension 10, i.e., the basis $\boldsymbol{Y} \in \mathbb{R}^{32256 \times 10}$. We apply the ADM + LP algorithm to find the sparsest elements in such a subspace, by randomly selecting 10% rows of $\boldsymbol{Y}$ as initializations for $\boldsymbol{q}$. We judge the sparsity in the ℓ^1/ℓ^2 sense, that is, the sparsest vector $\widehat{\boldsymbol{x}}_0 = \boldsymbol{Y}\boldsymbol{q}^\star$ should produce the smallest $\|\boldsymbol{Y}\boldsymbol{q}\|_1 \,/\, \|\boldsymbol{Y}\boldsymbol{q}\|$ among all results. Once some sparse vectors are found, we project the subspace onto orthogonal complement of the sparse vectors already found[1], and continue the seeking process in the projected subspace. Fig. 5.3(Top) shows the first four sparse vectors we get from the data. We can see they correspond well to different extreme illumination conditions. We also implemented the spectral method (with the LP post-processing) proposed in [HSSS15] for comparison under the same protocol. The result is presented as Fig. 5.3(Bottom): the ratios $\|\cdot\|_{\ell^1} \,/\, \|\cdot\|_{\ell^2}$ are significantly higher, and the ratios $\|\cdot\|_{\ell^4} \,/\, \|\cdot\|_{\ell^2}$ (this is the metric to be maximized in [HSSS15] to promote sparsity) are significantly lower. By these two criteria the spectral method with LP rounding consistently produces vectors with higher sparsity levels under our evaluation protocol. Moreover, the resulting images are harder to interpret physically.

[1] The idea is to build a sparse, orthonormal basis for the subspace in a greedy manner.

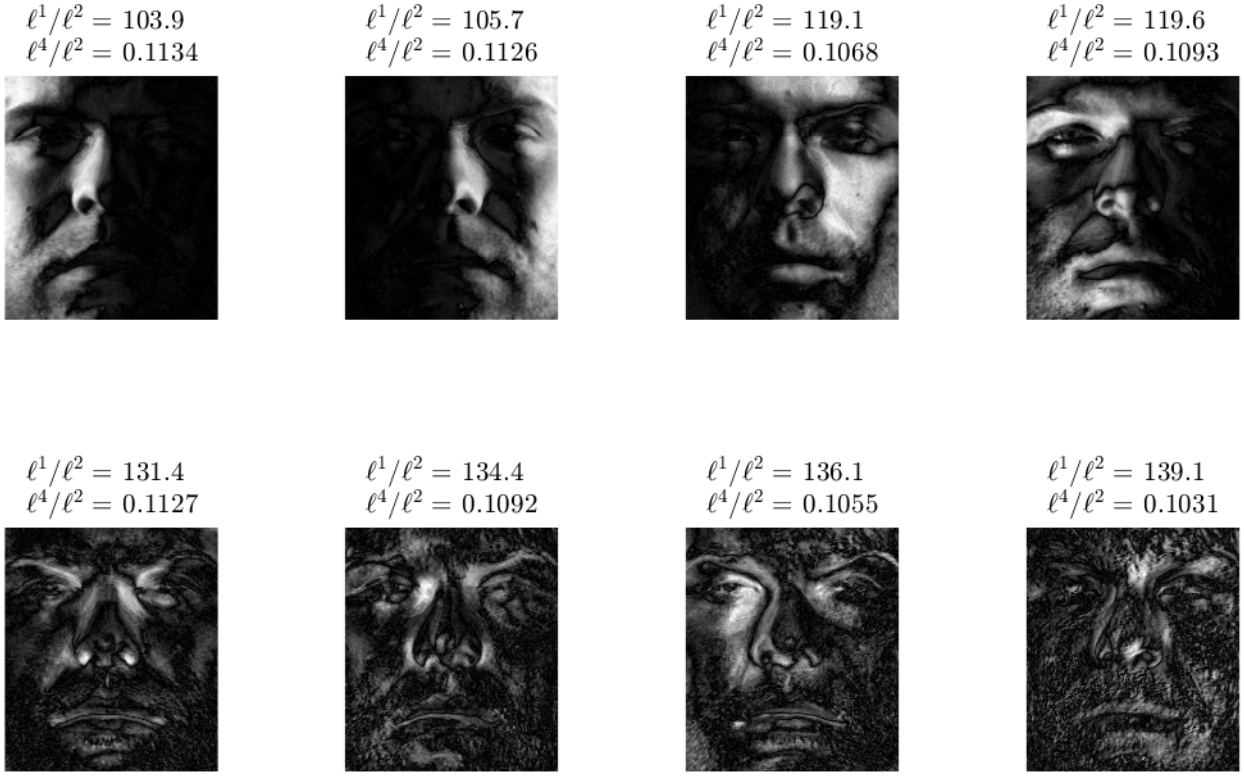


Figure 5.3: The first four sparse vectors extracted for one person in the Yale B database under different illuminations. (Top) by our ADM algorithm; (Bottom) by the speeding-up SOS algorithm proposed in [HSSS15].

Second, we manually select ten different persons' faces under the normal lighting condition. Again, the dimension of the subspace is 10 and $\boldsymbol{Y} \in \mathbb{R}^{32256\times 10}$. We repeat the same experiment as stated above. Fig. 5.4 shows four sparse vectors we get from the data. Interestingly, the sparse vectors roughly correspond to differences of face images concentrated around facial parts that different people tend to differ from each other, e.g., eye brows, forehead hair, nose, etc. By comparison, the vectors returned by the spectral method [HSSS15] are relatively denser and the sparsity patterns in the images are less structured physically.

In sum, our algorithm seems to find useful sparse vectors for potential applications, such as peculiarity discovery in first setting, and locating differences in second setting. Nevertheless, the main goal of this experiment is to invite readers to think about similar pattern discovery problems that might be cast as the problem of seeking sparse vectors in a subspace. The experiment also demonstrates in a concrete way the practicality of our algorithm, both in handling data sets of realistic size and in producing meaningful results even beyond the (idealized) planted sparse model that we adopted for analysis.

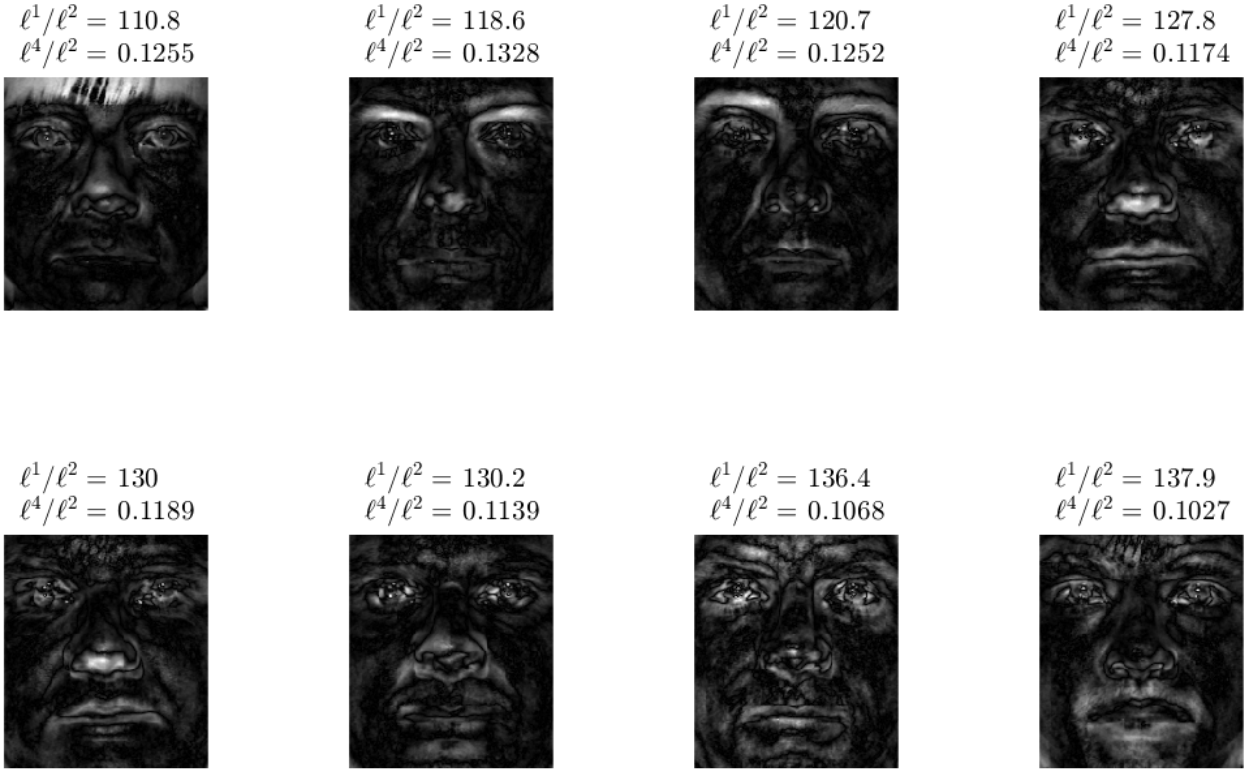


Figure 5.4: The first four sparse vectors extracted for 10 persons in the Yale B database under normal illuminations. (Top) by our ADM algorithm; (Bottom) by the speeding-up SOS algorithm proposed in [HSSS15].

Chapter 6

Discussion

6.1 Connections and Discussion

For the planted sparse model, there is a substantial performance gap in terms of p-n relationship between the our optimality theorem (Theorem 2.1), empirical simulations, and guarantees we have obtained via efficient algorithm (Theorem 4.1). More careful and tighter analysis based on decoupling [DlPG99] and chaining [Tal14b, LV15b] and geometrical analysis described below can probably help bridge the gap between our theoretical and empirical results. Matching the theoretical limit depicted in Theorem 2.1 seems to require novel algorithmic ideas. The random models we assume for the subspace can be extended to other random models, particularly for dictionary learning where all the bases are sparse (e.g., Bernoulli-Gaussian random model).

This work is part of a recent surge of research efforts on deriving provable and practical nonconvex algorithms to central problems in modern signal processing and machine learning. These problems include low-rank matrix recovery/completion [JNS13, Har13, HW14, Har14, JN14, NNS+14, ZL15, TBSR15, CW15], tensor recovery/decomposition [JO14, AGJ14b, AGJ14a, AJSN15, GHJY15], phase retrieval [NJS13, CLS14, CC15, SQW16], dictionary learning [AGM13, AAJ+13, AAN13, ABGM14, AGMM15, SQW15a], and so on.[1] Our approach, like the others, is to start with a carefully chosen, problem-specific initialization, and then perform a local analysis of the subsequent iterates to guarantee convergence to a good solution. In comparison, our subsequent work on complete dictionary learning [SQW15a] and generalized phase retrieval [SQW16]

[1] The webpage http://sunju.org/research/nonconvex/ maintained by the second author contains pointers to the growing list of work in this direction.

has taken a geometrical approach by characterizing the function landscape and designing efficient algorithm accordingly. The geometric approach has allowed provable recovery via efficient algorithms, with an *arbitrary initialization*. The article [SQW15d] summarizes the geometric approach and its applicability to several other problems of interest.

A hybrid of the initialization and the geometric approach discussed above is likely to be a powerful computational framework. To see it in action for the current planted sparse vector problem, in Fig. 6.1 we

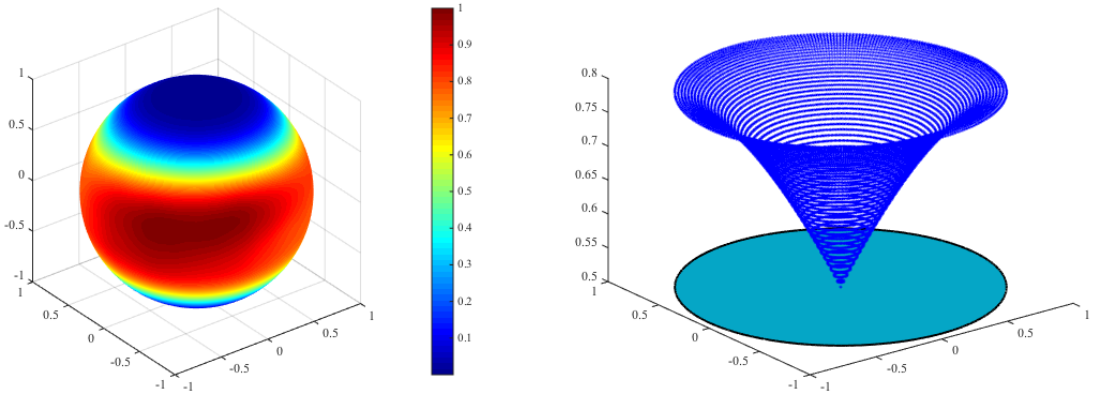


Figure 6.1: Function landscape of $f(\boldsymbol{q})$ with $\theta = 0.4$ for $n = 3$. (Left) $f(\boldsymbol{q})$ over the sphere $\mathbb{S}^2$. Note that near the spherical caps around the north and south poles, there are no critical points and the gradients are always nonzero; (Right) Projected function landscape by projecting the upper hemisphere onto the equatorial plane. Mathematically the function $g(\boldsymbol{w}) : \boldsymbol{e}_3^{\perp} \mapsto \mathbb{R}$ obtained via the reparameterization $\boldsymbol{q}(\boldsymbol{w}) = [\boldsymbol{w}; \sqrt{1 - \|\boldsymbol{w}\|^2}]$. Corresponding to the left, there is no undesired critical point around $\mathbf{0}$ within a large radius.

provide the asymptotic function landscape (i.e., $p \to \infty$) of the Huber loss on the sphere $\mathbb{S}^2$ (aka the relaxed formulation we tried to solve (3.0.1)). It is clear that with an initialization that is biased towards either the north or the south pole, we are situated in a region where the gradients are always nonzero and points to the favorable directions such that many reasonable optimization algorithms can take the gradient information and make steady progress towards the target. This will probably ease the algorithm development and analysis, and help yield tight performance guarantees.

We provide a very efficient algorithm for finding a sparse vector in a subspace, with strong guarantee. Our algorithm is practical for handling large datasets—in the experiment on the face dataset, we successfully extracted some meaningful features from the human face images. However, the potential of seeking sparse/structured element in a subspace seems largely unexplored, despite the cases we mentioned at the start. We hope this work could inspire more application ideas.

Chapter 7

Proof of Technical Results

7.1 Proof of ℓ^1/ℓ^2 Global Optimality

In this appendix, we prove the ℓ^1/ℓ^2 global optimality condition in Theorem 2.1 of Section 2.

Proof [Proof of Theorem 2.1] We will first analyze a canonical version, in which the input orthonormal basis is $\boldsymbol{Y}$ as defined in (3.0.6) of Section 3:

$$\min_{\boldsymbol{q}\in\mathbb{R}^n} \|\boldsymbol{Y}\boldsymbol{q}\|_1, \qquad \text{s.t. } \|\boldsymbol{q}\| = 1.$$

Let $\boldsymbol{q} = \begin{bmatrix} q_1 \\ \boldsymbol{q}_2 \end{bmatrix}$ and let $\mathcal{I}$ be the support set of $\boldsymbol{x}_0$, we have

$$\begin{aligned}
\|\boldsymbol{Y}\boldsymbol{q}\|_1 &= \|\boldsymbol{Y}_{\mathcal{I}}\boldsymbol{q}\|_1 + \|\boldsymbol{Y}_{\mathcal{I}^c}\boldsymbol{q}\|_1 \\
&\geq |q_1| \left\| \frac{\boldsymbol{x}_0}{\|\boldsymbol{x}_0\|} \right\|_1 - \|\boldsymbol{G}'_{\mathcal{I}}\boldsymbol{q}_2\|_1 + \|\boldsymbol{G}'_{\mathcal{I}^c}\boldsymbol{q}_2\|_1 \\
&\geq |q_1| \left\| \frac{\boldsymbol{x}_0}{\|\boldsymbol{x}_0\|} \right\|_1 - \|\boldsymbol{G}_{\mathcal{I}}\boldsymbol{q}_2\|_1 - \|(\boldsymbol{G}_{\mathcal{I}} - \boldsymbol{G}'_{\mathcal{I}})\,\boldsymbol{q}_2\|_1 + \|\boldsymbol{G}_{\mathcal{I}^c}\boldsymbol{q}_2\|_1 - \|(\boldsymbol{G}_{\mathcal{I}^c} - \boldsymbol{G}'_{\mathcal{I}^c})\,\boldsymbol{q}_2\|_1 \\
&\geq |q_1| \left\| \frac{\boldsymbol{x}_0}{\|\boldsymbol{x}_0\|} \right\|_1 - \|\boldsymbol{G}_{\mathcal{I}}\boldsymbol{q}_2\|_1 + \|\boldsymbol{G}_{\mathcal{I}^c}\boldsymbol{q}_2\|_1 - \|\boldsymbol{G} - \boldsymbol{G}'\|_{\ell^2\to\ell^1} \|\boldsymbol{q}_2\|,
\end{aligned}$$

where $\boldsymbol{G}$ and $\boldsymbol{G}'$ are defined in (A.2.1) and (A.2.2) of Appendix A.2. By Lemma A.14 and intersecting with $\mathcal{E}_0$ defined in (A.2.3), we have that as long as $p \geq C_1 n$,

$$\|\boldsymbol{G}_{\mathcal{I}}\boldsymbol{q}_2\|_1 \leq \frac{2\theta p}{\sqrt{p}} \|\boldsymbol{q}_2\| = 2\theta\sqrt{p}\,\|\boldsymbol{q}_2\| \text{ for all } \boldsymbol{q}_2 \in \mathbb{R}^{n-1},$$

$$\|\boldsymbol{G}_{\mathcal{I}^c}\boldsymbol{q}_2\|_1 \geq \frac{1}{2}\frac{p-2\theta p}{\sqrt{p}}\|\boldsymbol{q}_2\| = \frac{1}{2}\sqrt{p}\,(1-2\theta)\,\|\boldsymbol{q}_2\| \text{ for all } \boldsymbol{q}_2 \in \mathbb{R}^{n-1},$$

hold with probability at least $1-c_2p^{-2}$. Moreover, by Lemma A.17,

$$\|\boldsymbol{G}-\boldsymbol{G}'\|_{\ell^2\to\ell^1} \leq 4\sqrt{n}+7\sqrt{\log(2p)}$$

holds with probability at least $1-c_3p^{-2}$ when $p \geq C_4 n$ and $\theta > 1/\sqrt{n}$. So we obtain that

$$\|\boldsymbol{Y}\boldsymbol{q}\|_1 \geq g(\boldsymbol{q}) \doteq |q_1|\left\|\frac{\boldsymbol{x}_0}{\|\boldsymbol{x}_0\|}\right\|_1 + \|\boldsymbol{q}_2\|\left(\frac{1}{2}\sqrt{p}\,(1-2\theta)-2\theta\sqrt{p}-4\sqrt{n}-7\sqrt{\log(2p)}\right)$$

holds with probability at least $1-c_5p^{-2}$. Assuming $\mathcal{E}_0$, we observe

$$\left\|\frac{\boldsymbol{x}_0}{\|\boldsymbol{x}_0\|}\right\|_1 \leq \sqrt{|\mathcal{I}|}\left\|\frac{\boldsymbol{x}_0}{\|\boldsymbol{x}_0\|}\right\| \leq \sqrt{2\theta p}.$$

Now $g(\boldsymbol{q})$ is a linear function in $|q_1|$ and $\|\boldsymbol{q}_2\|$. Thus, whenever θ is sufficiently small and $p \geq C_6 n$ such that

$$\sqrt{2\theta p} < \frac{1}{2}\sqrt{p}\,(1-2\theta)-2\theta\sqrt{p}-4\sqrt{n}-7\sqrt{\log(2p)},$$

$\pm\boldsymbol{e}_1$ are the unique minimizers of $g(\boldsymbol{q})$ under the constraint $q_1^2+\|\boldsymbol{q}_2\|^2=1$. In this case, because $\|\boldsymbol{Y}(\pm\boldsymbol{e}_1)\|_1 = g(\pm\boldsymbol{e}_1)$, and we have

$$\|Yq\|_1 \geq g(\boldsymbol{q}) > g(\pm\boldsymbol{e}_1)$$

for all $\boldsymbol{q} \neq \pm\boldsymbol{e}_1$, $\pm\boldsymbol{e}_1$ are the unique minimizers of $\|\boldsymbol{Y}q\|_1$ under the spherical constraint. Thus there exists a universal constant $\theta_0 > 0$, such that for all $1/\sqrt{n} \leq \theta \leq \theta_0$, $\pm\boldsymbol{e}_1$ are the only global minimizers of (2.0.2) if the input basis is $\boldsymbol{Y}$.

Any other input basis can be written as $\widehat{\boldsymbol{Y}} = \boldsymbol{Y}\boldsymbol{U}$, for some orthogonal matrix $\boldsymbol{U}$. The program now is written as

$$\min_{\boldsymbol{q}\in\mathbb{R}^n}\left\|\widehat{\boldsymbol{Y}}\boldsymbol{q}\right\|_1, \qquad \text{s.t. } \|\boldsymbol{q}\|=1,$$

which is equivalent to

$$\min_{\boldsymbol{q}\in\mathbb{R}^n}\left\|\widehat{\boldsymbol{Y}}\boldsymbol{q}\right\|_1, \qquad \text{s.t. } \|\boldsymbol{U}\boldsymbol{q}\|=1,$$

which is obviously equivalent to the canonical program we analyzed above by a simple change of variable, i.e., $\overline{\boldsymbol{q}} \doteq \boldsymbol{U}\boldsymbol{q}$, completing the proof. ■

7.2 Good Initialization

In this appendix, we prove Proposition 4.3. We show that the initializations produced by the procedure described in Section 3 are biased towards the optimal.

Proof [Proof of Proposition 4.3] Our previous calculation has shown that $\theta p/2 \leq |\mathcal{I}| \leq 2\theta p$ with probability at least $1 - c_1 p^{-2}$ provided $p \geq C_2 n$ and $\theta > 1/\sqrt{n}$. Let $\boldsymbol{Y} = \left[\boldsymbol{y}^1, \cdots, \boldsymbol{y}^p\right]^\top$ as defined in (3.0.6). Consider any $i \in \mathcal{I}$. Then $x_0(i) = \frac{1}{\sqrt{\theta p}}$, and

$$\begin{aligned}\left\langle \boldsymbol{e}_1, \boldsymbol{y}^i / \left\|\boldsymbol{y}^i\right\| \right\rangle = \frac{1/\sqrt{\theta p}}{\|\boldsymbol{x}_0\| \left\|\boldsymbol{y}^i\right\|} &\geq \frac{1/\sqrt{\theta p}}{\|\boldsymbol{x}_0\| \left(\|\boldsymbol{x}_0\|_\infty / \|\boldsymbol{x}_0\| + \|(\boldsymbol{g}')^i\|\right)} \\ &\geq \frac{1/\sqrt{\theta p}}{\|\boldsymbol{x}_0\| \left(\|\boldsymbol{x}_0\|_\infty / \|\boldsymbol{x}_0\| + \|\boldsymbol{g}^i\| + \|\boldsymbol{G} - \boldsymbol{G}'\|_{\ell^2 \to \ell^\infty}\right)},\end{aligned}$$

where $\boldsymbol{g}^i$ and $(\boldsymbol{g}')^i$ are the i-th rows of $\boldsymbol{G}$ and $\boldsymbol{G}'$, respectively. Since such $\boldsymbol{g}^i$'s are independent Gaussian vectors in $\mathbb{R}^{n-1}$ distributed as $\mathcal{N}(\boldsymbol{0}, 1/p)$, by Gaussian concentration inequality and the fact that $|\mathcal{I}| \geq p\theta/2$ w.h.p.,

$$\mathbb{P}\exists i \in \mathcal{I} : \left\|\boldsymbol{g}^i\right\| \leq 2\sqrt{n/p} \geq 1 - \exp\left(-c_3 n\theta p\right) \leq c_4 p^{-2},$$

provided $p \geq C_5 n$ and $\theta > 1/\sqrt{n}$. Moreover,

$$\|\boldsymbol{x}_0\| = \sqrt{|\mathcal{I}| \times \frac{1}{\theta p}} \leq \sqrt{2\theta p \times \frac{1}{\theta p}} = \sqrt{2}.$$

Combining the above estimates and result of Lemma A.18, we obtain that provided $p \geq C_6 n$ and $\theta > 1/\sqrt{n}$, with probability at least $1 - c_7 p^{-2}$, there exists an $i \in [p]$, such that if we set $\boldsymbol{q}^{(0)} = \boldsymbol{y}^i / \left\|\boldsymbol{y}^i\right\|$, it holds that

$$\begin{aligned}\left|q_1^{(0)}\right| &\geq \frac{1/\sqrt{\theta p}}{1/\sqrt{\theta p} + 2\sqrt{2}\sqrt{n/p} + \sqrt{2}\left(4n/p + 8\sqrt{2}\log(2p)/p + 21\sqrt{n\log(2p)}/p\right)} \\ &\geq \frac{1/\sqrt{\theta p}}{1/\sqrt{\theta p} + 6\sqrt{2}\sqrt{n/p}} \quad \text{(using } p \geq C_6 n \text{ to simplifiy the above line)} \\ &= \frac{1}{1 + 6\sqrt{2}\sqrt{\theta n}} \\ &\geq \frac{1}{(1 + 6\sqrt{2})\sqrt{\theta n}} \quad \text{(as } \theta > 1/\sqrt{n}\text{)} \\ &\geq \frac{1}{10\sqrt{\theta n}},\end{aligned}$$

completing the proof. ∎

We will next show that for an arbitrary orthonormal basis $\widehat{\boldsymbol{Y}} \doteq \boldsymbol{Y}\boldsymbol{U}$ the initialization still biases towards

the target solution. To see this, suppose w.l.o.g. $\left(\boldsymbol{y}^i\right)^\top$ is a row of $\boldsymbol{Y}$ with nonzero first coordinate. We have shown above that with high probability $\left|\left\langle \frac{\boldsymbol{y}^i}{\|\boldsymbol{y}^i\|}, \boldsymbol{e}_1 \right\rangle\right| \geq \frac{1}{10\sqrt{\theta n}}$ if $\boldsymbol{Y}$ is the input orthonormal basis. For $\boldsymbol{Y}$, as $\boldsymbol{x}_0 = \boldsymbol{Y}\boldsymbol{e}_1 = \boldsymbol{Y}\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{e}_1$, we know $\boldsymbol{q}_\star = \boldsymbol{U}^\top\boldsymbol{e}_1$ is the target solution corresponding to $\widehat{\boldsymbol{Y}}$. Observing that

$$\left|\left\langle \boldsymbol{U}^\top\boldsymbol{e}_1, \frac{\left(\boldsymbol{e}_i^\top\widehat{\boldsymbol{Y}}\right)^\top}{\left\|\left(\boldsymbol{e}_i^\top\widehat{\boldsymbol{Y}}\right)^\top\right\|} \right\rangle\right| = \left|\left\langle \boldsymbol{U}^\top\boldsymbol{e}_1, \frac{\boldsymbol{U}^\top\boldsymbol{Y}^\top\boldsymbol{e}_i}{\|\boldsymbol{U}^\top\boldsymbol{Y}^\top\boldsymbol{e}_i\|} \right\rangle\right| = \left|\left\langle \boldsymbol{e}_1, \frac{(\boldsymbol{Y})^\top\boldsymbol{e}_i}{\|\boldsymbol{Y}^\top\boldsymbol{e}_i\|} \right\rangle\right| = \left|\left\langle \boldsymbol{e}_1, \frac{\boldsymbol{y}^i}{\|\boldsymbol{y}^i\|} \right\rangle\right| \geq \frac{1}{10\sqrt{n\theta}},$$

corroborating our claim.

7.3 Lower Bounding Finite Sample Gap $G(q)$

In this Section, we prove Proposition 4.4. In particular, we show that the gap $\boldsymbol{G}(\boldsymbol{q})$ defined in (4.2.6) is strictly positive over a large portion of the sphere $\mathbb{S}^{n-1}$.

Proof [Proof of Proposition 4.4] Without loss of generality, we work with the "canonical" orthonormal basis $\boldsymbol{Y}$ defined in (3.0.6). Recall that $\boldsymbol{Y}$ is the orthogonalization of the planted sparse basis $\overline{\boldsymbol{Y}}$ as defined in (3.0.5). We define the processes $\overline{\boldsymbol{Q}}(\boldsymbol{q})$ and $\boldsymbol{Q}(\boldsymbol{q})$ on $\boldsymbol{q} \in \mathbb{S}^{n-1}$, via

$$\overline{\boldsymbol{Q}}(\boldsymbol{q}) = \frac{1}{p}\sum_{i=1}^{p} \overline{\boldsymbol{y}}^i S_\lambda\left[\boldsymbol{q}^\top\overline{\boldsymbol{y}}^i\right], \quad \boldsymbol{Q}(\boldsymbol{q}) = \frac{1}{p}\sum_{i=1}^{p} \boldsymbol{y}^i S_\lambda\left[\boldsymbol{q}^\top\boldsymbol{y}^i\right].$$

Thus, we can separate $\overline{\boldsymbol{Q}}(\boldsymbol{q})$ as $\overline{\boldsymbol{Q}}(\boldsymbol{q}) = \begin{bmatrix} \overline{Q}_1(\boldsymbol{q}) \\ \overline{\boldsymbol{Q}}_2(\boldsymbol{q}) \end{bmatrix}$, where

$$\overline{Q}_1(\boldsymbol{q}) = \frac{1}{p}\sum_{i=1}^{p} x_{0i} S_\lambda\left[\boldsymbol{q}^\top\overline{\boldsymbol{y}}^i\right] \quad \text{and} \quad \overline{\boldsymbol{Q}}_2(\boldsymbol{q}) = \frac{1}{p}\sum_{i=1}^{p} \boldsymbol{g}_i S_\lambda\left[\boldsymbol{q}^\top\overline{\boldsymbol{y}}^i\right], \tag{7.3.1}$$

and separate $\boldsymbol{Q}(\boldsymbol{q})$ correspondingly. Our task is to lower bound the gap $G(\boldsymbol{q})$ for finite samples as defined in (4.2.6). Since we can deterministically constrain $|q_1|$ and $\|\boldsymbol{q}_2\|$ over the set Γ as defined in (4.2.5) (e.g., $\frac{1}{10\sqrt{n\theta}} \leq |q_1| \leq 3\sqrt{\theta}$ and $\|\boldsymbol{q}_2\| \geq \frac{1}{10}$, where the choice of $\frac{1}{10}$ for $\boldsymbol{q}_2$ is arbitrary here, as we can always take a sufficiently small θ), the challenge lies in lower bounding $|Q_1(\boldsymbol{q})|$ and upper bounding $\|\boldsymbol{Q}_2(\boldsymbol{q})\|$, which depend on the orthonormal basis $\boldsymbol{Y}$. The unnormalized basis $\overline{\boldsymbol{Y}}$ is much easier to work with than $\boldsymbol{Y}$. Our proof will follow the observation that

$$\begin{aligned} |Q_1(\boldsymbol{q})| &\geq \left|\mathbb{E}\overline{Q}_1(\boldsymbol{q})\right| - \left|\overline{Q}_1(\boldsymbol{q}) - \mathbb{E}\overline{Q}_1(\boldsymbol{q})\right| - \left|Q_1(\boldsymbol{q}) - \overline{Q}_1(\boldsymbol{q})\right|, \\ \|\boldsymbol{Q}_2(\boldsymbol{q})\| &\leq \left\|\mathbb{E}\overline{\boldsymbol{Q}}_2(\boldsymbol{q})\right\| + \left\|\overline{\boldsymbol{Q}}_2(\boldsymbol{q}) - \mathbb{E}\overline{\boldsymbol{Q}}_2(\boldsymbol{q})\right\| + \left\|\boldsymbol{Q}_2(\boldsymbol{q}) - \overline{\boldsymbol{Q}}_2(\boldsymbol{q})\right\|. \end{aligned}$$

In particular, we show the following:

- Section 7.3.1 shows that the expected gap is lower bounded for all $\boldsymbol{q} \in \mathbb{S}^{n-1}$ with $|q_1| \leq 3\sqrt{\theta}$:

$$\overline{G}(\boldsymbol{q}) \doteq \frac{\left|\mathbb{E}\overline{Q}_1(\boldsymbol{q})\right|}{|q_1|} - \frac{\left\|\mathbb{E}\overline{\boldsymbol{Q}}_2(\boldsymbol{q})\right\|}{\|\boldsymbol{q}_2\|} \geq \frac{1}{50}\frac{q_1^2}{\theta p}.$$

  As $|q_1| \geq \frac{1}{10\sqrt{n\theta}}$, we have

$$\inf_{\boldsymbol{q}\in\Gamma} \frac{\left|\mathbb{E}\overline{Q}_1(\boldsymbol{q})\right|}{|q_1|} - \frac{\left\|\mathbb{E}\overline{\boldsymbol{Q}}_2(\boldsymbol{q})\right\|}{\|\boldsymbol{q}_2\|} \geq \frac{1}{5000}\frac{1}{\theta^2 np}.$$

- Section 7.3.2, as summarized in Proposition 7.8, shows that whenever $p \geq \Omega\left(n^4 \log n\right)$, it holds with high probability that

$$\begin{aligned}&\sup_{\boldsymbol{q}\in\Gamma} \frac{\left|\overline{Q}_1(\boldsymbol{q}) - \mathbb{E}\overline{Q}_1(\boldsymbol{q})\right|}{|q_1|} + \frac{\left\|\overline{\boldsymbol{Q}}_2(\boldsymbol{q}) - \mathbb{E}\overline{\boldsymbol{Q}}_2(\boldsymbol{q})\right\|}{\|\boldsymbol{q}_2\|} \\ \leq &\frac{10\sqrt{\theta n}}{4\times 10^5 \theta^{5/2} n^{3/2} p} + \frac{10}{4\times 10^5 \theta^2 np} = \frac{1}{2\times 10^4 \theta^2 np}.\end{aligned}$$

- Section 7.3.4 shows that whenever $p \geq \Omega\left(n^4 \log n\right)$, it holds with high probability that

$$\begin{aligned}&\sup_{\boldsymbol{q}\in\Gamma} \frac{\left|\overline{Q}_1(\boldsymbol{q}) - Q_1(\boldsymbol{q})\right|}{|q_1|} + \frac{\left\|\overline{\boldsymbol{Q}}_2(\boldsymbol{q}) - \boldsymbol{Q}_2(\boldsymbol{q})\right\|}{\|\boldsymbol{q}_2\|} \\ \leq &\frac{10\sqrt{\theta n}}{4\times 10^5 \theta^{5/2} n^{3/2} p} + \frac{10}{4\times 10^5 \theta^2 np} = \frac{1}{2\times 10^4 \theta^2 np}.\end{aligned}$$

Observing that

$$\begin{aligned}\inf_{\boldsymbol{q}\in\Gamma} G(\boldsymbol{q}) \geq \inf_{\boldsymbol{q}\in\Gamma} &\left(\frac{\left|\mathbb{E}\overline{Q}_1(\boldsymbol{q})\right|}{|q_1|} - \frac{\left\|\mathbb{E}\overline{\boldsymbol{Q}}_2(\boldsymbol{q})\right\|}{\|\boldsymbol{q}_2\|}\right) - \sup_{\boldsymbol{q}\in\Gamma}\left(\frac{\left|\overline{Q}_1(\boldsymbol{q}) - \mathbb{E}\overline{Q}_1(\boldsymbol{q})\right|}{|q_1|} + \frac{\left\|\overline{\boldsymbol{Q}}_2(\boldsymbol{q}) - \mathbb{E}\overline{\boldsymbol{Q}}_2(\boldsymbol{q})\right\|}{\|\boldsymbol{q}_2\|}\right) \\ &- \sup_{\boldsymbol{q}\in\Gamma}\left(\frac{\left|\overline{Q}_1(\boldsymbol{q}) - Q_1(\boldsymbol{q})\right|}{|q_1|} + \frac{\left\|\overline{\boldsymbol{Q}}_2(\boldsymbol{q}) - \boldsymbol{Q}_2(\boldsymbol{q})\right\|}{\|\boldsymbol{q}_2\|}\right),\end{aligned}$$

we obtain the result as desired. ∎

For the general case when the input orthonormal basis is $\widehat{\boldsymbol{Y}} = \boldsymbol{Y}\boldsymbol{U}$ with target solution $\boldsymbol{q}_\star = \boldsymbol{U}^\top \boldsymbol{e}_1$, a straightforward extension of the definition for the gap would be:

$$G\left(\boldsymbol{q}; \widehat{\boldsymbol{Y}} = \boldsymbol{Y}\boldsymbol{U}\right) \doteq \frac{\left|\left\langle \boldsymbol{Q}\left(\boldsymbol{q}; \widehat{\boldsymbol{Y}}\right), \boldsymbol{U}^\top \boldsymbol{e}_1\right\rangle\right|}{|\langle \boldsymbol{q}, \boldsymbol{U}^\top \boldsymbol{e}_1\rangle|} - \frac{\left\|\left(\boldsymbol{I} - \boldsymbol{U}^\top \boldsymbol{e}_1 \boldsymbol{e}_1^\top \boldsymbol{U}\right) \boldsymbol{Q}\left(\boldsymbol{q}; \widehat{\boldsymbol{Y}}\right)\right\|}{\left\|\left(\boldsymbol{I} - \boldsymbol{U}^\top \boldsymbol{e}_1 \boldsymbol{e}_1^\top \boldsymbol{U}\right) \boldsymbol{q}\right\|}.$$

Since $\boldsymbol{Q}\left(\boldsymbol{q};\widehat{\boldsymbol{Y}}\right)=\frac{1}{p}\sum_{k=1}^{p}\boldsymbol{U}^{\top}\boldsymbol{y}^{k}S_{\lambda}\left(\boldsymbol{q}^{\top}\boldsymbol{U}^{\top}\boldsymbol{y}^{k}\right)$, we have

$$\boldsymbol{U}\boldsymbol{Q}\left(\boldsymbol{q};\widehat{\boldsymbol{Y}}\right)=\frac{1}{p}\sum_{k=1}^{p}\boldsymbol{U}\boldsymbol{U}^{\top}\boldsymbol{y}^{k}S_{\lambda}\left(\boldsymbol{q}^{\top}\boldsymbol{U}^{\top}\boldsymbol{y}^{k}\right)=\frac{1}{p}\sum_{k=1}^{p}\boldsymbol{y}^{k}S_{\lambda}\left[\left(\boldsymbol{U}\boldsymbol{q}\right)^{\top}\boldsymbol{y}^{k}\right]=\boldsymbol{Q}\left(\boldsymbol{U}\boldsymbol{q};\boldsymbol{Y}\right). \tag{7.3.2}$$

Hence we have

$$G\left(\boldsymbol{q};\widehat{\boldsymbol{Y}}=\boldsymbol{Y}\boldsymbol{U}\right)=\frac{\left|\left\langle\boldsymbol{Q}\left(\boldsymbol{U}\boldsymbol{q};\boldsymbol{Y}\right),\boldsymbol{e}_{1}\right\rangle\right|}{\left|\left\langle\boldsymbol{U}\boldsymbol{q},\boldsymbol{e}_{1}\right\rangle\right|}-\frac{\left\|\left(\boldsymbol{I}-\boldsymbol{e}_{1}\boldsymbol{e}_{1}^{\top}\right)\boldsymbol{Q}\left(\boldsymbol{U}\boldsymbol{q};\boldsymbol{Y}\right)\right\|}{\left\|\left(\boldsymbol{I}-\boldsymbol{e}_{1}\boldsymbol{e}_{1}^{\top}\right)\boldsymbol{U}\boldsymbol{q}\right\|}.$$

Therefore, from Proposition 4.4 above, we conclude that under the same technical conditions as therein,

$$\inf_{\boldsymbol{q}\in\mathbb{S}^{n-1}:\frac{1}{10\sqrt{\theta}n}\leq\left|\left\langle\boldsymbol{U}\boldsymbol{q},\boldsymbol{e}_{1}\right\rangle\right|\leq3\sqrt{\theta}}G\left(\boldsymbol{q};\widehat{\boldsymbol{Y}}\right)\geq\frac{1}{10^{4}\theta^{2}np}$$

with high probability.

7.3.1 Lower Bounding the Expected Gap $\overline{G}(q)$

In this section, we provide a nontrivial lower bound for the gap

$$\overline{G}(\boldsymbol{q})=\frac{\left|\mathbb{E}\left[\overline{Q}_{1}(\boldsymbol{q})\right]\right|}{\left|q_{1}\right|}-\frac{\left\|\mathbb{E}\left[\overline{\boldsymbol{Q}}_{2}(\boldsymbol{q})\right]\right\|}{\left\|\boldsymbol{q}_{2}\right\|}. \tag{7.3.3}$$

More specifically, we show that:

Proposition 7.1 *There exists some numerical constant $\theta_0>0$, such that for all $\theta\in(0,\theta_0)$, it holds that*

$$\overline{G}(\boldsymbol{q})\geq\frac{1}{50}\frac{q_{1}^{2}}{\theta p} \tag{7.3.4}$$

for all $\boldsymbol{q}\in\mathbb{S}^{n-1}$ with $|q_1|\leq3\sqrt{\theta}$.

Estimating the gap $\overline{G}(\boldsymbol{q})$ requires delicate estimates for $\mathbb{E}\left[\overline{Q}_{1}(\boldsymbol{q})\right]$ and $\mathbb{E}\left[\overline{\boldsymbol{Q}}_{2}(\boldsymbol{q})\right]$. We first outline the main proof in Section 7.3.1.1, and delay these detailed technical calculations to the subsequent subsections.

7.3.1.1 Sketch of the Proof

W.l.o.g., we only consider the situation that $q_1>0$, because the case of $q_1<0$ can be similarly shown by symmetry. By (7.3.1), we have

$$\begin{aligned}\mathbb{E}\left[\overline{Q}_{1}(\boldsymbol{q})\right]&=\mathbb{E}\left[x_{0}S_{\lambda}\left[x_{0}q_{1}+\boldsymbol{q}_{2}^{\top}\boldsymbol{g}\right]\right],\\ \mathbb{E}\left[\overline{\boldsymbol{Q}}_{2}(\boldsymbol{q})\right]&=\mathbb{E}\left[\boldsymbol{g}S_{\lambda}\left[x_{0}q_{1}+\boldsymbol{q}_{2}^{\top}\boldsymbol{g}\right]\right],\end{aligned}$$

where $\boldsymbol{g} \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{p}\boldsymbol{I}\right)$, and $x_0 \sim \frac{1}{\sqrt{\theta p}}\text{Ber}(\theta)$. Let us decompose

$$\boldsymbol{g} = \boldsymbol{g}_{\|} + \boldsymbol{g}_{\perp},$$

with $\boldsymbol{g}_{\|} = \mathcal{P}_{\|}\boldsymbol{g} = \frac{\boldsymbol{q}_2\boldsymbol{q}_2^\top}{\|\boldsymbol{q}_2\|^2}\boldsymbol{g}$, and $\boldsymbol{g}_{\perp} = (\boldsymbol{I} - \mathcal{P}_{\|})\boldsymbol{g}$. In this notation, we have

$$\begin{aligned}
\mathbb{E}\left[\overline{\boldsymbol{Q}}_2(\boldsymbol{q})\right] &= \mathbb{E}\left[\boldsymbol{g}_{\|}S_\lambda\left[x_0 q_1 + \boldsymbol{q}_2^\top \boldsymbol{g}_{\|}\right]\right] + \mathbb{E}\left[\boldsymbol{g}_{\perp}S_\lambda\left[x_0 q_1 + \boldsymbol{q}_2^\top \boldsymbol{g}_{\|}\right]\right] \\
&= \mathbb{E}\left[\boldsymbol{g}_{\|}S_\lambda\left[x_0 q_1 + \boldsymbol{q}_2^\top \boldsymbol{g}\right]\right] + \mathbb{E}\left[\boldsymbol{g}_{\perp}\right]\mathbb{E}\left[S_\lambda\left[x_0 q_1 + \boldsymbol{q}_2^\top \boldsymbol{g}\right]\right] \\
&= \frac{\boldsymbol{q}_2}{\|\boldsymbol{q}_2\|^2}\mathbb{E}\left[\boldsymbol{q}_2^\top \boldsymbol{g} S_\lambda\left[x_0 q_1 + \boldsymbol{q}_2^\top \boldsymbol{g}\right]\right],
\end{aligned}$$

where we used the facts that $\boldsymbol{q}_2^\top \boldsymbol{g} = \boldsymbol{q}_2^\top \boldsymbol{g}_{\|}$, $\boldsymbol{g}_{\perp}$ and $\boldsymbol{g}_{\|}$ are uncorrelated Gaussian vectors and therefore independent, and $\mathbb{E}\left[\boldsymbol{g}_{\perp}\right] = \mathbf{0}$. Let $Z \doteq \boldsymbol{g}^\top \boldsymbol{q}_2 \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = \|\boldsymbol{q}_2\|^2/p$, by partial evaluation of the expectations with respect to x_0, we get

$$\mathbb{E}\left[\overline{Q}_1(\boldsymbol{q})\right] = \sqrt{\frac{\theta}{p}}\mathbb{E}\left[S_\lambda\left[\frac{q_1}{\sqrt{\theta p}} + Z\right]\right], \tag{7.3.5}$$

$$\mathbb{E}\left[\overline{\boldsymbol{Q}}_2(\boldsymbol{q})\right] = \frac{\theta \boldsymbol{q}_2}{\|\boldsymbol{q}_2\|^2}\mathbb{E}\left[Z S_\lambda\left[\frac{q_1}{\sqrt{\theta p}} + Z\right]\right] + \frac{(1-\theta)\boldsymbol{q}_2}{\|\boldsymbol{q}_2\|^2}\mathbb{E}\left[Z S_\lambda\left[Z\right]\right]. \tag{7.3.6}$$

Straightforward integration based on Lemma A.1 gives a explicit form of the expectations as follows

$$\mathbb{E}\left[\overline{Q}_1(\boldsymbol{q})\right] = \sqrt{\frac{\theta}{p}}\left\{\left[\alpha\Psi\left(-\frac{\alpha}{\sigma}\right) + \beta\Psi\left(\frac{\beta}{\sigma}\right)\right] + \sigma\left[\psi\left(-\frac{\beta}{\sigma}\right) - \psi\left(-\frac{\alpha}{\sigma}\right)\right]\right\}, \tag{7.3.7}$$

$$\mathbb{E}\left[\overline{\boldsymbol{Q}}_2(\boldsymbol{q})\right] = \left\{\frac{2(1-\theta)}{p}\Psi\left(-\frac{\lambda}{\sigma}\right) + \frac{\theta}{p}\left[\Psi\left(-\frac{\alpha}{\sigma}\right) + \Psi\left(\frac{\beta}{\sigma}\right)\right]\right\} q_2, \tag{7.3.8}$$

where the scalars α and β are defined as

$$\alpha = \frac{q_1}{\sqrt{\theta p}} + \lambda, \qquad \beta = \frac{q_1}{\sqrt{\theta p}} - \lambda,$$

and $\psi(t)$ and $\Psi(t)$ are *pdf* and *cdf* for standard normal distribution, respectively, as defined in Lemma A.1. Plugging (7.3.7) and (7.3.8) into (7.3.3), by some simplifications, we obtain

$$\begin{aligned}
\overline{G}(\boldsymbol{q}) = &\frac{1}{q_1}\sqrt{\frac{\theta}{p}}\left[\alpha\Psi\left(-\frac{\alpha}{\sigma}\right) + \beta\Psi\left(\frac{\beta}{\sigma}\right) - \frac{2q_1}{\sqrt{\theta p}}\Psi\left(-\frac{\lambda}{\sigma}\right)\right] - \frac{\theta}{p}\left[\Psi\left(-\frac{\alpha}{\sigma}\right) + \Psi\left(\frac{\beta}{\sigma}\right) - 2\Psi\left(-\frac{\lambda}{\sigma}\right)\right] \\
&+ \frac{\sigma}{q_1}\sqrt{\frac{\theta}{p}}\left[\psi\left(\frac{\beta}{\sigma}\right) - \psi\left(-\frac{\alpha}{\sigma}\right)\right].
\end{aligned} \tag{7.3.9}$$

With $\lambda = 1/\sqrt{p}$ and $\sigma^2 = \|\boldsymbol{q}_2\|^2/p = (1-q_1^2)/p$, we have

$$-\frac{\alpha}{\sigma} = -\frac{\delta+1}{\sqrt{1-q_1^2}}, \qquad \frac{\beta}{\sigma} = \frac{\delta-1}{\sqrt{1-q_1^2}}, \qquad \frac{\lambda}{\sigma} = \frac{1}{\sqrt{1-q_1^2}},$$

where $\delta = q_1/\sqrt{\theta}$ for $q_1 \leq 3\sqrt{\theta}$. To proceed, it is natural to consider estimating the gap $\overline{G}(\boldsymbol{q})$ by Taylor's expansion. More specifically, we approximate $\Psi\left(-\frac{\alpha}{\sigma}\right)$ and $\psi\left(-\frac{\alpha}{\sigma}\right)$ around $-1-\delta$, and approximate $\Psi\left(\frac{\beta}{\sigma}\right)$ and $\psi\left(\frac{\beta}{\sigma}\right)$ around $-1+\delta$. Applying the estimates for the relevant quantities established in Lemma 7.2, we obtain

$$\begin{aligned}\overline{G}(\boldsymbol{q}) \geq\ & \frac{1-\theta}{p}\Phi_1(\delta) - \frac{1}{\delta p}\Phi_2(\delta) + \frac{1-\theta}{p}\psi(-1)q_1^2 + \frac{1}{p}\left(\sigma\sqrt{p} + \frac{\theta}{2} - 1\right)\eta_2(\delta)q_1^2 \\ & + \frac{1}{2\delta p}\left[1+\delta^2-\theta\delta^2-\sigma\left(1+\delta^2\right)\sqrt{p}\right]q_1^2\eta_1(\delta) + \frac{\sigma}{\delta\sqrt{p}}\eta_1(\delta) - \frac{5C_T\sqrt{\theta}q_1^3}{p}(\delta+1)^3,\end{aligned}$$

where we define

$$\begin{aligned}\Phi_1(\delta) &= \Psi(-1-\delta) + \Psi(-1+\delta) - 2\Psi(-1), & \Phi_2(\delta) &= \Psi(-1+\delta) - \Psi(-1-\delta), \\ \eta_1(\delta) &= \psi(-1+\delta) - \psi(-1-\delta), & \eta_2(\delta) &= \psi(-1+\delta) + \psi(-1-\delta),\end{aligned}$$

and C_T is as defined in Lemma 7.2. Since $1-\sigma\sqrt{p} \geq 0$, dropping those small positive terms $\frac{q_1^2}{p}(1-\theta)\psi(-1)$, $\frac{\theta q_1^2}{2p}\eta_2(\delta)$, and $\left(1+\delta^2\right)\left(1-\sigma\sqrt{p}\right)q_1^2\eta_1(\delta)/(2\delta p)$, and using the fact that $\delta = q_1/\sqrt{\theta}$, we obtain

$$\begin{aligned}\overline{G}(\boldsymbol{q}) &\geq \frac{1-\theta}{p}\Phi_1(\delta) - \frac{1}{\delta p}\left[\Phi_2(\delta) - \sigma\sqrt{p}\eta_1(\delta)\right] - \frac{q_1^2}{p}\left(1-\sigma\sqrt{p}\right)\eta_2(\delta) - \frac{\sqrt{\theta}}{2p}q_1^3\eta_1(\delta) - \frac{C_1\sqrt{\theta}q_1^3}{p}\max\left(\frac{q_1^3}{\theta^{3/2}}, 1\right) \\ &\geq \frac{1-\theta}{p}\Phi_1(\delta) - \frac{1}{\delta p}\left[\Phi_2(\delta) - \eta_1(\delta)\right] - \frac{q_1^2}{p}\frac{\eta_1(\delta)}{\delta} - \frac{q_1^2}{\theta p}\left(\frac{2\theta}{\sqrt{2\pi}} + \frac{3\theta^2}{2\sqrt{2\pi}} + C_1\theta^2\right),\end{aligned}$$

for some constant $C_1 > 0$, where we have used $q_1 \leq 3\sqrt{\theta}$ to simplify the bounds and the fact $\sigma\sqrt{p} = \sqrt{1-q_1^2} \geq 1-q_1^2$ to simplify the expression. Substituting the estimates in Lemma 7.4 and use the fact $\delta \mapsto \eta_1(\delta)/\delta$ is bounded, we obtain

$$\begin{aligned}\overline{G}(p) &\geq \frac{1}{p}\left(\frac{1}{40} - \frac{1}{\sqrt{2\pi}}\theta\right)\delta^2 - \frac{q_1^2}{\theta p}\left(c_1\theta + c_2\theta^2\right) \\ &\geq \frac{q_1^2}{\theta p}\left(\frac{1}{40} - \frac{1}{\sqrt{2\pi}}\theta - c_1\theta - c_2\theta^2\right)\end{aligned}$$

for some positive constants c_1 and c_2. We obtain the claimed result once θ_0 is made sufficiently small.

7.3.1.2 Auxiliary Results Used in the Proof

Lemma 7.2 *Let $\delta \doteq q_1/\sqrt{\theta}$. There exists some universal constant $C_T > 0$ such that we have the follow polynomial approximations hold for all $q_1 \in \left(0, \frac{1}{2}\right)$:*

$$\begin{aligned}
\left|\psi\left(-\frac{\alpha}{\sigma}\right) - \left[1 - \frac{1}{2}(1+\delta)^2 q_1^2\right]\psi(-1-\delta)\right| &\leq C_T\,(1+\delta)^2\, q_1^4,\\
\left|\psi\left(\frac{\beta}{\sigma}\right) - \left[1 - \frac{1}{2}(\delta-1)^2 q_1^2\right]\psi(\delta-1)\right| &\leq C_T\,(\delta-1)^2\, q_1^4,\\
\left|\Psi\left(-\frac{\alpha}{\sigma}\right) - \left[\Psi(-1-\delta) - \frac{1}{2}\psi(-1-\delta)(1+\delta)q_1^2\right]\right| &\leq C_T\,(1+\delta)^2\, q_1^4,\\
\left|\Psi\left(\frac{\beta}{\sigma}\right) - \left[\Psi(\delta-1) + \frac{1}{2}\psi(\delta-1)(\delta-1)q_1^2\right]\right| &\leq C_T\,(\delta-1)^2\, q_1^4,\\
\left|\Psi\left(-\frac{\lambda}{\sigma}\right) - \left[\Psi(-1) - \frac{1}{2}\psi(-1)q_1^2\right]\right| &\leq C_T q_1^4.
\end{aligned}$$

Proof First observe that for any $q_1 \in \left(0, \frac{1}{2}\right)$ it holds that

$$0 \leq \frac{1}{\sqrt{1-q_1^2}} - \left(1 + \frac{q_1^2}{2}\right) \leq q_1^4.$$

Hence we have

$$\begin{aligned}
-(1+\delta)\left(1 + \frac{1}{2}q_1^2 + q_1^4\right) \leq -\frac{\alpha}{\sigma} \leq -(1+\delta)\left(1 + \frac{1}{2}q_1^2\right),\\
(\delta-1)\left(1 + \frac{1}{2}q_1^2\right) \leq \frac{\beta}{\sigma} \leq (\delta-1)\left(1 + \frac{1}{2}q_1^2 + q_1^4\right), \text{ when } \delta \geq 1\\
(\delta-1)\left(1 + \frac{1}{2}q_1^2 + q_1^4\right) \leq \frac{\beta}{\sigma} \leq (\delta-1)\left(1 + \frac{1}{2}q_1^2\right), \text{ when } \delta \leq 1.
\end{aligned}$$

So we have

$$\psi\left(-(1+\delta)\left(1 + \frac{1}{2}q_1^2 + q_1^4\right)\right) \leq \psi\left(-\frac{\alpha}{\sigma}\right) \leq \psi\left(-(1+\delta)\left(1 + \frac{1}{2}q_1^2\right)\right).$$

By Taylor expansion of the left and right sides of the above two-side inequality around $-1-\delta$ using Lemma A.2, we obtain

$$\left|\psi\left(-\frac{\alpha}{\sigma}\right) - \psi(-1-\delta) - \frac{1}{2}(1+\delta)^2 q_1^2 \psi(-1-\delta)\right| \leq C_T\,(1+\delta)^2\, q_1^4,$$

for some numerical constant $C_T > 0$ sufficiently large. In the same way, we can obtain other claimed results.

■

Lemma 7.3 *For any $\delta \in [0,3]$, it holds that*

$$\Phi_2(\delta) - \eta_1(\delta) \;\geq\; \frac{\eta_1(3)}{9}\delta^3 \geq \frac{1}{20}\delta^3. \tag{7.3.10}$$

Proof Let us define

$$h(\delta) \;=\; \Phi_2(\delta) - \eta_1(\delta) - C\delta^3$$

for some $C > 0$ to be determined later. Then it is obvious that $h(0) = 0$. Direct calculation shows that

$$\frac{d}{d\delta}\Phi_1(\delta) = \eta_1(\delta), \quad \frac{d}{d\delta}\Phi_2(\delta) = \eta_2(\delta), \quad \frac{d}{d\delta}\eta_1(\delta) = \eta_2(\delta) - \delta\eta_1(\delta). \tag{7.3.11}$$

Thus, to show (7.3.10), it is sufficient to show that $h'(\delta) \geq 0$ for all $\delta \in [0,3]$. By differentiating $h(\delta)$ with respect to δ and use the results in (7.3.11), it is sufficient to have

$$h'(\delta) = \delta\eta_1(\delta) - 3C\delta^2 \geq 0 \Longleftrightarrow \eta_1(\delta) \geq 3C\delta$$

for all $\delta \in [0,3]$. We obtain the claimed result by observing that $\delta \mapsto \eta_1(\delta)/3\delta$ is monotonically decreasing over $\delta \in [0,3]$ as justified below.

Consider the function

$$p(\delta) \doteq \frac{\eta_1(\delta)}{3\delta} = \frac{1}{3\sqrt{2\pi}} \exp\left(-\frac{\delta^2+1}{2}\right) \frac{e^{\delta} - e^{-\delta}}{\delta}.$$

To show it is monotonically decreasing, it is enough to show $p'(\delta)$ is always nonpositive for $\delta \in (0,3)$, or equivalently

$$g(\delta) \doteq \left(e^{\delta} + e^{-\delta}\right)\delta - \left(\delta^2+1\right)\left(e^{\delta} - e^{-\delta}\right) \leq 0$$

for all $\delta \in (0,3)$, which can be easily verified by noticing that $g(0) = 0$ and $g'(\delta) \leq 0$ for all $\delta \geq 0$. ■

Lemma 7.4 *For any $\delta \in [0,3]$, we have*

$$(1-\theta)\Phi_1(\delta) - \frac{1}{\delta}\left[\Phi_2(\delta) - \eta_1(\delta)\right] \geq \left(\frac{1}{40} - \frac{1}{\sqrt{2\pi}}\theta\right)\delta^2. \tag{7.3.12}$$

Proof Let us define

$$g(\delta) \;=\; (1-\theta)\Phi_1(\delta) - \frac{1}{\delta}\left[\Phi_2(\delta) - \eta_1(\delta)\right] - c_0(\theta)\,\delta^2,$$

where $c_0(\theta) > 0$ is a function of θ. Thus, by the results in (7.3.11) and L'Hospital's rule, we have

$$\lim_{\delta\to 0}\frac{\Phi_2(\delta)}{\delta} = \lim_{\delta\to 0}\eta_2(\delta) = 2\psi(-1), \quad \lim_{\delta\to 0}\frac{\eta_1(\delta)}{\delta} = \lim_{\delta\to 0}[\eta_2(\delta) - \delta\eta_1(\delta)] = 2\psi(-1).$$

Combined that with the fact that $\Phi_1(0) = 0$, we conclude $g(0) = 0$. Hence, to show (7.3.12), it is sufficient to show that $g'(\delta) \geq 0$ for all $\delta \in [0, 3]$. Direct calculation using the results in (7.3.11) shows that

$$g'(\delta) = \frac{1}{\delta^2}[\Phi_2(\delta) - \eta_1(\delta)] - \theta\eta_1(\delta) - 2c_0(\theta)\delta.$$

Since $\eta_1(\delta)/\delta$ is monotonically decreasing as shown in Lemma 7.3, we have that for all $\delta \in (0, 3)$

$$\eta_1(\delta) \leq \delta\lim_{\delta\to 0}\frac{\eta(\delta)}{\delta} \leq \frac{2}{\sqrt{2\pi}}\delta.$$

Using the above bound and the main result from Lemma 7.3 again, we obtain

$$g'(\delta) \geq \frac{1}{20}\delta - \frac{2}{\sqrt{2\pi}}\theta\delta - 2c_0\delta.$$

Choosing $c_0(\theta) = \frac{1}{40} - \frac{1}{\sqrt{2\pi}}\theta$ completes the proof. ■

7.3.2 Finite Sample Concentration

In the following two subsections, we estimate the deviations around the expectations $\mathbb{E}[\overline{Q}_1(\boldsymbol{q})]$ and $\mathbb{E}[\overline{\boldsymbol{Q}}_2(\boldsymbol{q})]$, i.e., $|\overline{Q}_1(\boldsymbol{q}) - \mathbb{E}[\overline{Q}_1(\boldsymbol{q})]|$ and $\|\overline{\boldsymbol{Q}}_2(\boldsymbol{q}) - \mathbb{E}[\overline{\boldsymbol{Q}}_2(\boldsymbol{q})]\|$, and show that the total deviations fit into the gap $\overline{G}(\boldsymbol{q})$ we derived in Section 7.3.1. Our analysis is based on the scalar and vector Bernstein's inequalities with moment conditions. Finally, in Section 7.3.3, we uniformize the bound by applying the classical discretization argument.

7.3.2.1 Concentration for $\overline{Q}_1(\boldsymbol{q})$

Lemma 7.5 (Bounding $|\overline{Q}_1(\boldsymbol{q}) - \mathbb{E}[\overline{Q}_1(\boldsymbol{q})]|$) *For each $\boldsymbol{q} \in \mathbb{S}^{n-1}$, it holds for all $t > 0$ that*

$$\mathbb{P}\left[|\overline{Q}_1(\boldsymbol{q}) - \mathbb{E}[\overline{Q}_1(\boldsymbol{q})]| \geq t\right] \leq 2\exp\left(-\frac{\theta p^3 t^2}{8 + 4pt}\right).$$

Proof By (7.3.1), we know that

$$\overline{Q}_1(\boldsymbol{q}) = \frac{1}{p}\sum_{k=1}^{p} X_k^1, \quad X_k^1 = x_0(k)\mathcal{S}_\lambda[x_0(k)q_1 + Z_k]$$

where $Z_k = \boldsymbol{q}_2^\top \boldsymbol{g}_k \sim \mathcal{N}\left(0, \frac{\|\boldsymbol{q}_2\|^2}{p}\right)$. Thus, for any $m \geq 2$, by Lemma A.4, we have

$$\begin{aligned}
\mathbb{E}\left[\left|X_k^1\right|^m\right] &\leq \theta\left(\frac{1}{\sqrt{\theta p}}\right)^m \mathbb{E}\left[\left|\frac{q_1}{\sqrt{\theta p}} + Z_k\right|^m\right] \\
&= \theta\left(\frac{1}{\sqrt{\theta p}}\right)^m \sum_{l=0}^{m} \binom{m}{l} \left(\frac{q_1}{\sqrt{\theta p}}\right)^l \mathbb{E}\left[|Z_k|^{m-l}\right] \\
&= \theta\left(\frac{1}{\sqrt{\theta p}}\right)^m \sum_{l=0}^{m} \binom{m}{l} \left(\frac{q_1}{\sqrt{\theta p}}\right)^l (m-l-1)!! \left(\frac{\|\boldsymbol{q}_2\|}{\sqrt{p}}\right)^{m-l} \\
&\leq \frac{m!}{2}\theta\left(\frac{1}{\sqrt{\theta p}}\right)^m \left(\frac{q_1}{\sqrt{\theta p}} + \frac{\|\boldsymbol{q}_2\|}{\sqrt{p}}\right)^m \\
&\leq \frac{m!}{2}\theta\left(\frac{2}{\theta p}\right)^m = \frac{m!}{2}\frac{4}{\theta p^2}\left(\frac{2}{\theta p}\right)^{m-2}
\end{aligned}$$

let $\sigma_X^2 = 4/(\theta p^2)$ and $R = 2/(\theta p)$, apply Lemma A.7, we get

$$\mathbb{P}\left[\left|\overline{Q}_1(\boldsymbol{q}) - \mathbb{E}\left[\overline{Q}_1(\boldsymbol{q})\right]\right| \geq t\right] \leq 2\exp\left(-\frac{\theta p^3 t^2}{8 + 4pt}\right).$$

as desired. ■

7.3.2.2 Concentration for $\overline{\boldsymbol{Q}}_2(\boldsymbol{q})$

Lemma 7.6 (Bounding $\left\|\overline{\boldsymbol{Q}}_2(\boldsymbol{q}) - \mathbb{E}\left[\overline{\boldsymbol{Q}}_2(\boldsymbol{q})\right]\right\|$) *For each $\boldsymbol{q} \in \mathbb{S}^{n-1}$, it holds for all $t > 0$ that*

$$\mathbb{P}\left[\left\|\overline{\boldsymbol{Q}}_2(\boldsymbol{q}) - \mathbb{E}\left[\overline{\boldsymbol{Q}}_2(\boldsymbol{q})\right]\right\| > t\right] \leq 2(n+1)\exp\left(-\frac{\theta p^3 t^2}{128n + 16\sqrt{\theta n p t}}\right).$$

Before proving Lemma 7.6, we record the following useful results.

Lemma 7.7 *For any positive integer $s, l > 0$, we have*

$$\mathbb{E}\left[\left\|\boldsymbol{g}^k\right\|^s \left|\boldsymbol{q}_2^\top \boldsymbol{g}^k\right|^l\right] \leq \frac{(l+s)!!}{2}\left\|\boldsymbol{q}_2\right\|^l \frac{\left(2\sqrt{n}\right)^s}{\left(\sqrt{p}\right)^{s+l}}.$$

In particular, when $s = l$, we have

$$\mathbb{E}\left[\left\|\boldsymbol{g}^k\right\|^l \left|\boldsymbol{q}_2^\top \boldsymbol{g}^k\right|^l\right] \leq \frac{l!}{2}\left\|\boldsymbol{q}_2\right\|^l \left(\frac{4\sqrt{n}}{p}\right)^l$$

Proof Let $\mathcal{P}_{\boldsymbol{q}_2^\parallel} = \frac{\boldsymbol{q}_2\boldsymbol{q}_2^\top}{\|\boldsymbol{q}_2\|^2}$ and $\mathcal{P}_{\boldsymbol{q}_2^\perp} = \left(\boldsymbol{I} - \frac{1}{\|\boldsymbol{q}_2\|^2}\boldsymbol{q}_2\boldsymbol{q}_2^\top\right)$ denote the projection operators onto $\boldsymbol{q}_2$ and its orthogonal complement, respectively. By Lemma A.4, we have

$$\mathbb{E}\left[\left\|\boldsymbol{g}^k\right\|^s \left|\boldsymbol{q}_2^\top \boldsymbol{g}^k\right|^l\right] \leq \mathbb{E}\left[\left(\left\|\mathcal{P}_{\boldsymbol{q}_2^\parallel}\boldsymbol{g}^k\right\| + \left\|\mathcal{P}_{\boldsymbol{q}_2^\perp}\boldsymbol{g}^k\right\|\right)^s \left|\boldsymbol{q}_2^\top \boldsymbol{g}^k\right|^l\right]$$

$$
\begin{aligned}
&= \sum_{i=0}^{s} \binom{s}{i} \mathbb{E}\left[\left\|\mathcal{P}_{\boldsymbol{q}_2^{\perp}} \boldsymbol{g}^k\right\|^i\right] \mathbb{E}\left[\left|\boldsymbol{q}_2^{\top} \boldsymbol{g}^k\right|^l \left\|\mathcal{P}_{\boldsymbol{q}_2^{\|}} \boldsymbol{g}^k\right\|^{s-i}\right] \\
&= \sum_{i=0}^{s} \binom{s}{i} \mathbb{E}\left[\left\|\mathcal{P}_{\boldsymbol{q}_2^{\perp}} \boldsymbol{g}^k\right\|^i\right] \mathbb{E}\left[\left|\boldsymbol{q}_2^{\top} \boldsymbol{g}^k\right|^{l+s-i}\right] \frac{1}{\|\boldsymbol{q}_2\|^{s-i}} \\
&\leq \|\boldsymbol{q}_2\|^l \sum_{i=0}^{s} \binom{s}{i} \mathbb{E}\left[\left\|\mathcal{P}_{\boldsymbol{q}_2^{\perp}} \boldsymbol{g}^k\right\|^i\right] \left(\frac{1}{\sqrt{p}}\right)^{l+s-i} (l+s-i-1)!!.
\end{aligned}
$$

Using Lemma A.5 and the fact that $\left\|\mathcal{P}_{\boldsymbol{q}_2^{\perp}} \boldsymbol{g}^k\right\| \leq \left\|\boldsymbol{g}^k\right\|$, we obtain

$$
\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{g}^k\right\|^s \left|\boldsymbol{q}_2^{\top} \boldsymbol{g}^k\right|^l\right] &\leq \|\boldsymbol{q}_2\|^l \sum_{i=0}^{s} \binom{s}{i} \left(\frac{\sqrt{n}}{\sqrt{p}}\right)^i i!! \left(\frac{1}{\sqrt{p}}\right)^{l+s-i} (l+s-i-1)!! \\
&\leq \|\boldsymbol{q}_2\|^l \left(\frac{1}{\sqrt{p}}\right)^l \frac{(l+s)!!}{2} \left(\frac{\sqrt{n}}{\sqrt{p}} + \frac{1}{\sqrt{p}}\right)^s \\
&\leq \frac{(l+s)!!}{2} \|\boldsymbol{q}_2\|^l \frac{(2\sqrt{n})^s}{\left(\sqrt{p}\right)^{s+l}}.
\end{aligned}
$$

■

Now, we are ready to prove Lemma 7.6,

Proof By (7.3.1), note that

$$
\overline{\boldsymbol{Q}}_2 = \frac{1}{p} \sum_{k=1}^{p} \boldsymbol{X}_k^2, \quad \boldsymbol{X}_k^2 = \boldsymbol{g}^k \mathcal{S}_\lambda \left[x_0(k) q_1 + Z_k\right]
$$

where $Z_k = \boldsymbol{q}_2^{\top} \boldsymbol{g}^k$. Thus, for any $m \geq 2$, by Lemma 7.7, we have

$$
\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{X}_k^2\right\|^m\right] &\leq \theta \mathbb{E}\left[\left\|\boldsymbol{g}^k\right\|^m \left|\frac{q_1}{\sqrt{\theta p}} + \boldsymbol{q}_2^{\top} \boldsymbol{g}^k\right|^m\right] + (1-\theta) \mathbb{E}\left[\left\|\boldsymbol{g}^k\right\|^m \left|\boldsymbol{q}_2^{\top} \boldsymbol{g}^k\right|^m\right] \\
&\leq \theta \sum_{l=0}^{m} \binom{m}{l} \mathbb{E}\left[\left|\boldsymbol{q}_2^{\top} \boldsymbol{g}^k\right|^l \left\|\boldsymbol{g}^k\right\|^m\right] \left|\frac{q_1}{\sqrt{\theta p}}\right|^{m-l} + (1-\theta) \mathbb{E}\left[\left\|\boldsymbol{g}^k\right\|^m \left|\boldsymbol{q}_2^{\top} \boldsymbol{g}^k\right|^m\right] \\
&\leq \theta \left(\frac{2\sqrt{n}}{\sqrt{p}}\right)^m \sum_{l=0}^{m} \binom{m}{l} \frac{(m+l)!!}{2} \left(\frac{\|\boldsymbol{q}_2\|}{\sqrt{p}}\right)^l \left|\frac{q_1}{\sqrt{\theta p}}\right|^{m-l} + (1-\theta) \frac{m!}{2} \|\boldsymbol{q}_2\|^m \left(\frac{4\sqrt{n}}{p}\right)^m \\
&\leq \theta \frac{m!}{2} \left(\frac{4\sqrt{n}}{\sqrt{p}}\right)^m \left(\frac{\|\boldsymbol{q}_2\|}{\sqrt{p}} + \frac{q_1}{\sqrt{\theta p}}\right)^m + (1-\theta) \frac{m!}{2} \|\boldsymbol{q}_2\|^m \left(\frac{4\sqrt{n}}{p}\right)^m \\
&\leq \frac{m!}{2} \left(\frac{8\sqrt{n}}{\sqrt{\theta p}}\right)^m.
\end{aligned}
$$

Taking $\sigma_X^2 = 64n/(\theta p^2)$ and $R = 8\sqrt{n}/(\sqrt{\theta} p)$ and using vector Bernstein's inequality in Lemma A.8, we obtain

$$
\mathbb{P}\left[\left\|\overline{\boldsymbol{Q}}_2(\boldsymbol{q}) - \mathbb{E}\left[\overline{\boldsymbol{Q}}_2(\boldsymbol{q})\right]\right\| \geq t\right] \leq 2(n+1) \exp\left(-\frac{\theta p^3 t^2}{128n + 16\sqrt{\theta n} p t}\right),
$$

as desired. ■

7.3.3 Union Bound

Proposition 7.8 (Uniformizing the Bounds) *Suppose that $\theta > 1/\sqrt{n}$. Given any $\xi > 0$, there exists some constant $C(\xi)$, such that whenever $p \geq C(\xi) n^4 \log n$, we have*

$$\begin{aligned} \left|\overline{Q}_1(\boldsymbol{q}) - \mathbb{E}\left[\overline{Q}_1(\boldsymbol{q})\right]\right| &\leq \frac{2\xi}{\theta^{5/2} n^{3/2} p}, \\ \left\|\overline{\boldsymbol{Q}}_2(\boldsymbol{q}) - \mathbb{E}\left[\overline{\boldsymbol{Q}}_2(\boldsymbol{q})\right]\right\| &\leq \frac{2\xi}{\theta^2 np} \end{aligned}$$

hold uniformly for all $\boldsymbol{q} \in \mathbb{S}^{n-1}$, with probability at least $1 - c(\xi)p^{-2}$ for a positive constant $c(\xi)$.

Proof We apply the standard covering argument. For any $\varepsilon \in (0,1)$, by Lemma A.12, the unit hemisphere of interest can be covered by an ε-net $\mathcal{N}_\varepsilon$ of cardinality at most $(3/\varepsilon)^n$. For any $\boldsymbol{q} \in \mathbb{S}^{n-1}$, it can be written as

$$\boldsymbol{q} = \boldsymbol{q}' + \boldsymbol{e}$$

where $\boldsymbol{q}' \in \mathcal{N}_\varepsilon$ and $\|\boldsymbol{e}\| \leq \varepsilon$. Let a row of $\overline{\boldsymbol{Y}}$ be $\overline{\boldsymbol{y}}^k = \left[x_0(k), \boldsymbol{g}^k\right]^\top$, which is an independent copy of $\overline{\boldsymbol{y}} = [x_0, \boldsymbol{g}]^\top$. By (7.3.1), we have

$$\begin{aligned} &\left|\overline{Q}_1(\boldsymbol{q}) - \mathbb{E}\left[\overline{Q}_1(\boldsymbol{q})\right]\right| \\ &= \left|\frac{1}{p}\sum_{k=1}^{p}\left\{x_0(k)\mathcal{S}_\lambda\left[\left\langle\overline{\boldsymbol{y}}^k, \boldsymbol{q}' + \boldsymbol{e}\right\rangle\right] - \mathbb{E}\left[x_0(k)\mathcal{S}_\lambda\left[\left\langle\overline{\boldsymbol{y}}^k, \boldsymbol{q}' + \boldsymbol{e}\right\rangle\right]\right]\right\}\right| \\ &\leq \left|\frac{1}{p}\sum_{k=1}^{p} x_0(k)\mathcal{S}_\lambda\left[\left\langle\overline{\boldsymbol{y}}^k, \boldsymbol{q}' + \boldsymbol{e}\right\rangle\right] - \frac{1}{p}\sum_{k=1}^{p} x_0(k)\mathcal{S}_\lambda\left[\left\langle\overline{\boldsymbol{y}}^k, \boldsymbol{q}'\right\rangle\right]\right| + \left|\frac{1}{p}\sum_{k=1}^{p} x_0(k)\mathcal{S}_\lambda\left[\left\langle\overline{\boldsymbol{y}}^k, \boldsymbol{q}'\right\rangle\right] - \mathbb{E}\left[x_0\mathcal{S}_\lambda\left[\left\langle\overline{\boldsymbol{y}}, \boldsymbol{q}'\right\rangle\right]\right]\right| \\ &+ \left|\mathbb{E}\left[x_0\mathcal{S}_\lambda\left[\left\langle\overline{\boldsymbol{y}}, \boldsymbol{q}'\right\rangle\right]\right] - \mathbb{E}\left[x_0\mathcal{S}_\lambda\left[\left\langle\overline{\boldsymbol{y}}, \boldsymbol{q}' + \boldsymbol{e}\right\rangle\right]\right]\right|. \end{aligned}$$

Using Cauchy-Schwarz inequality and the fact that $\mathcal{S}_\lambda[\cdot]$ is a nonexpansive operator, we have

$$\begin{aligned} \left|\overline{Q}_1(\boldsymbol{q}) - \mathbb{E}\left[\overline{Q}_1(\boldsymbol{q})\right]\right| &\leq \left|\overline{Q}_1(\boldsymbol{q}') - \mathbb{E}\left[\overline{Q}_1(\boldsymbol{q}')\right]\right| + \left(\frac{1}{p}\sum_{k=1}^{p}|x_0(k)|\left\|\overline{\boldsymbol{y}}^k\right\| + \mathbb{E}\left[|x_0|\,\|\overline{\boldsymbol{y}}\|\right]\right)\|\boldsymbol{e}\| \\ &\leq \left|\overline{Q}_1(\boldsymbol{q}') - \mathbb{E}\left[\overline{Q}_1(\boldsymbol{q}')\right]\right| + \varepsilon\frac{1}{\sqrt{\theta p}}\left(\frac{2}{\sqrt{\theta p}} + \max_{k\in[p]}\left\|\boldsymbol{g}^k\right\| + \mathbb{E}\left[\|\boldsymbol{g}\|\right]\right). \end{aligned}$$

By Lemma A.10, $\max_{k\in[p]}\left\|\boldsymbol{g}^k\right\| \leq \sqrt{n/p} + 2\sqrt{2\log(2p)/p}$ with probability at least $1 - c_1 p^{-3}$. Also $\mathbb{E}\left[\|\boldsymbol{g}\|\right] \leq \left(\mathbb{E}\left[\|\boldsymbol{g}\|^2\right]\right)^{1/2} \leq \sqrt{n/p}$. Taking $t = \xi\theta^{-5/2}n^{-3/2}p^{-1}$ in Lemma 7.5 and applying a union bound with $\varepsilon =$

$\xi\theta^{-2}n^{-2}(\log 2p)^{-1/2}/7$, and combining with the above estimates, we obtain that

$$\left|\overline{Q}_1(\boldsymbol{q}) - \mathbb{E}\left[\overline{Q}_1(\boldsymbol{q})\right]\right| \leq \frac{\xi}{\theta^{5/2}n^{3/2}p} + \frac{\xi}{7\theta^{5/2}n^2\sqrt{\log(2p)p}}\left(4\sqrt{n} + 2\sqrt{2\log(2p)}\right) \leq \frac{2\xi}{\theta^{5/2}n^{3/2}p}$$

holds for all $\boldsymbol{q} \in \mathbb{S}^{n-1}$, with probability at least

$$1 - c_1 p^{-3} - 2\exp\left(-c_3(\xi)\, p/(\theta^4 n^3) + c_4(\xi)\, n\log n + c_5(\xi) n\log\log(2p)\right).$$

Similarly, by (7.3.1), we have

$$\begin{aligned}
\left\|\overline{\boldsymbol{Q}}_2(\boldsymbol{q}) - \mathbb{E}\left[\overline{\boldsymbol{Q}}_2(\boldsymbol{q})\right]\right\| &= \left\|\frac{1}{p}\sum_{k=1}^{p}\left\{\boldsymbol{g}^k\mathcal{S}_\lambda\left[\langle\overline{\boldsymbol{y}}^k, \boldsymbol{q}' + \boldsymbol{e}\rangle\right] - \mathbb{E}\left[\boldsymbol{g}\mathcal{S}_\lambda\left[\langle\overline{\boldsymbol{y}}, \boldsymbol{q}' + \boldsymbol{e}\rangle\right]\right]\right\}\right\| \\
&\leq \left\|\overline{\boldsymbol{Q}}_2(\boldsymbol{q}') - \mathbb{E}\left[\overline{\boldsymbol{Q}}_2(\boldsymbol{q}')\right]\right\| + \left(\frac{1}{p}\sum_{k=1}^{p}\left\|\boldsymbol{g}^k\right\|\left\|\overline{\boldsymbol{y}}^k\right\| + \mathbb{E}\left[\|\boldsymbol{g}\|\,\|\overline{\boldsymbol{y}}\|\right]\right)\|\boldsymbol{e}\| \\
&\leq \left\|\overline{\boldsymbol{Q}}_2(\boldsymbol{q}') - \mathbb{E}\left[\overline{\boldsymbol{Q}}_2(\boldsymbol{q}')\right]\right\| + \varepsilon\left[\max_{k\in[p]}\left\|\boldsymbol{g}^k\right\|\left(\frac{1}{\sqrt{\theta p}} + \max_{k\in[p]}\left\|\boldsymbol{g}^k\right\|\right) + \frac{\sqrt{n}}{\sqrt{\theta}p} + \frac{n}{p}\right].
\end{aligned}$$

Applying the above estimates for $\max_{k\in[p]}\left\|\boldsymbol{g}^k\right\|$, and taking $t = \xi\theta^{-2}n^{-1}p^{-1}$ in Lemma 7.6 and applying a union bound with $\varepsilon = \xi\theta^{-2}n^{-2}\log^{-1}(2p)/30$, we obtain that

$$\begin{aligned}
\left\|\overline{\boldsymbol{Q}}_2(\boldsymbol{q}) - \mathbb{E}\left[\overline{\boldsymbol{Q}}_2(\boldsymbol{q})\right]\right\| &\leq \frac{\xi}{\theta^2 np} + \frac{\xi}{30\theta^2 n^2\log(2p)}\left\{4\left(\sqrt{\frac{n}{p}} + \sqrt{\frac{2\log(2p)}{p}}\right)^2 + \frac{2n}{p}\right\} \\
&\leq \frac{\xi}{\theta^2 np} + \frac{\xi}{30\theta^2 n^2\log(2p)}\left\{\frac{16\log(2p)}{p} + \frac{10n}{p}\right\} \\
&\leq \frac{2\xi}{\theta^2 np}
\end{aligned}$$

holds for all $\boldsymbol{q} \in \mathbb{S}^{n-1}$, with probability at least

$$1 - c_1 p^{-3} - \exp\left(-c_6(\xi)\, p/(\theta^3 n^3) + c_7(\xi) n\log n + c_8(\xi) n\log\log(2p)\right).$$

Taking $p \geq C_9(\xi) n^4\log n$ and simplifying the probability terms complete the proof. ∎

7.3.4 $Q(q)$ approximates $\overline{Q}(q)$

Proposition 7.9 *Suppose $\theta > 1/\sqrt{n}$. For any $\xi > 0$, there exists some constant $C(\xi)$, such that whenever $p \geq C(\xi)\, n^4\log n$, the following bounds*

$$\sup_{\boldsymbol{q}\in\mathbb{S}^{n-1}}\left|Q_1(\boldsymbol{q}) - \overline{Q}_1(\boldsymbol{q})\right| \leq \frac{\xi}{\theta^{5/2}n^{3/2}p} \tag{7.3.13}$$

$$\sup_{\boldsymbol{q}\in\mathbb{S}^{n-1}} \left\|\boldsymbol{Q}_2(\boldsymbol{q}) - \overline{\boldsymbol{Q}}_2(\boldsymbol{q})\right\| \leq \frac{\xi}{\theta^2 np}, \tag{7.3.14}$$

hold with probability at least $1 - c(\xi)p^{-2}$ *for a positive constant* $c(\xi)$.

Proof First, for any $\boldsymbol{q} \in \mathbb{S}^{n-1}$, from (7.3.1), we know that

$$\begin{aligned}
&\left|\overline{Q}_1(\boldsymbol{q}) - Q_1(\boldsymbol{q})\right| \\
=& \left|\frac{1}{p}\sum_{k=1}^{p} x_0(k)\mathcal{S}_\lambda\left[\boldsymbol{q}^\top\overline{\boldsymbol{y}}^k\right] - \frac{1}{p}\sum_{k=1}^{p}\frac{x_0(k)}{\|\boldsymbol{x}_0\|}\mathcal{S}_\lambda\left[\boldsymbol{q}^\top\boldsymbol{y}^k\right]\right| \\
\leq& \left|\frac{1}{p}\sum_{k=1}^{p} x_0(k)\mathcal{S}_\lambda\left[\boldsymbol{q}^\top\overline{\boldsymbol{y}}^k\right] - \frac{1}{p}\sum_{k=1}^{p} x_0(k)\mathcal{S}_\lambda\left[\boldsymbol{q}^\top\boldsymbol{y}^k\right]\right| + \left|\frac{1}{p}\sum_{k=1}^{p} x_0(k)\mathcal{S}_\lambda\left[\boldsymbol{q}^\top\boldsymbol{y}^k\right] - \frac{1}{p}\sum_{k=1}^{p}\frac{x_0(k)}{\|\boldsymbol{x}_0\|}\mathcal{S}_\lambda\left[\boldsymbol{q}^\top\boldsymbol{y}^k\right]\right| \\
\leq& \frac{1}{p}\sum_{k=1}^{p}|x_0(k)|\left|\mathcal{S}_\lambda\left[\boldsymbol{q}^\top\overline{\boldsymbol{y}}^k\right] - \mathcal{S}_\lambda\left[\boldsymbol{q}^\top\boldsymbol{y}^k\right]\right| + \frac{1}{p}\sum_{k=1}^{p}|x_0(k)|\left|1 - \frac{1}{\|\boldsymbol{x}_0\|}\right|\left|\mathcal{S}_\lambda\left[\boldsymbol{q}^\top\boldsymbol{y}^k\right]\right|.
\end{aligned}$$

For any $\mathcal{I} = \mathrm{supp}(\boldsymbol{x}_0)$, using the fact that $\mathcal{S}_\lambda[\cdot]$ is a nonexpansive operator, we have

$$\begin{aligned}
\sup_{\boldsymbol{q}\in\mathbb{S}^{n-1}}\left|\overline{Q}_1(\boldsymbol{q}) - Q_1(\boldsymbol{q})\right| &\leq \frac{1}{p}\sup_{\boldsymbol{q}\in\mathbb{S}^{n-1}}\sum_{k\in\mathcal{I}}|x_0(k)|\left|\boldsymbol{q}^\top\left(\overline{\boldsymbol{y}}^k - \boldsymbol{y}^k\right)\right| + \left|1 - \frac{1}{\|\boldsymbol{x}_0\|}\right|\frac{1}{p}\sup_{\boldsymbol{q}\in\mathbb{S}^{n-1}}\sum_{k\in\mathcal{I}}|x_0(k)|\left|\boldsymbol{q}^\top\boldsymbol{y}^k\right| \\
&= \frac{1}{\sqrt{\theta}p^{3/2}}\left(\left\|\overline{\boldsymbol{Y}}_{\mathcal{I}} - \boldsymbol{Y}_{\mathcal{I}}\right\|_{\ell^2\to\ell^1} + \left|1 - \frac{1}{\|\boldsymbol{x}_0\|}\right|\left\|\boldsymbol{Y}_{\mathcal{I}}\right\|_{\ell^2\to\ell^1}\right).
\end{aligned}$$

By Lemma A.15 and Lemma A.17 in Appendix A.2, we have the following holds

$$\sup_{\boldsymbol{q}\in\mathbb{S}^{n-1}}\left|\overline{Q}_1(\boldsymbol{q}) - Q_1(\boldsymbol{q})\right| \leq \frac{1}{\sqrt{\theta}p^{3/2}}\left(20\sqrt{\frac{n\log p}{\theta}} + \frac{4\sqrt{2}}{5}\sqrt{\frac{n\log p}{\theta^2 p}} \times 7\sqrt{2\theta p}\right) \leq \frac{32}{\theta p^{3/2}}\sqrt{n\log p},$$

with probability at least $1 - c_1 p^{-2}$, provided $p \geq C_2 n$ and $\theta > 1/\sqrt{n}$. Simple calculation shows that it is enough to have $p \geq C_3(\xi)\, n^4 \log n$ for some sufficiently large $C_1(\xi)$ to obtain the claimed result in (7.3.13). Similarly, by Lemma A.17 and Lemma A.18 in Appendix A.2, we have

$$\begin{aligned}
&\sup_{\boldsymbol{q}\in\mathbb{S}^{n-1}}\left\|\overline{\boldsymbol{Q}}_2(\boldsymbol{q}) - \boldsymbol{Q}_2(\boldsymbol{q})\right\| \\
=& \sup_{\boldsymbol{q}\in\mathbb{S}^{n-1}}\left\|\frac{1}{p}\sum_{k=1}^{p}\boldsymbol{g}^k\mathcal{S}_\lambda\left[\boldsymbol{q}^\top\overline{\boldsymbol{y}}^k\right] - \frac{1}{p}\sum_{k=1}^{p}\boldsymbol{g}'^k\mathcal{S}_\lambda\left[\boldsymbol{q}^\top\boldsymbol{y}^k\right]\right\| \\
\leq& \sup_{\boldsymbol{q}\in\mathbb{S}^{n-1}}\left\|\frac{1}{p}\sum_{k=1}^{p}\boldsymbol{g}^k\mathcal{S}_\lambda\left[\boldsymbol{q}^\top\overline{\boldsymbol{y}}^k\right] - \frac{1}{p}\sum_{k=1}^{p}\boldsymbol{g}'^k\mathcal{S}_\lambda\left[\boldsymbol{q}^\top\overline{\boldsymbol{y}}^k\right]\right\| + \left\|\frac{1}{p}\sum_{k=1}^{p}\boldsymbol{g}'^k\mathcal{S}_\lambda\left[\boldsymbol{q}^\top\overline{\boldsymbol{y}}^k\right] - \frac{1}{p}\sum_{k=1}^{p}\boldsymbol{g}'^k\mathcal{S}_\lambda\left[\boldsymbol{q}^\top\boldsymbol{y}^k\right]\right\| \\
\leq& \frac{1}{p}\sup_{\boldsymbol{q}\in\mathbb{S}^{n-1}}\sum_{k=1}^{p}\left\|\boldsymbol{g}^k - \boldsymbol{g}'^k\right\|\left|\boldsymbol{q}^\top\overline{\boldsymbol{y}}^k\right| + \frac{1}{p}\sup_{\boldsymbol{q}\in\mathbb{S}^{n-1}}\sum_{k=1}^{p}\left\|\boldsymbol{g}'^k\right\|\left|\boldsymbol{q}^\top\left(\overline{\boldsymbol{y}}^k - \boldsymbol{y}^k\right)\right| \\
\leq& \frac{1}{p}\left(\|\boldsymbol{G} - \boldsymbol{G}'\|_{\ell^2\to\ell^\infty}\left\|\overline{\boldsymbol{Y}}\right\|_{\ell^2\to\ell^1} + \|\boldsymbol{G}'\|_{\ell^2\to\ell^\infty}\left\|\overline{\boldsymbol{Y}} - \boldsymbol{Y}\right\|_{\ell^2\to\ell^1}\right)
\end{aligned}$$

$$\leq \frac{1}{p}\left(\frac{120\max(n,\log(2p))}{\sqrt{p}}+\frac{300\sqrt{n\log(2p)}\max(\sqrt{n},\sqrt{\log(2p)})}{\sqrt{\theta p}}\right) \leq \frac{420\sqrt{n\log(2p)}\max(\sqrt{n},\sqrt{\log(2p)})}{\theta^{1/2}p^{3/2}}$$

with probability at least $1-c_4p^{-2}$ provided $p \geq C_4 n$ and $\theta > 1/\sqrt{n}$. It is sufficient to have $p \geq C_5(\xi)\, n^4 \log n$ to obtain the claimed result (7.3.14). ■

7.4 Large $|q_1|$ Iterates Staying in Safe Region for Rounding

In this appendix, we prove Proposition 4.5 in Section 4.

Proof [Proof of Proposition 4.5] For notational simplicity, w.l.o.g. we will proceed to prove assuming $q_1 > 0$. The proof for $q_1 < 0$ is similar by symmetry. It is equivalent to show that

$$\frac{\|\boldsymbol{Q}_2(\boldsymbol{q})\|}{|Q_1(\boldsymbol{q})|} < \sqrt{\frac{1}{4\theta}-1},$$

which is implied by

$$\mathcal{L}(\boldsymbol{q}) \doteq \frac{\|\mathbb{E}[\overline{\boldsymbol{Q}}_2(\boldsymbol{q})]\| + \|\boldsymbol{Q}_2(\boldsymbol{q}) - \mathbb{E}[\overline{\boldsymbol{Q}}_2(\boldsymbol{q})]\|}{\mathbb{E}[\overline{Q}_1(\boldsymbol{q})] - |Q_1(\boldsymbol{q}) - \mathbb{E}[\overline{Q}_1(\boldsymbol{q})]|} < \sqrt{\frac{1}{4\theta}-1}$$

for any $\boldsymbol{q} \in \mathbb{S}^{n-1}$ satisfying $q_1 > 3\sqrt{\theta}$. Recall from (7.3.7) that

$$\mathbb{E}[\overline{Q}_1(\boldsymbol{q})] = \sqrt{\frac{\theta}{p}}\left\{\left[\alpha\Psi\left(-\frac{\alpha}{\sigma}\right)+\beta\Psi\left(\frac{\beta}{\sigma}\right)\right]+\sigma\left[\psi\left(\frac{\beta}{\sigma}\right)-\psi\left(-\frac{\alpha}{\sigma}\right)\right]\right\},$$

where

$$\alpha = \frac{1}{\sqrt{p}}\left(\frac{q_1}{\sqrt{\theta}}+1\right), \quad \beta = \frac{1}{\sqrt{p}}\left(\frac{q_1}{\sqrt{\theta}}-1\right), \quad \sigma = \|\boldsymbol{q}_2\|/\sqrt{p}.$$

Noticing the fact that

$$\psi\left(\frac{\beta}{\sigma}\right)-\psi\left(-\frac{\alpha}{\sigma}\right) \geq 0,$$
$$\Psi\left(\frac{\beta}{\sigma}\right) = \Psi\left(\frac{1}{\sqrt{1-q_1^2}}\left(\frac{q_1}{\sqrt{\theta}}-1\right)\right) \geq \Psi(2) \geq \frac{19}{20} \quad \text{for } q_1 > 3\sqrt{\theta},$$

we have

$$\mathbb{E}[\overline{Q}_1(\boldsymbol{q})] \geq \frac{\sqrt{\theta}}{p}\left\{\frac{q_1}{\sqrt{\theta}}\left[\Psi\left(-\frac{\alpha}{\sigma}\right)+\Psi\left(\frac{\beta}{\sigma}\right)\right]+\Psi\left(-\frac{\alpha}{\sigma}\right)-\Psi\left(\frac{\beta}{\sigma}\right)\right\} \geq \frac{2\sqrt{\theta}}{p}\Psi\left(\frac{\beta}{\sigma}\right) \geq \frac{19}{10}\frac{\sqrt{\theta}}{p}.$$

Moreover, from (7.3.8), we have

$$\begin{aligned}\left\|\mathbb{E}\left[\overline{\boldsymbol{Q}}_2\left(\boldsymbol{q}\right)\right]\right\| &= \|\boldsymbol{q}_2\|\left\{\frac{2\left(1-\theta\right)}{p}\Psi\left(-\frac{\lambda}{\sigma}\right)+\frac{\theta}{p}\left[\Psi\left(-\frac{\alpha}{\sigma}\right)+\Psi\left(\frac{\beta}{\sigma}\right)\right]\right\}\\ &\leq \frac{2\left(1-\theta\right)}{p}\Psi\left(-1\right)+\frac{\theta}{p}\left[\Psi\left(-1\right)+1\right]\leq\frac{2}{p}\Psi\left(-1\right)+\frac{\theta}{p}\leq\frac{2}{5p}+\frac{\theta}{p},\end{aligned}$$

where we have used the fact that $-\lambda/\sigma\leq-1$ and $-\alpha/\sigma\leq-1$. Moreover, from results in Proposition 7.8 and Proposition 7.9 in Appendix 7.3, we know that

$$\begin{aligned}\sup_{\boldsymbol{q}\in\mathbb{S}^{n-1}}\left|Q_1(\boldsymbol{q})-\mathbb{E}\left[\overline{Q}_1(\boldsymbol{q})\right]\right| &\leq \sup_{\boldsymbol{q}\in\mathbb{S}^{n-1}}\left|Q_1(\boldsymbol{q})-\overline{Q}_1(\boldsymbol{q})\right|+\sup_{\boldsymbol{q}\in\mathbb{S}^{n-1}}\left|\overline{Q}_1(\boldsymbol{q})-\mathbb{E}\left[\overline{Q}_1(\boldsymbol{q})\right]\right|\leq\frac{1}{2\times10^5\theta^{5/2}n^{3/2}p},\\ \sup_{\boldsymbol{q}\in\mathbb{S}^{n-1}}\left\|\boldsymbol{Q}(\boldsymbol{q})-\mathbb{E}\left[\overline{\boldsymbol{Q}}(\boldsymbol{q})\right]\right\| &\leq \sup_{\boldsymbol{q}\in\mathbb{S}^{n-1}}\left\|\boldsymbol{Q}(\boldsymbol{q})-\overline{\boldsymbol{Q}}(\boldsymbol{q})\right\|+\sup_{\boldsymbol{q}\in\mathbb{S}^{n-1}}\left\|\overline{\boldsymbol{Q}}(\boldsymbol{q})-\mathbb{E}\left[\overline{\boldsymbol{Q}}(\boldsymbol{q})\right]\right\|\leq\frac{1}{2\times10^5\theta^2np}\end{aligned}$$

hold with probability at least $1-c_1p^{-2}$ provided that $p\geq\Omega\left(n^4\log n\right)$. Hence, with high probability, we have

$$\mathcal{L}\left(\boldsymbol{q}\right)\leq\frac{2/(5p)+\theta/p+(2\times10^5\theta^2np)^{-1}}{19\sqrt{\theta}/(10p)-(2\times10^5\theta^{5/2}n^{3/2}p)^{-1}}\leq\frac{3/5}{18\sqrt{\theta}/10}\leq\frac{1}{3\sqrt{\theta}}<\sqrt{\frac{1}{4\theta}-1},$$

whenever θ is sufficiently small. This completes the proof. ■

Now, keep the notation in Appendix 7.3 for general orthonormal basis $\widehat{\boldsymbol{Y}}=\boldsymbol{Y}\boldsymbol{U}$. For any current iterate $\boldsymbol{q}\in\mathbb{S}^{n-1}$ that is close enough to the target solution, i.e., $\left|\left\langle\boldsymbol{q},\boldsymbol{U}^\top\boldsymbol{e}_1\right\rangle\right|=\left|\langle\boldsymbol{U}\boldsymbol{q},\boldsymbol{e}_1\rangle\right|\geq3\sqrt{\theta}$, we have

$$\frac{\left|\left\langle\boldsymbol{Q}\left(\boldsymbol{q};\widehat{\boldsymbol{Y}}\right),\boldsymbol{U}^\top\boldsymbol{e}_1\right\rangle\right|}{\left\|\boldsymbol{Q}\left(\boldsymbol{q};\widehat{\boldsymbol{Y}}\right)\right\|}=\frac{\left|\left\langle\boldsymbol{U}\boldsymbol{Q}\left(\boldsymbol{q};\widehat{\boldsymbol{Y}}\right),\boldsymbol{e}_1\right\rangle\right|}{\left\|\boldsymbol{U}\boldsymbol{Q}\left(\boldsymbol{q};\widehat{\boldsymbol{Y}}\right)\right\|}=\frac{\left|\langle\boldsymbol{Q}\left(\boldsymbol{U}\boldsymbol{q};\boldsymbol{Y}\right),\boldsymbol{e}_1\rangle\right|}{\left\|\boldsymbol{Q}\left(\boldsymbol{U}\boldsymbol{q};\boldsymbol{Y}\right)\right\|},$$

where we have applied the identity proved in (7.3.2). Taking $\boldsymbol{U}\boldsymbol{q}\in\mathbb{S}^{n-1}$ as the object of interest, by Proposition 4.5, we conclude that

$$\frac{\left|\langle\boldsymbol{Q}\left(\boldsymbol{U}\boldsymbol{q};\boldsymbol{Y}\right),\boldsymbol{e}_1\rangle\right|}{\left\|\boldsymbol{Q}\left(\boldsymbol{U}\boldsymbol{q};\boldsymbol{Y}\right)\right\|}\geq2\sqrt{\theta}$$

with high probability.

7.5 Bounding Iteration Complexity

In this appendix, we prove Proposition 4.6 in Section 4.

Proof [Proof of Proposition 4.6] Recall from Proposition 4.4 in Section 4, the gap

$$G(\boldsymbol{q})=\frac{\left|Q_1(\boldsymbol{q})\right|}{\left|q_1\right|}-\frac{\left\|\boldsymbol{Q}_2(\boldsymbol{q})\right\|}{\left\|\boldsymbol{q}\right\|}\ \geq\ \frac{1}{10^4\theta^2np}$$

holds uniformly over $\boldsymbol{q} \in \mathbb{S}^{n-1}$ satisfying $\frac{1}{10\sqrt{\theta n}} \leq |q_1| \leq 3\sqrt{\theta}$, with probability at least $1 - c_1 p^{-2}$, provided $p \geq C_2 n^4 \log n$. The gap $G(\boldsymbol{q})$ implies that

$$\begin{aligned}
&\left|\widetilde{Q}_1(\boldsymbol{q})\right| \doteq \frac{|Q_1(\boldsymbol{q})|}{\|\boldsymbol{Q}(\boldsymbol{q})\|} \geq \frac{|q_1| \|\boldsymbol{Q}_2(\boldsymbol{q})\|}{\|\boldsymbol{q}\| \|\boldsymbol{Q}(\boldsymbol{q})\|} + \frac{|q_1|}{10^4\theta^2 np \|\boldsymbol{Q}(\boldsymbol{q})\|} \\
\Longleftrightarrow &\left|\widetilde{Q}_1(\boldsymbol{q})\right| \geq \frac{|q_1|}{\|\boldsymbol{q}_2\|}\sqrt{1 - \left|\widetilde{Q}_1(\boldsymbol{q})\right|^2} + \frac{|q_1|}{10^4\theta^2 np \|\boldsymbol{Q}(\boldsymbol{q})\|} \\
\Longrightarrow &\left|\widetilde{Q}_1(\boldsymbol{q})\right|^2 \geq |q_1|^2 \left(1 + \frac{\|\boldsymbol{q}_2\|^2}{10^8\theta^4 n^2 p^2 \|\boldsymbol{Q}(\boldsymbol{q})\|^2}\right).
\end{aligned}$$

Given the set Γ defined in (4.2.5), now we know that

$$\begin{aligned}
\sup_{\boldsymbol{q}\in\Gamma} \|\boldsymbol{Q}(\boldsymbol{q})\| \leq &\sup_{\boldsymbol{q}\in\Gamma} \left|\mathbb{E}\overline{Q}_1(\boldsymbol{q})\right| + \sup_{\boldsymbol{q}\in\mathbb{S}^{n-1}} \left|\mathbb{E}\overline{Q}_1(\boldsymbol{q}) - \overline{Q}_1(\boldsymbol{q})\right| + \sup_{\boldsymbol{q}\in\mathbb{S}^{n-1}} \left|Q_1(\boldsymbol{q}) - \overline{Q}_1(\boldsymbol{q})\right| \\
&+ \sup_{\boldsymbol{q}\in\Gamma} \left\|\mathbb{E}\overline{\boldsymbol{Q}}_2(\boldsymbol{q})\right\| + \sup_{\boldsymbol{q}\in\mathbb{S}^{n-1}} \left\|\mathbb{E}\overline{\boldsymbol{Q}}_2(\boldsymbol{q}) - \overline{\boldsymbol{Q}}_2(\boldsymbol{q})\right\| + \sup_{\boldsymbol{q}\in\mathbb{S}^{n-1}} \left\|\boldsymbol{Q}_2(\boldsymbol{q}) - \overline{\boldsymbol{Q}}_2(\boldsymbol{q})\right\| \\
\leq &\sup_{\boldsymbol{q}\in\Gamma} \left|\mathbb{E}\overline{Q}_1(\boldsymbol{q})\right| + \sup_{\boldsymbol{q}\in\Gamma} \left|\mathbb{E}\overline{\boldsymbol{Q}}_2(\boldsymbol{q})\right| + \frac{1}{pn}
\end{aligned}$$

with probability at least $1 - c_3 p^{-2}$ provided $p \geq C_4 n^4 \log n$ and $\theta > 1/\sqrt{n}$. Here we have used Proposition 7.8 and Proposition 7.9 to bound the magnitudes of the four difference terms. To bound the magnitudes of the expectations, we have

$$\begin{aligned}
\left|\mathbb{E}\overline{Q}_1(\boldsymbol{q})\right| &= \left|\mathbb{E}\frac{1}{p}\sum_{k=1}^{p} x_0(k) S_\lambda\left[x_0(k) q_1 + \boldsymbol{q}_2^\top \boldsymbol{g}^k\right]\right| \leq \frac{1}{\sqrt{\theta p}}\left(\frac{1}{\sqrt{\theta p}} + \mathbb{E}\|\boldsymbol{g}\|\right) \leq \frac{3\sqrt{n}}{\sqrt{\theta p}} \leq \frac{3n}{p}, \\
\left\|\mathbb{E}\overline{\boldsymbol{Q}}_2(\boldsymbol{q})\right\| &= \left\|\mathbb{E}\frac{1}{p}\sum_{k=1}^{p} \boldsymbol{g}^k S_\lambda\left[x_0(k) q_1 + \boldsymbol{q}_2^\top \boldsymbol{g}^k\right]\right\| \leq \frac{1}{\sqrt{\theta p}}\mathbb{E}\|\boldsymbol{g}\| + \mathbb{E}\|\boldsymbol{g}\|^2 \leq \frac{3n}{p}
\end{aligned}$$

hold uniformly for all $\boldsymbol{q} \in \Gamma$, provided $\theta > 1/\sqrt{n}$. Thus, we obtain that

$$\sup_{\boldsymbol{q}\in\Gamma} \|\boldsymbol{Q}(\boldsymbol{q})\| \leq \frac{3n}{p} + \frac{3n}{p} + \frac{1}{np} \leq \frac{7n}{p}$$

with probability at least $1 - c_3 p^{-2}$ provided $p \geq C_4 n^4 \log n$ and $\theta > 1/\sqrt{n}$. So we conclude that

$$\frac{\left|\widetilde{Q}_1(\boldsymbol{q})\right|}{|q_1|} \geq \sqrt{1 + \frac{1 - 9\theta}{10^8 \times 7^2 \times \theta^4 n^4}}.$$

Thus, starting with any $\boldsymbol{q} \in \mathbb{S}^{n-1}$ such that $|q_1| \geq \frac{1}{10\sqrt{\theta n}}$, we will need at most

$$T = \frac{2\log\left(3\sqrt{\theta}/\frac{1}{10\sqrt{\theta n}}\right)}{\log\left(1 + \frac{1-9\theta}{10^8\times 7^2\times\theta^4 n^4}\right)} = \frac{2\log\left(30\theta\sqrt{n}\right)}{\log\left(1 + \frac{1-9\theta}{10^8\times 7^2\times\theta^4 n^4}\right)} \leq \frac{2\log\left(30\theta\sqrt{n}\right)}{(\log 2)\frac{1-9\theta}{10^8\times 7^2\times\theta^4 n^4}} \leq C_5 n^4 \log n$$

steps to arrive at a $\overline{\boldsymbol{q}} \in \mathbb{S}^{n-1}$ with $|\bar{q}_1| \geq 3\sqrt{\theta}$ for the first time. Here we have assumed $\theta_0 < 1/9$ and used

the fact that $\log(1+x) \geq x \log 2$ for $x \in [0,1]$ to simplify the final result. ■

7.6 Rounding to the Desired Solution

In this appendix, we prove Proposition 4.7 in Section 4. For convenience, we will assume the notations we used in Appendix A.2. Then the rounding scheme can be written as

$$\min_{\boldsymbol{q}} \|\boldsymbol{Y}\boldsymbol{q}\|, \quad \text{s.t. } \langle \overline{\boldsymbol{q}}, \boldsymbol{q} \rangle = 1. \tag{7.6.1}$$

We will show the rounding procedure get us to the desired solution with high probability, regardless of the particular orthonormal basis used.

Proof [Proof of Proposition 4.7] The rounding program (7.6.1) can be written as

$$\inf_{\boldsymbol{q}} \|\boldsymbol{Y}\boldsymbol{q}\|_1, \quad \text{s.t. } \overline{q}_1 q_1 + \langle \overline{\boldsymbol{q}}_2, \boldsymbol{q}_2 \rangle = 1. \tag{7.6.2}$$

Consider its relaxation

$$\inf_{\boldsymbol{q}} \|\boldsymbol{Y}\boldsymbol{q}\|_1, \quad \text{s.t. } \overline{q}_1 q_1 + \|\overline{\boldsymbol{q}}_2\| \, \|\boldsymbol{q}_2\| \geq 1. \tag{7.6.3}$$

It is obvious that the feasible set of (7.6.3) contains that of (7.6.2). So if $\boldsymbol{e}_1/\overline{q}_1$ is the unique optimal solution (UOS) of (7.6.3), it is also the UOS of (7.6.2). Let $\mathcal{I} = \operatorname{supp}(\boldsymbol{x}_0)$, and consider a modified problem

$$\inf_{\boldsymbol{q}} \left\| \frac{\boldsymbol{x}_0}{\|\boldsymbol{x}_0\|} \right\|_1 |q_1| - \|\boldsymbol{G}'_{\mathcal{I}}\boldsymbol{q}_2\|_1 + \|\boldsymbol{G}'_{\mathcal{I}^c}\boldsymbol{q}_2\|_1, \quad \text{s.t. } \overline{q}_1 q_1 + \|\overline{\boldsymbol{q}}_2\| \, \|\boldsymbol{q}_2\| \geq 1. \tag{7.6.4}$$

The objective value of (7.6.4) lower bounds the objective value of (7.6.3), and are equal when $\boldsymbol{q} = \boldsymbol{e}_1/\overline{q}_1$. So if $\boldsymbol{q} = \boldsymbol{e}_1/\overline{q}_1$ is the UOS to (7.6.4), it is also UOS to (7.6.3), and hence UOS to (7.6.2) by the argument above. Now

$$\begin{aligned} -\|\boldsymbol{G}'_{\mathcal{I}}\boldsymbol{q}_2\|_1 + \|\boldsymbol{G}'_{\mathcal{I}^c}\boldsymbol{q}_2\|_1 &\geq -\|\boldsymbol{G}_{\mathcal{I}}\boldsymbol{q}_2\|_1 + \|\boldsymbol{G}_{\mathcal{I}^c}\boldsymbol{q}_2\|_1 - \|(\boldsymbol{G}-\boldsymbol{G}')\,\boldsymbol{q}_2\|_1 \\ &\geq -\|\boldsymbol{G}_{\mathcal{I}}\boldsymbol{q}_2\|_1 + \|\boldsymbol{G}_{\mathcal{I}^c}\boldsymbol{q}_2\|_1 - \|\boldsymbol{G}-\boldsymbol{G}'\|_{\ell^2\to\ell^1} \|\boldsymbol{q}_2\|. \end{aligned}$$

When $p \geq C_1 n$, by Lemma A.14 and Lemma A.17, we know that

$$\begin{aligned} &-\|\boldsymbol{G}_{\mathcal{I}}\boldsymbol{q}_2\|_1 + \|\boldsymbol{G}_{\mathcal{I}^c}\boldsymbol{q}_2\|_1 - \|\boldsymbol{G}-\boldsymbol{G}'\|_{\ell^2\to\ell^1} \|\boldsymbol{q}_2\| \\ &\qquad \geq -\frac{6}{5}\sqrt{\frac{2}{\pi}}2\theta\sqrt{p}\,\|\boldsymbol{q}_2\| + \frac{24}{25}\sqrt{\frac{2}{\pi}}(1-2\theta)\sqrt{p}\,\|\boldsymbol{q}_2\| - 4\sqrt{n}\,\|\boldsymbol{q}_2\| - 7\sqrt{\log(2p)}\,\|\boldsymbol{q}_2\| \doteq \zeta\,\|\boldsymbol{q}_2\| \end{aligned}$$

holds with probability at least $1 - c_2 p^{-2}$. Thus, we make a further relaxation of problem (7.6.2) by

$$\inf_{\boldsymbol{q}} \left\| \frac{\boldsymbol{x}_0}{\|\boldsymbol{x}_0\|} \right\|_1 |q_1| + \zeta \|\boldsymbol{q}_2\|, \quad \text{s.t. } \overline{q}_1 q_1 + \|\overline{\boldsymbol{q}}_2\| \|\boldsymbol{q}_2\| \geq 1, \tag{7.6.5}$$

whose objective value lower bounds that of (7.6.4). By similar arguments, if $\boldsymbol{e}_1/\overline{q}_1$ is UOS to (7.6.5), it is UOS to (7.6.2). At the optimal solution to (7.6.5), notice that it is necessary to have $\text{sign}(q_1) = \text{sign}(\overline{q}_1)$ and $\overline{q}_1 q_1 + \|\overline{\boldsymbol{q}}_2\| \|\boldsymbol{q}_2\| = 1$. So (7.6.5) is equivalent to

$$\inf_{\boldsymbol{q}} \left\| \frac{\boldsymbol{x}_0}{\|\boldsymbol{x}_0\|} \right\|_1 |q_1| + \zeta \|\boldsymbol{q}_2\|, \quad \text{s.t. } \overline{q}_1 q_1 + \|\overline{\boldsymbol{q}}_2\| \|\boldsymbol{q}_2\| = 1. \tag{7.6.6}$$

which is further equivalent to

$$\inf_{q_1} \left\| \frac{\boldsymbol{x}_0}{\|\boldsymbol{x}_0\|} \right\|_1 |q_1| + \zeta \frac{1 - |\overline{q}_1| |q_1|}{\|\overline{\boldsymbol{q}}\|}, \quad \text{s.t. } |q_1| \leq \frac{1}{|\overline{q}_1|}. \tag{7.6.7}$$

Notice that the problem in (7.6.7) is linear in $|q_1|$ with a compact feasible set. Since the objective is also monotonic in $|q_1|$, it indicates that the optimal solution only occurs at the boundary points $|q_1| = 0$ or $|q_1| = 1/|\overline{q}_1|$ Therefore, $\boldsymbol{q} = \boldsymbol{e}_1/\overline{q}_1$ is the UOS of (7.6.7) if and only if

$$\frac{1}{|\overline{q}_1|} \left\| \frac{\boldsymbol{x}_0}{\|\boldsymbol{x}_0\|} \right\|_1 < \frac{\zeta}{\|\overline{\boldsymbol{q}}_2\|}.$$

Since $\left\| \frac{\boldsymbol{x}_0}{\|\boldsymbol{x}_0\|} \right\|_1 \leq \sqrt{2\theta p}$ conditioned on $\mathcal{E}_0$, it is sufficient to have

$$\frac{\sqrt{2\theta p}}{2\sqrt{\theta}} \leq \zeta = \frac{24}{25}\sqrt{\frac{2}{\pi}}\sqrt{p}\left(1 - \frac{9}{2}\theta - \frac{25}{6}\sqrt{\frac{\pi}{2}}\sqrt{\frac{n}{p}} - \frac{175}{24}\sqrt{\frac{\pi}{2}}\sqrt{\frac{\log(2p)}{p}}\right).$$

Therefore there exists a constant $\theta_0 > 0$, such that whenever $\theta \leq \theta_0$ and $p \geq C_3(\theta_0)n$, the rounding returns $\boldsymbol{e}_1/\overline{q}_1$. A bit of thought suggests one can take a universal C_3 for all possible choice of θ_0, completing the proof. ∎

When the input basis is $\widehat{\boldsymbol{Y}} = \boldsymbol{Y}\boldsymbol{U}$ for some orthogonal matrix $\boldsymbol{U} \neq \boldsymbol{I}$, if the ADM algorithm produces some $\overline{\boldsymbol{q}} = \boldsymbol{U}^\top \boldsymbol{q}'$, such that $q_1' > 2\sqrt{\theta}$. It is not hard to see that now the rounding (7.6.1) is equivalent to

$$\min_{\boldsymbol{q}} \|\boldsymbol{Y}\boldsymbol{U}\boldsymbol{q}\|_1, \quad \text{s.t. } \langle \boldsymbol{q}', \boldsymbol{U}\boldsymbol{q} \rangle = 1.$$

Renaming $\boldsymbol{U}\boldsymbol{q}$, it follows from the above argument that at optimum $\boldsymbol{q}_\star$ it holds that $\boldsymbol{U}\boldsymbol{q}_\star = \gamma \boldsymbol{e}_1$ for some constant γ with high probability.

Part III

Complete Dictionary Learning

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 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 results based on efficient algorithms either only guarantee recovery when $\boldsymbol{X}_0$ has $O(\sqrt{n})$ zeros per column, or require multiple rounds of SDP relaxation to work when $\boldsymbol{X}_0$ has $O(n)$ nonzeros per column.

Our algorithmic pipeline centers around solving a certain nonconvex optimization problem with a spherical constraint. In this paper, we provide a geometric characterization of the objective landscape. In particular, we show that the problem is highly structured: with high probability, (1) there are no "spurious" local minimizers; and (2) around all saddle points the objective has a negative directional curvature. This distinctive structure makes the problem amenable to efficient optimization algorithms. We design a second-order trust-region algorithm over the sphere that provably converges to a local minimizer from arbitrary initializations, despite the presence of saddle points.

This part is organized as follows. In Chapter 8 we motivate the dictionary learning problem and overview main ingredients of our nonconvex approach. In Chapter 9 we present the nonconvex formulation of the complete dictionary learning problem. In Chapter 10 we present our main geometric results that confirm the central nonconvex problem is a ridable saddle function. In Chapter 11 we present the results for convergence of the Riemannian trust-region algorithm over the sphere. In Chapter 12, we present simulations on both synthetic and real data to corroborate our theory. Finally, we conclude this part of thesis in Chapter 13.

All the detailed proofs are omitted in this part of thesis. We refer the readers to our paper [SQW15b] and [SQW15c] for detailed proofs.

Chapter 8

Introduction

Given p signal samples from $\mathbb{R}^n$, i.e., $\boldsymbol{Y} \doteq [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p]$, is it possible to construct a "dictionary" $\boldsymbol{A} \doteq [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m]$ 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 compression, and also prove useful to 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 successful; 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 basis 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. [1]

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

[1] As Donoho et al [DVDD98] put it, "...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."

classical image processing, visual recognition, compressive signal acquisition, and also recent deep architectures for signal classification (see, e.g., [Ela10, MBP14] for review of this development).

8.1 Theoretical and Algorithmic Challenges

In contrast to the above empirical successes, theoretical study of dictionary learning is still developing. For applications in which dictionary learning 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 8.4, 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 $\boldsymbol{A} \in \mathbb{R}^{n\times m}$ and coefficients $\boldsymbol{X} \in \mathbb{R}^{m\times p}$ which best trade-off sparsity and fidelity to the observed data:

$$\text{minimize}_{\boldsymbol{A},\boldsymbol{X}} \quad \lambda \left\|\boldsymbol{X}\right\|_1 + \frac{1}{2}\left\|\boldsymbol{A}\boldsymbol{X} - \boldsymbol{Y}\right\|_F^2, \quad \text{subject to} \quad \boldsymbol{A} \in \mathcal{A}. \tag{8.1.1}$$

Here, $\left\|\boldsymbol{X}\right\|_1 \doteq \sum_{i,j} |X_{ij}|$ 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)[2], while the most daunting nonconvexity comes from the bilinear mapping: $(\boldsymbol{A}, \boldsymbol{X}) \mapsto \boldsymbol{A}\boldsymbol{X}$. 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 conceptual formulation (8.1.1), where $\boldsymbol{\Pi}$ is any permutation matrix, and $\boldsymbol{\Sigma}$ any diagonal matrix with diagonal entries in $\{\pm 1\}$, and $(\cdot)^*$ denotes matrix transpose. Thus, we should expect the problem to have combinatorially many global minimizers. These global minimizers are generally isolated, likely jeopardizing natural convex relaxation (see similar discussions in, e.g., [GS10] and [GW11]).[3] This contrasts sharply with problems in sparse recovery and compressed sensing, in which simple convex relaxations are often provably effective

[2]For example, in nonlinear approximation and harmonic analysis, orthonormal basis or (tight-)frames are preferred; to fix the scale ambiguity discussed in the text, a common practice is to require that $\boldsymbol{A}$ to be column-normalized.

[3]Simple convex relaxations normally replace the objective function with a convex surrogate, and the constraint set with its convex hull. When there are multiple isolated global minimizers for the original nonconvex problem, any point in the convex hull of these global minimizers are necessarily feasible for the relaxed version, and such points tend to produce smaller or equal values than that of the original global minimizers by the relaxed objective function, due to convexity. This implies such relaxations are bound to be loose. Semidefinite programming (SDP) lifting may be one useful general strategy to convexify bilinear inverse problems, see, e.g., [ARR14, CM14a]. However, for problems with general nonlinear constraints, it is unclear whether the lifting always yields tight relaxation; consider, e.g., [BKS13a, BR14, CM14a] and the identification issue in blind deconvolution [**? ?**].

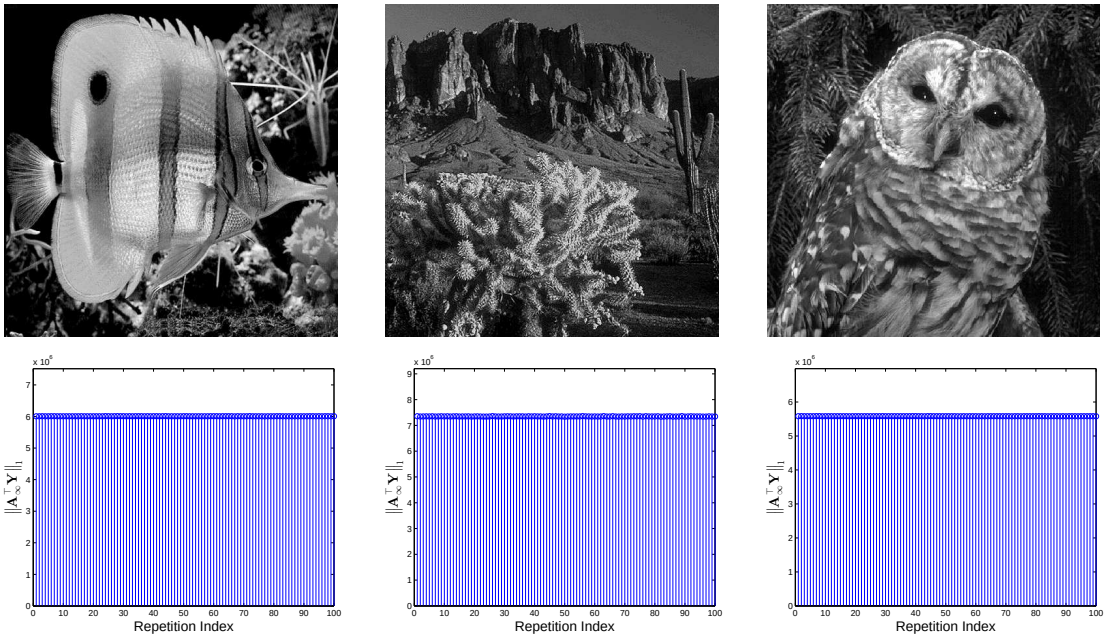


Figure 8.1: Alternating direction method for (8.2.1) on uncompressed real images seems to always produce the same solution! Top: Each image is 512×512 in resolution and encoded in the uncompressed `pgm` format (uncompressed images to prevent possible bias towards standard bases used for compression, 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 (8.2.1) 100 times with independent and randomized (uniform over the orthogonal group) initialization $\boldsymbol{A}_0$. Let $\boldsymbol{A}_\infty$ denote the value of $\boldsymbol{A}$ at convergence (we set the maximally allowable number of ADM iterations to be 10^4 and $\lambda = 2$). 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}!

[DT09, OH10, CLMW11b, DGM13, MT14, MHWG13, CRPW12, CSV13, ALMT14, Can14]. Is there any hope to obtain global solutions to the DL problem?

8.2 An Intriguing Numerical Experiment with Real Images

We provide empirical evidence in support of a positive answer to the above question. Specifically, we learn orthogonal bases (orthobases) for real images patches. Orthobases are of interest because typical hand-designed dictionaries such as discrete cosine (DCT) and wavelet bases are orthogonal, and orthobases seem competitive in performance for applications such as image denoising, as compared to overcomplete dictionaries [BCJ13][4].

We divide a given greyscale image into 8×8 non-overlapping patches, which are converted into 64-dimensional vectors and stacked column-wise into a data matrix $\boldsymbol{Y}$. Specializing (8.1.1) to this setting, we

[4]See Section 8.3 for more detailed discussions of this point. [LGBB05] also gave motivations and algorithms for learning (union of) orthobases as dictionaries.

obtain the optimization problem:

$$\begin{aligned} \text{minimize}_{\boldsymbol{A},\boldsymbol{X}} \quad & \lambda \|\boldsymbol{X}\|_1 + \frac{1}{2}\|\boldsymbol{A}\boldsymbol{X} - \boldsymbol{Y}\|_F^2, \\ \text{subject to} \quad & \boldsymbol{A} \in O_n, \end{aligned} \tag{8.2.1}$$

where O_n is the set of order n orthogonal matrices, i.e., order-n orthogonal group. To derive a concrete algorithm for (8.2.1), one can deploy the alternating direction method (ADM)[5], i.e., alternately minimizing the objective function with respect to (w.r.t.) one variable while fixing the other. The iteration sequence actually takes very simple form: for $k = 1, 2, 3, \ldots,$

$$\boldsymbol{X}_k = \mathcal{S}_\lambda\left[\boldsymbol{A}_{k-1}^* \boldsymbol{Y}\right], \qquad \boldsymbol{A}_k = \boldsymbol{U}\boldsymbol{V}^*$$

where $\mathcal{S}_\lambda[\cdot]$ denotes the well-known soft-thresholding operator acting elementwise on matrices, i.e., $\mathcal{S}_\lambda[x] \doteq \text{sign}(x)\max(|x| - \lambda, 0)$ for any scalar x, and $\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^* = \texttt{SVD}(\boldsymbol{Y}\boldsymbol{X}_k^*)$.

Fig. 8.1 shows what we obtained using the simple ADM algorithm, with *independent and randomized initializations*: *The algorithm seems to always produce the same optimal value, regardless of the initialization.*

This observation is consistent with the possibility that the heuristic ADM algorithm may *always converge to a global minimizer*! [6] Equally surprising is that the phenomenon has been observed on real images[7]. One may imagine only random data typically have "favorable" structures; in fact, almost all existing theories for DL pertain only to random data [SWW12a, AAJ+13, AGM13, AAN13, ABGM14, AGMM15].

8.3 Dictionary Recovery and Our Results

In this thesis, we take a step towards explaining the surprising effectiveness of nonconvex optimization heuristics 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:

[5] 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.

[6] Technically, the convergence to global solutions is surprising because even convergence of ADM to critical points is not guaranteed in general, see, e.g., [ABRS10, BST14] and references therein.

[7] Actually the same phenomenon is also observed for simulated data when the coefficient matrix obeys the Bernoulli-Gaussian model, which is defined later. The result on real images supports that previously claimed empirical successes over two decades may be non-incidental.

- 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. [8]

- The coefficient matrix $\boldsymbol{X}_0$ follows the Bernoulli-Gaussian (BG) model with rate θ: $[X_0]_{ij} = \Omega_{ij} V_{ij}$, with $\Omega_{ij} \sim \mathrm{Ber}(\theta)$ and $V_{ij} \sim \mathcal{N}(0,1)$, where all the different random variables are jointly independent. We write compactly $\boldsymbol{X}_0 \sim_{i.i.d.} \mathrm{BG}(\theta)$. This BG model, or the Bernoulli-Subgaussian model as used in [SWW12a], is a reasonable first model for generic sparse coefficients: the Bernoulli process enables explicit control on the (hard) sparsity level, and the (sub)-Gaussian process seems plausible for modeling variations in magnitudes. Real signals may admit encoding coefficients with additional or different characteristics. We will focus on generic sparse encoding coefficients as a first step towards theoretical understanding.

In this paper, we provide a nonconvex formulation for the DR problem, and characterize the geometric structure of the formulation that allows development of efficient algorithms for optimization. In the companion paper [SQW15c], we derive an efficient algorithm taking advantage of the structure, and describe a complete algorithmic pipeline for efficient recovery. Together, we prove the following result:

> **Theorem 8.1 (Informal statement of our results, a detailed version included in [SQW15c])** *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 (up to sign, scale, and permutation) $\boldsymbol{A}_0$ 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 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.). Ensuring 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 an efficient algorithm is possible. The parameter θ controls the sparsity level of $\boldsymbol{X}_0$. Intuitively, the recovery problem

[8] 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 a 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. See Section **??** for relevant discussion.

is easy for small θ and becomes harder for large θ.[9] 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 DL, our result gives the first efficient algorithm that provably recovers complete $\boldsymbol{A}_0$ and $\boldsymbol{X}_0$ when $\boldsymbol{X}_0$ has $O(n)$ nonzeros per column under appropriate probability model. Section 8.4 provides detailed comparison of our result with other recent recovery results for complete and overcomplete dictionaries.

8.4 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 dictionary learning, 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 optimizer of natural recovery criteria. 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.

[SWW12a] 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 (under the Bernoulli-Subgaussian model) with $O(\sqrt{n})$ nonzeros per column (and the method provably breaks down when $\boldsymbol{X}_0$ contains slightly more than $\Omega(\sqrt{n})$ nonzeros per column). [AAJ+13, AAN13] and [AGM13, AGMM15] gave efficient algorithms that provably recover overcomplete ($m \geq n$), incoherent dictionaries, based on a combination of {clustering

[9] 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 $\boldsymbol{A}_0$.

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] provided 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 (quasipolynomial) time when the sparsity level goes up to $O(n)$. Similarly, [ABGM14] also proposed a super-polynomial time algorithm that guarantees recovery with (almost) $O(n)$ nonzeros per column. Detailed models for those methods dealing with overcomplete dictionaries all differ from one another; nevertheless, they all assume each column of $\boldsymbol{X}_0$ has bounded sparsity levels, and the nonzero coefficients have certain sub-Gaussian magnitudes[11]. 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, under the BG model. After our initial submission, the very recent work [MSS16] assumed the same model as in [BKS14] and provided the first polynomial-time algorithm that guarantees to recover overcomplete dictionaries when the coefficients have up to near sparsity. The improvement is based on a refinement to the rounding procedure for the SOS proposed in [BKS14].

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 [SWW12a] 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[12], [DH14] showed that solving a sequence of linear programs similar to [SWW12a] can recover sparse vectors with sparsity up to $O(p/\sqrt{n})$, sublinear in the vector dimension. The work in Part II of this thesis improved the recovery limit to $O(p)$ by solving a nonconvex sphere-constrained problem similar to (9.0.3)[13] via an ADM 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 (9.0.3) here and that employed in Chapter II.

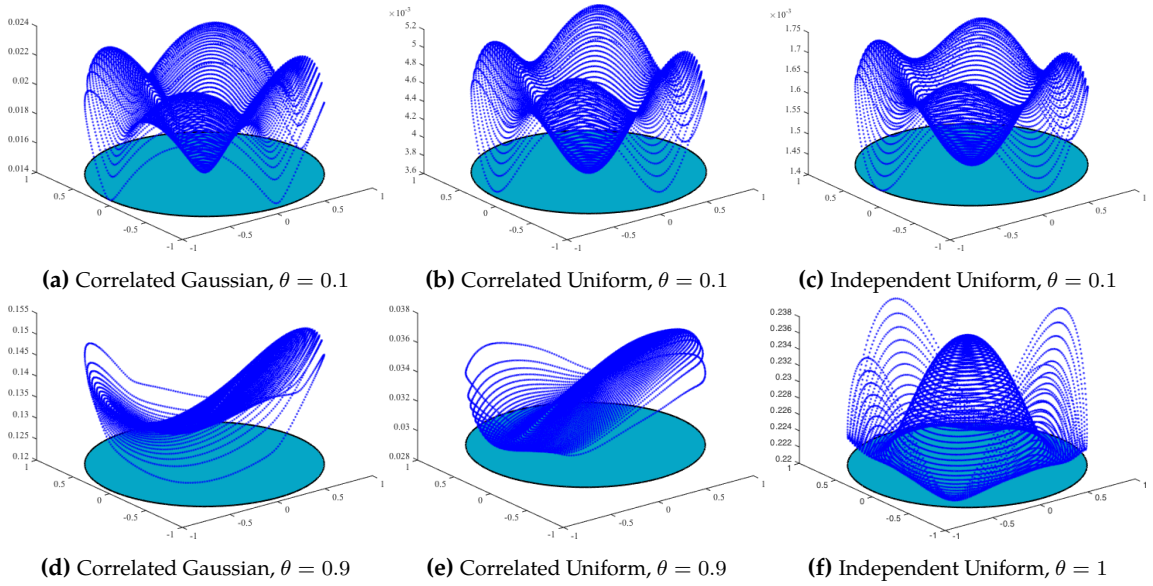
[10] The $\widetilde{O}$ suppresses some logarithm factors.

[11] Thus, one may anticipate that the performances of those methods do not change much qualitatively, if the BG model for the coefficients had been assumed.

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

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

Nonconvex Optimization Problems For other nonconvex optimization problems of recovery of structured signals[14], 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 and optimization [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 property of the optimizing objective to ensure that the local refinement succeeds.[15] By comparison, we provide a complete characterization of the global geometry, which admits efficient algorithms without any special initialization.



(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 8.2: Asymptotic function landscapes when rows of $\boldsymbol{X}_0$ are not independent. 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 $\boldsymbol{v}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}^2)$ 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 $\log\cosh$ function as in (9.0.3).

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 fac-

[14] 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].

[15] 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.

torization (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, SA14, 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 $\boldsymbol{X}$ achieve maximal statistical independence [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]. Algorithms and analysis with another popular contrast function, the fourth-order cumulants, however, indeed overlap with ours considerably [FJK96, AGMS12][16]. While this interesting connection potentially helps port our analysis to ICA, it is a fundamental question to ask what is playing a more vital role for DR, sparsity or independence.

Fig. 8.2 helps shed some light in this direction, where we again plot the asymptotic objective landscape with the natural reparameterization as in Section 10. From the left and central panels, it is evident that even without independence, $\boldsymbol{X}_0$ with sparse columns induces the familiar geometric structures we saw in Fig. 10.1; 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[17], it seems that 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.[18] This suggests that sparsity, instead of independence, makes our current algorithm for DR work.

Nonconvex Problems with Similar Geometric Structure Besides ICA discussed above, it turns out that a handful of other practical problems arising in signal processing and machine learning induce the "no spurious minimizers, all saddles are second-order" structure under natural setting, including the eigenvalue

[16] 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.

[17] 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.

[18] For solving the ICA problem, this suggests the $\log\cosh$ contrast function, that works well empirically [Hyv99], may not work for all distributions (rotation-invariant Gaussian excluded of course), at least when one does not process the data (say perform certain whitening or scaling).

problem, generalized phase retrieval [SQW16], orthogonal tensor decomposition [GHJY15], low-rank matrix recovery/completion [BNS16, GLM16], noisy phase synchronization and community detection [BVB16, Bou16, BBV16], linear neural nets learning [BH89, Kaw16, SC16]. [SQW15d] gave a review of these problems, and discussed how the methodology developed in this and the companion paper [SQW15c] can be generalized to solve those problems.

Chapter 9

Nonconvex Problem 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 fact, for $\boldsymbol{X}_0 \sim_{i.i.d.} \mathrm{BG}\,(\theta)$, rows of $\boldsymbol{X}_0$ are the n *sparsest* vectors (directions) in $\mathrm{row}\,(\boldsymbol{Y})$ w.h.p. whenever $p \geq \Omega\,(n \log n)$ [SWW12a]. Thus, recovering rows of $\boldsymbol{X}_0$ is equivalent to finding the sparsest vectors/directions (due to the scale ambiguity) in $\mathrm{row}(\boldsymbol{Y})$. Since any vector in $\mathrm{row}(\boldsymbol{Y})$ can be written as $\boldsymbol{q}^* \boldsymbol{Y}$ for a certain $\boldsymbol{q}$, one might try to solve

$$\text{minimize } \|\boldsymbol{q}^* \boldsymbol{Y}\|_0 \quad \text{subject to } \quad \boldsymbol{q}^* \boldsymbol{Y} \neq \boldsymbol{0} \tag{9.0.1}$$

to find the sparsest vector in $\mathrm{row}(\boldsymbol{Y})$. Once the sparsest one is found, one then appropriately reduces the subspace $\mathrm{row}(\boldsymbol{Y})$ by one dimension, and solves an analogous version of (9.0.1) to find the second sparsest vector. The process is continued recursively until all sparse vectors are obtained. The above idea of reducing the original recovery problem into finding sparsest vectors in a known subspace first appeared in [SWW12a].

The objective is discontinuous, and the domain is an open set. In particular, the homogeneous 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 [SWW12a, DH14] fix the scale by setting $\|\boldsymbol{q}^* \boldsymbol{Y}\|_\infty = 1$ and use $\|\cdot\|_1$ as a surrogate to $\|\cdot\|_0$, where $\|\cdot\|_\infty$ is the elementwise ℓ^∞ norm, leading to the optimization problem

$$\text{minimize } \|\boldsymbol{q}^* \boldsymbol{Y}\|_1 \quad \text{subject to } \quad \|\boldsymbol{q}^* \boldsymbol{Y}\|_\infty = 1. \tag{9.0.2}$$

The constraint means at least one coordinate of $\boldsymbol{q}^* \boldsymbol{Y}$ has unit magnitude[1]. Thus, (9.0.2) reduces to a sequence of convex (linear) programs. [SWW12a] has shown that (see also [DH14]) solving (9.0.2) recovers $(\boldsymbol{A}_0, \boldsymbol{X}_0)$ for very sparse $\boldsymbol{X}_0$, but the idea provably breaks down when θ is slightly above $O(1/\sqrt{n})$, or equivalently when each column of $\boldsymbol{X}_0$ has more than $O(\sqrt{n})$ nonzeros. Inspired by our previous image experiment, we work with a *nonconvex* alternative[2]:

$$\text{minimize } f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) \doteq \frac{1}{p} \sum_{k=1}^{p} h_\mu \left(\boldsymbol{q}^* \widehat{\boldsymbol{y}}_k \right), \text{ subject to } \|\boldsymbol{q}\| = 1, \tag{9.0.3}$$

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 $\|\cdot\|$ is the usual ℓ^2 norm for vectors. Here $h_\mu(\cdot)$ is chosen to be a convex smooth approximation to $|\cdot|$, namely,

$$h_\mu(z) = \mu \log \left(\frac{e^{z/\mu} + e^{-z/\mu}}{2} \right) = \mu \log \cosh(z/\mu), \tag{9.0.4}$$

which is infinitely differentiable and μ controls the smoothing level.[3] An illustration of the $h_\mu(\cdot)$ function

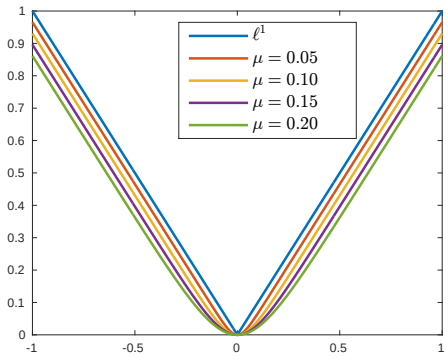


Figure 9.1: The smooth ℓ^1 surrogate defined in (9.0.4) vs. the ℓ^1 function, for varying values of μ. The surrogate approximates the ℓ^1 function more closely when μ gets smaller.

vs. the ℓ^1 function is provided in Fig. 9.1. The spherical constraint is nonconvex. Hence, a-priori, it is unclear whether (9.0.3) admits efficient algorithms that attain global optima. Surprisingly, simple descent algorithms for (9.0.3) exhibit very striking behavior: on many practical numerical examples[4], they appear to produce global solutions. Our next section will uncover interesting geometrical structures underlying the phenomenon.

[1] The sign ambiguity is tolerable here.

[2] A similar formulation has been proposed in [ZP01] in the context of blind source separation; see also Chapter II.

[3] In fact, there is nothing special about this choice and we believe that any valid smooth (twice continuously differentiable) approximation to $|\cdot|$ 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.

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

Chapter 10

The High-dimensional Function Landscape

For the moment, suppose $\boldsymbol{A}_0 = \boldsymbol{I}$ and take $\widehat{\boldsymbol{Y}} = \boldsymbol{Y} = \boldsymbol{A}_0\boldsymbol{X}_0 = \boldsymbol{X}_0$ in (9.0.3). Fig. 10.1 (left) plots $\mathbb{E}_{\boldsymbol{X}_0}\left[f\left(\boldsymbol{q};\boldsymbol{X}_0\right)\right]$ over $\boldsymbol{q} \in \mathbb{S}^2$ ($n = 3$). Remarkably, $\mathbb{E}_{\boldsymbol{X}_0}\left[f\left(\boldsymbol{q};\boldsymbol{X}_0\right)\right]$ has no spurious local minimizers. In fact, every local minimizer $\widehat{\boldsymbol{q}}$ is one of the signed standard basis vectors, i.e., $\pm\boldsymbol{e}_i$'s where $i \in \{1,2,3\}$. Hence, $\widehat{\boldsymbol{q}}^*\boldsymbol{Y}$ reproduces a certain row of $\boldsymbol{X}_0$, and all minimizers reproduce all rows of $\boldsymbol{X}_0$.

Let $\boldsymbol{e}_3^\perp$ be the equatorial section that is orthogonal to $\boldsymbol{e}_3$, i.e., $\boldsymbol{e}_3^\perp \doteq \operatorname{span}(\boldsymbol{e}_1,\boldsymbol{e}_2) \cap \mathbb{B}^3$. To better illustrate the above point, we project the upper hemisphere above $\boldsymbol{e}_3^\perp$ 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. Fig. 10.1 (right) 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 strong gradient region, and a region where at each point one can always find a direction of negative curvature. 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 a global minimizer, irrespective of initialization.

Two challenges stand out when implementing this idea. For geometry, one has to show similar structure 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 11, we will describe a Riemannian trust region method which

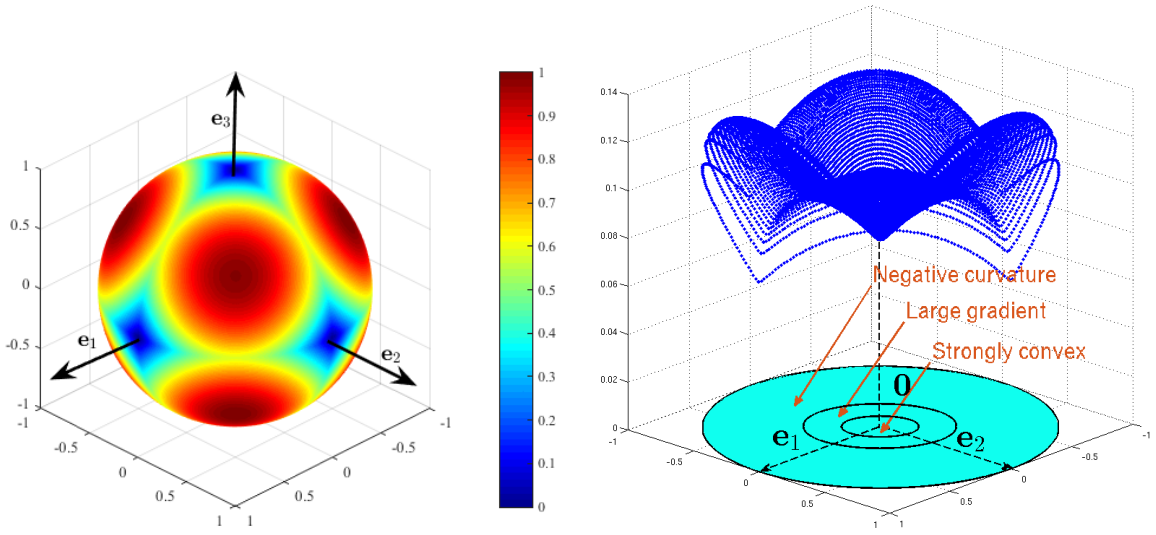


Figure 10.1: Why is dictionary learning over $\mathbb{S}^{n-1}$ tractable? Assume the target dictionary $\boldsymbol{A}_0 = \boldsymbol{I}$. **Left:** Large sample objective function $\mathbb{E}_{\boldsymbol{X}_0}[f(\boldsymbol{q})]$. The only local minimizers are the signed basis vectors $\pm\boldsymbol{e}_i$. **Right:** A visualization of the function as a height above the equatorial section $\boldsymbol{e}_3^\perp$, i.e., $\mathrm{span}\{\boldsymbol{e}_1, \boldsymbol{e}_2\} \cap \mathbb{B}^3$. The derived function is obtained by assigning values of points on the upper hemisphere to their corresponding projections on the equatorial section $\boldsymbol{e}_3^\perp$. The minimizers for the derived function are $\boldsymbol{0}, \pm\boldsymbol{e}_1, \pm\boldsymbol{e}_2$. Around $\boldsymbol{0}$ in $\boldsymbol{e}_3^\perp$, the function exhibits a small region of strong convexity, a region of large gradient, and finally a region in which the direction away from $\boldsymbol{0}$ is a direction of negative curvature.

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 $f(\boldsymbol{q}; \boldsymbol{A}_0\boldsymbol{X}_0) = f(\boldsymbol{A}_0^*\boldsymbol{q}; \boldsymbol{X}_0)$, the landscape of $f(\boldsymbol{q}; \boldsymbol{A}_0\boldsymbol{X}_0)$ is simply a rotated version of that of $f(\boldsymbol{q}; \boldsymbol{X}_0)$, 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 $\pm\boldsymbol{e}_1, \ldots, \pm\boldsymbol{e}_n$, we work with the symmetric section around $\boldsymbol{e}_n$ as an exemplar. An illustration of the symmetric sections and the exemplar we choose to work with on $\mathbb{S}^2$ is provided in Fig. 10.2. The result will carry over to all sections with the same argument; together this provides a complete characterization of the function $f(\boldsymbol{q}; \boldsymbol{X}_0)$ over $\mathbb{S}^{n-1}$.

To study the function on this exemplar region, we again invoke the projection trick described above, this time onto the equatorial section $\boldsymbol{e}_n^\perp$. This can be formally captured by the reparameterization mapping:

$$\boldsymbol{q}(\boldsymbol{w}) = \left(\boldsymbol{w}, \sqrt{1 - \|\boldsymbol{w}\|^2}\right), \ \boldsymbol{w} \in \mathbb{B}^{n-1}, \tag{10.0.1}$$

where $\boldsymbol{w}$ is the new variable and $\mathbb{B}^{n-1}$ is the unit ball in $\mathbb{R}^{n-1}$. We first study the composition $g(\boldsymbol{w}; \boldsymbol{X}_0) \doteq$

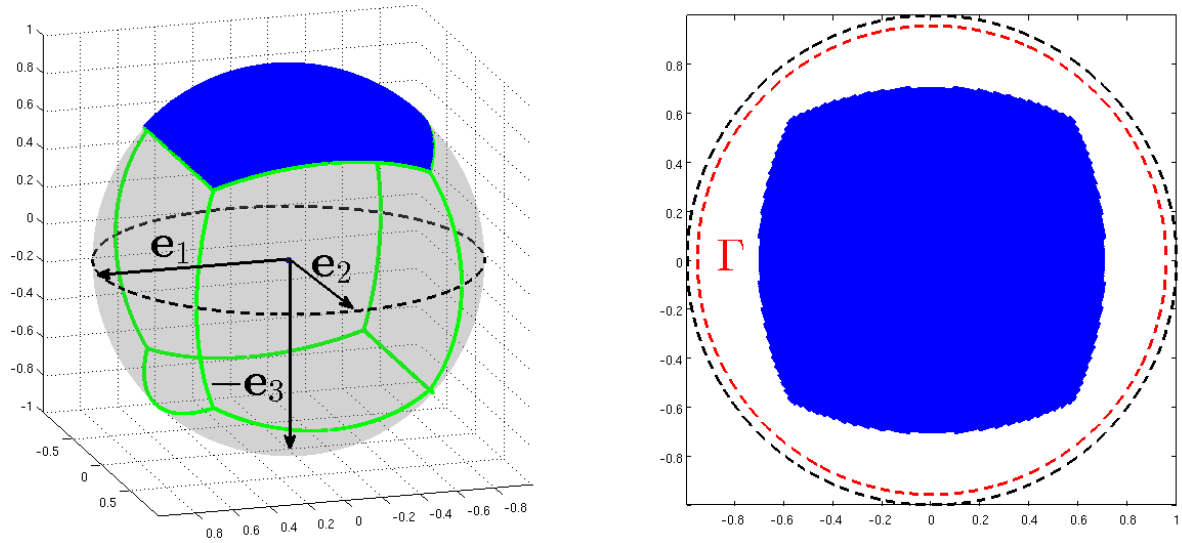


Figure 10.2: Illustration of the six symmetric sections on $\mathbb{S}^2$ and the exemplar we work with. **Left:** The six symmetric sections on $\mathbb{S}^2$, as divided by the green curves. The signed basis vectors, $\pm \boldsymbol{e}_i$'s, are centers of these sections. We choose to work with the exemplar that is centered around $\boldsymbol{e}_3$ that is shaded in blue. **Right:** Projection of the upper hemisphere onto the equatorial section $\boldsymbol{e}_3^{\perp}$. The blue region is projection of the exemplar under study. The larger region enclosed by the red circle is the Γ set on which we characterize the reparametrized function g.

$f\left(\boldsymbol{q}\left(\boldsymbol{w}\right);\boldsymbol{X}_0\right)$ over the set

$$\Gamma \doteq \left\{\boldsymbol{w} : \|\boldsymbol{w}\| < \sqrt{\tfrac{4n-1}{4n}}\right\} \subsetneq \mathbb{B}^{n-1}. \tag{10.0.2}$$

It can be verified the exemplar we chose to work with is strictly contained in this set[1]. This is illustrated for the case $n = 3$ in Fig. 10.2 (right). Our analysis characterizes the properties of $g\left(\boldsymbol{w};\boldsymbol{X}_0\right)$ by studying three quantities

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

respectively over three consecutive regions moving away from the origin, corresponding to the three regions in Fig. 10.1 (right). In particular, through typical expectation-concentration style arguments, we show that there exists a positive constant c such that

$$\nabla^2 g\left(\boldsymbol{w};\boldsymbol{X}_0\right) \succeq \frac{1}{\mu} c\theta \boldsymbol{I}, \quad \frac{\boldsymbol{w}^*\nabla g\left(\boldsymbol{w};\boldsymbol{X}_0\right)}{\|\boldsymbol{w}\|} \geq c\theta, \quad \frac{\boldsymbol{w}^*\nabla^2 g\left(\boldsymbol{w};\boldsymbol{X}_0\right)\boldsymbol{w}}{\|\boldsymbol{w}\|^2} \leq -c\theta \tag{10.0.3}$$

[1]Indeed, if $\langle \boldsymbol{q}, \boldsymbol{e}_n\rangle \geq |\langle \boldsymbol{q}, \boldsymbol{e}_i\rangle|$ for all $i \neq n$, $1 - \|\boldsymbol{w}\|^2 = q_n^2 \geq 1/n$, implying $\|\boldsymbol{w}\|^2 \leq \frac{n-1}{n} < \frac{4n-1}{4n}$. The reason we have defined an open set instead of a closed (compact) one is to avoid potential trivial local minimizers located on the boundary. We study behavior of g over this slightly larger set Γ, instead of just the projection of the chosen symmetric section, to conveniently deal with the boundary effect: if we choose to work with just projection of the chosen symmetric section, there would be considerable technical subtleties at the boundaries when we call the union argument to cover the whole sphere.

over the respective regions w.h.p., confirming our low-dimensional observations described above. In particular, 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 $g(\boldsymbol{w}; \boldsymbol{X}_0)$ over Γ is very close to $\boldsymbol{0}$, within a distance of $O(\mu)$[2]. More specifically, our result can be stated as follows.

Theorem 10.1 (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 constants $c_\star$ and C, such that for any $\theta \in (0, 1/2)$ and $\mu < c_a \min\left\{\theta n^{-1}, n^{-5/4}\right\}$, whenever*

$$p \geq \frac{C}{\mu^2\theta^2} n^3 \log\frac{n}{\mu\theta}, \tag{10.0.4}$$

the following hold simultaneously with probability at least $1 - c_b p^{-6}$:

$$\nabla^2 g(\boldsymbol{w}; \boldsymbol{X}_0) \succeq \frac{c_\star\theta}{\mu}\boldsymbol{I}, \quad \text{if } \|\boldsymbol{w}\| \leq \frac{\mu}{4\sqrt{2}}, \tag{10.0.5}$$

$$\frac{\boldsymbol{w}^*\nabla g(\boldsymbol{w}; \boldsymbol{X}_0)}{\|\boldsymbol{w}\|} \geq c_\star\theta, \quad \text{if } \frac{\mu}{4\sqrt{2}} \leq \|\boldsymbol{w}\| \leq \frac{1}{20\sqrt{5}} \tag{10.0.6}$$

$$\frac{\boldsymbol{w}^*\nabla^2 g(\boldsymbol{w}; \boldsymbol{X}_0)\boldsymbol{w}}{\|\boldsymbol{w}\|^2} \leq -c_\star\theta, \quad \text{if } \frac{1}{20\sqrt{5}} \leq \|\boldsymbol{w}\| \leq \sqrt{\frac{4n-1}{4n}}, \tag{10.0.7}$$

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*

$$\|\boldsymbol{w}_\star - \boldsymbol{0}\| \;\leq\; \min\left\{\frac{c_c\mu}{\theta}\sqrt{\frac{n\log p}{p}}, \frac{\mu}{16}\right\}. \tag{10.0.8}$$

Here c_a through c_c are all positive constants.

Here $\boldsymbol{q}(\boldsymbol{0}) = \boldsymbol{e}_n$, which exactly recovers the last row of $\boldsymbol{X}_0$, $(\boldsymbol{x}_0)^n$. Though the unique local minimizer $\boldsymbol{w}_\star$ may not be $\boldsymbol{0}$, it is very near to $\boldsymbol{0}$. 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 equatorial section $\boldsymbol{e}_n^\perp$ to $\boldsymbol{e}_i^\perp$. Doing this $2n$ times (and multiplying the failure probability in Theorem 10.1 by $2n$), we obtain a characterization of $f(\boldsymbol{q}; \boldsymbol{X}_0)$ over the entirety of $\mathbb{S}^{n-1}$.[3] The result is captured by the next corollary.

[2]When $p \to \infty$, the local minimizer is exactly $\boldsymbol{0}$; deviation from $\boldsymbol{0}$ that we described is due to finite-sample perturbation. The deviation distance depends both the $h_\mu(\cdot)$ and p; see Theorem 10.1 for example.

[3]In fact, it is possible to pull the very detailed geometry captured in (10.0.5) through (10.0.7) 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.

Corollary 10.2 *Suppose $\boldsymbol{A}_0 = \boldsymbol{I}$ and hence $\boldsymbol{Y} = \boldsymbol{A}_0\boldsymbol{X}_0 = \boldsymbol{X}_0$. There exist positive constant C, such that for any $\theta \in (0, 1/2)$ and $\mu < c_a \min\left\{\theta n^{-1}, n^{-5/4}\right\}$, whenever $p \geq \frac{C}{\mu^2\theta^2} n^3 \log\frac{n}{\mu\theta}$, with probability at least $1 - c_b p^{-5}$, the function $f(\boldsymbol{q}; \boldsymbol{X}_0)$ 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*

$$\|\boldsymbol{q}_\star - \boldsymbol{b}\| \leq \sqrt{2}\min\left\{\frac{c_c\mu}{\theta}\sqrt{\frac{n\log p}{p}}, \frac{\mu}{16}\right\}. \tag{10.0.9}$$

Here c_a to c_c are positive constants.

We refer the readers to [SQW15b] for the detailed proofs of Theorem 10.1 and Corollary 10.2. Though the $2n$ isolated local minimizers may have different objective values, they are equally good in the sense each of them helps produce a close approximation to a certain row of $\boldsymbol{X}_0$. As discussed above, 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.

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 qualitative properties of the objective landscape. Write

$$\overline{\boldsymbol{Y}} = \left(\tfrac{1}{p\theta}\boldsymbol{Y}\boldsymbol{Y}^*\right)^{-1/2}\boldsymbol{Y}. \tag{10.0.10}$$

Note that for $\boldsymbol{X}_0 \sim_{i.i.d.} \mathrm{BG}(\theta)$, $\mathbb{E}\left[\boldsymbol{X}_0\boldsymbol{X}_0^*\right]/(p\theta) = \boldsymbol{I}$. Thus, one expects $\frac{1}{p\theta}\boldsymbol{Y}\boldsymbol{Y}^* = \frac{1}{p\theta}\boldsymbol{A}_0\boldsymbol{X}_0\boldsymbol{X}_0^*\boldsymbol{A}_0^*$ to behave roughly like $\boldsymbol{A}_0\boldsymbol{A}_0^*$ and hence $\overline{\boldsymbol{Y}}$ to behave like

$$\left(\boldsymbol{A}_0\boldsymbol{A}_0^*\right)^{-1/2}\boldsymbol{A}_0\boldsymbol{X}_0 = \left(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^*\right)^{-1/2}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*\boldsymbol{X}_0 = \boldsymbol{U}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^*\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*\boldsymbol{X}_0 = \boldsymbol{U}\boldsymbol{V}^*\boldsymbol{X}_0 \tag{10.0.11}$$

where $\mathtt{SVD}(\boldsymbol{A}_0) = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$. 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

$$\overline{\boldsymbol{Y}} = \boldsymbol{U}\boldsymbol{V}^*\boldsymbol{X}_0 + \boldsymbol{\Xi}\boldsymbol{X}_0, \tag{10.0.12}$$

where $\boldsymbol{\Xi}$ is a matrix with a small magnitude. Simple perturbation argument shows that the constant c in (10.0.3) 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. To characterize the function landscape of $f(\boldsymbol{q}; \boldsymbol{X}_0)$ over $\mathbb{S}^{n-1}$, we mostly work with the function

$$g(\boldsymbol{w}) \doteq f(\boldsymbol{q}(\boldsymbol{w}); \boldsymbol{X}_0) = \frac{1}{p}\sum_{k=1}^{p} h_\mu\left(\boldsymbol{q}(\boldsymbol{w})^* (\boldsymbol{x}_0)_k\right), \tag{10.0.13}$$

induced by the reparametrization

$$\boldsymbol{q}(\boldsymbol{w}) = \left(\boldsymbol{w}, \sqrt{1-\|\boldsymbol{w}\|^2}\right), \quad \boldsymbol{w} \in \mathbb{B}^{n-1}. \tag{10.0.14}$$

In particular, we focus our attention to the smaller set

$$\Gamma = \left\{\boldsymbol{w} : \|\boldsymbol{w}\| < \sqrt{\frac{4n-1}{4n}}\right\} \subsetneq \mathbb{B}^{n-1}, \tag{10.0.15}$$

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 similarly characterize other parts of f on $\mathbb{S}^{n-1}$ using projection onto other equatorial sections. Note that over Γ, $q_n = \sqrt{1-\|\boldsymbol{w}\|^2} \geq 1/(2\sqrt{n})$.

Theorem 10.3 (High-dimensional landscape - complete dictionary) *Suppose $\boldsymbol{A}_0$ is complete with its condition number $\kappa(\boldsymbol{A}_0)$. There exist positive constants $c_\star$ (particularly, the same constant as in Theorem 10.1) and C, such that for any $\theta \in (0, 1/2)$ and $\mu < c_a \min\{\theta n^{-1}, n^{-5/4}\}$, when*

$$p \geq \frac{C}{c_\star^2 \theta^2} \max\left\{\frac{n^4}{\mu^4}, \frac{n^5}{\mu^2}\right\} \kappa^8(\boldsymbol{A}_0) \log^4\left(\frac{\kappa(\boldsymbol{A}_0)\, n}{\mu\theta}\right) \tag{10.0.16}$$

and $\overline{\boldsymbol{Y}} \doteq \sqrt{p\theta}\,(\boldsymbol{Y}\boldsymbol{Y}^*)^{-1/2}\boldsymbol{Y}$, $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = \mathtt{SVD}(\boldsymbol{A}_0)$, the following hold simultaneously with probability at least $1 - c_b p^{-6}$:

$$\nabla^2 g(\boldsymbol{w}; \boldsymbol{V}\boldsymbol{U}^*\overline{\boldsymbol{Y}}) \succeq \frac{c_\star\theta}{2\mu}\boldsymbol{I}, \quad \text{if } \|\boldsymbol{w}\| \leq \frac{\mu}{4\sqrt{2}}, \tag{10.0.17}$$

$$\frac{\boldsymbol{w}^*\nabla g(\boldsymbol{w}; \boldsymbol{V}\boldsymbol{U}^*\overline{\boldsymbol{Y}})}{\|\boldsymbol{w}\|} \geq \frac{1}{2}c_\star\theta, \quad \text{if } \frac{\mu}{4\sqrt{2}} \leq \|\boldsymbol{w}\| \leq \frac{1}{20\sqrt{5}} \tag{10.0.18}$$

$$\frac{\boldsymbol{w}^*\nabla^2 g(\boldsymbol{w}; \boldsymbol{V}\boldsymbol{U}^*\overline{\boldsymbol{Y}})\boldsymbol{w}}{\|\boldsymbol{w}\|^2} \leq -\frac{1}{2}c_\star\theta, \quad \text{if } \frac{1}{20\sqrt{5}} \leq \|\boldsymbol{w}\| \leq \sqrt{\frac{4n-1}{4n}}, \tag{10.0.19}$$

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*

$$\|\boldsymbol{w}_\star - \boldsymbol{0}\| \leq \mu/7. \tag{10.0.20}$$

Here c_a, a_b are both positive constants.

Corollary 10.4 *Suppose $\boldsymbol{A}_0$ is complete with its condition number $\kappa(\boldsymbol{A}_0)$. There exist positive constants $c_\star$ (particularly, the same constant as in Theorem 10.1) and C, such that for any $\theta \in (0, 1/2)$ and $\mu < c_a \min\left\{\theta n^{-1}, n^{-5/4}\right\}$, when $p \geq \frac{C}{c_\star^2 \theta^2} \max\left\{\frac{n^4}{\mu^4}, \frac{n^5}{\mu^2}\right\} \kappa^8(\boldsymbol{A}_0) \log^4\left(\frac{\kappa(\boldsymbol{A}_0) n}{\mu\theta}\right)$ and $\overline{\boldsymbol{Y}} \doteq \sqrt{p\theta}\left(\boldsymbol{Y}\boldsymbol{Y}^*\right)^{-1/2}\boldsymbol{Y}$, $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = \mathtt{SVD}(\boldsymbol{A}_0)$, with probability at least $1 - c_b p^{-5}$, 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*

$$\|\boldsymbol{q}_\star - \boldsymbol{b}\| \leq \sqrt{2}\mu/7. \tag{10.0.21}$$

Here c_a, c_b are both positive constants.

From the above theorems, it is clear that for any saddle point in the $\boldsymbol{w}$ space, the Hessian has at least one negative eigenvalue with an associated eigenvector $\boldsymbol{w}/\|\boldsymbol{w}\|$. Now the question is whether all saddle points of f on $\mathbb{S}^{n-1}$ have analogous properties, we will show in Section 11 that we need to perform actual optimization in the $\boldsymbol{q}$ space. The arguments are put in the language of Riemannian geometry, and we can switch back and forth between $\boldsymbol{q}$ and $\boldsymbol{w}$ spaces in our algorithm analysis without stating this fact.

Chapter 11

Algorithm

To optimize the objective (9.0.3), as we do not know $\boldsymbol{A}_0$ ahead of time, so our algorithm needs to take advantage of the structure described in the previous chapter 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 saddle points with indefinite Hessian, which we call *ridable saddles* [1] (Fig. 10.1). We can use second-order information to guarantee to escape from such saddle points. In this chapter, we derive an algorithm based on the Riemannian trust region method (TRM) [ABG07, AMS09] for solving the complete dictionary learning problem. There are other algorithmic possibilities; see, e.g., [Gol80, GHJY15].

First, let us provide the basic intuition why a local minimizer can be retrieved by the second-order trust-region method. Consider an unconstrained optimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \phi\left(\boldsymbol{x}\right).$$

Typical (second-order) TRM proceeds by successively forming a second-order approximation to ϕ at the current iterate,

$$\widehat{\phi}(\boldsymbol{\delta};\boldsymbol{x}^{(r-1)}) \doteq \phi(\boldsymbol{x}^{(r-1)}) + \nabla^*\phi(\boldsymbol{x}^{(r-1)})\boldsymbol{\delta} + \tfrac{1}{2}\boldsymbol{\delta}^*\boldsymbol{Q}(\boldsymbol{x}^{(r-1)})\boldsymbol{\delta}, \tag{11.0.1}$$

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

[1] See [SQW15d] and [GHJY15].

can efficiently solved:

$$\boldsymbol{\delta}^{(r)} \doteq \underset{\boldsymbol{\delta}\in\mathbb{R}^n, \|\boldsymbol{\delta}\|_p \le \Delta}{\arg\min} \widehat{\phi}(\boldsymbol{\delta}; \boldsymbol{x}^{(r-1)}), \tag{11.0.2}$$

where Δ is called the trust-region radius that controls how far the movement can be made. If we take $\boldsymbol{Q}(\boldsymbol{x}^{(r-1)}) = \nabla^2\phi(\boldsymbol{x}^{(r-1)})$ for all r, then whenever the gradient is nonvanishing or the Hessian is indefinite, we expect to decrease the objective function by a concrete amount provided $\|\boldsymbol{\delta}\|$ is sufficiently small. Since the domain is compact, the iterate sequence ultimately moves into the strongly convex region, where the trust-region algorithm behaves like a typical Newton algorithm. In the following, we generalize those ideas to our objective (9.0.3) over the sphere and make it rigorous. We refer the readers to [SQW15c] for the detailed proofs.

11.1 Finding One Local Minimizer via the Riemannian Trust-Region Method

We are interested to seek a local minimizer of (9.0.3). The presence of saddle points have 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 modeling assumptions, this algorithm with an arbitrary initialization efficiently produces an accurate approximation[2] to one of the minimizers. Throughout the exposition, basic knowledge of Riemannian geometry is assumed. The reader can consult the excellent monograph [AMS09] for relevant background and details.

11.1.1 Some Basic Facts about the Sphere and f

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

$$T_{\boldsymbol{q}}\mathbb{S}^{n-1} = \left\{\boldsymbol{\delta} \in \mathbb{R}^n : \boldsymbol{q}^*\boldsymbol{\delta} = 0\right\},$$

$$\mathcal{P}_{T_{\boldsymbol{q}}\mathbb{S}^{n-1}} = \boldsymbol{I} - \boldsymbol{q}\boldsymbol{q}^* = \boldsymbol{U}\boldsymbol{U}^*,$$

[2]By "accurate" we mean one can achieve an arbitrary numerical accuracy $\varepsilon > 0$ with a reasonable amount of time. Here the running time of the algorithm is on the order of $\log\log(1/\varepsilon)$ in the target accuracy ε, and polynomial in other problem parameters.

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). Consider any $\boldsymbol{\delta} \in T_{\boldsymbol{q}}\mathbb{S}^{n-1}$. The map

$$\gamma(t) : t \mapsto \boldsymbol{q}\cos\left(t\left\|\boldsymbol{\delta}\right\|\right) + \frac{\boldsymbol{\delta}}{\left\|\boldsymbol{\delta}\right\|}\sin\left(t\left\|\boldsymbol{\delta}\right\|\right)$$

defines a smooth curve on the sphere that satisfies $\gamma(0) = \boldsymbol{q}$ and $\dot{\gamma}(0) = \boldsymbol{\delta}$. Geometrically, $\gamma(t)$ is a segment of the great circle that passes $\boldsymbol{q}$ and has $\boldsymbol{\delta}$ as its tangent vector at $\boldsymbol{q}$. The exponential map for $\boldsymbol{\delta}$ is defined as

$$\exp_{\boldsymbol{q}}(\boldsymbol{\delta}) \doteq \gamma(1) = \boldsymbol{q}\cos\left\|\boldsymbol{\delta}\right\| + \frac{\boldsymbol{\delta}}{\left\|\boldsymbol{\delta}\right\|}\sin\left\|\boldsymbol{\delta}\right\|.$$

It is a canonical way of pulling $\boldsymbol{\delta}$ to the sphere.

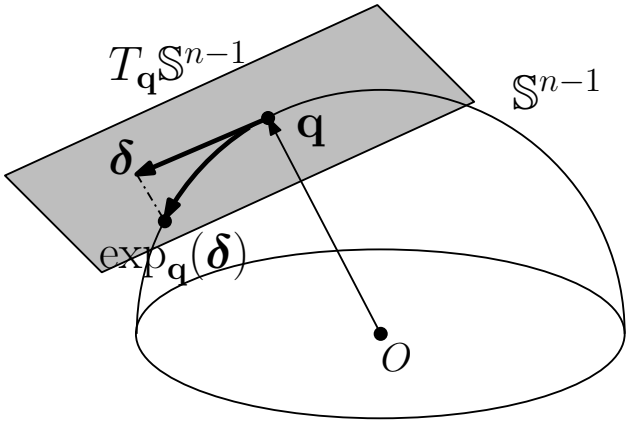


Figure 11.1: Illustrations of the tangent space $T_{\boldsymbol{q}}\mathbb{S}^{n-1}$ and exponential map $\exp_{\boldsymbol{q}}(\boldsymbol{\delta})$ defined on the sphere $\mathbb{S}^{n-1}$.

In this paper we are interested in the restriction of f to the unit sphere $\mathbb{S}^{n-1}$. For the sake of performing optimization, we need local approximations of f. Instead of directly approximating the function in $\mathbb{R}^n$, we form quadratic approximations of f in the tangent spaces of $\mathbb{S}^{n-1}$. We consider the smooth function $f \circ \exp_{\boldsymbol{q}}(\boldsymbol{\delta}) : T_{\boldsymbol{q}}\mathbb{S}^{n-1} \mapsto \mathbb{R}$, where $\circ$ is the usual function composition operator. An applications of vector space Taylor's theorem gives

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

when $\left\|\boldsymbol{\delta}\right\|$ is small. Thus, we form a quadratic approximation $\widehat{f}(\boldsymbol{\delta};\boldsymbol{q}) : T_{\boldsymbol{q}}\mathbb{S}^{n-1} \mapsto \mathbb{R}$ as

$$\widehat{f}(\boldsymbol{\delta};\boldsymbol{q},\widehat{\boldsymbol{Y}}) \doteq f(\boldsymbol{q};\widehat{\boldsymbol{Y}}) + \left\langle \underbrace{\nabla f(\boldsymbol{q};\widehat{\boldsymbol{Y}})}, \boldsymbol{\delta}\right\rangle + \frac{1}{2}\boldsymbol{\delta}^*\left(\underbrace{\nabla^2 f(\boldsymbol{q};\widehat{\boldsymbol{Y}}) - \left\langle \nabla f(\boldsymbol{q};\widehat{\boldsymbol{Y}}), \boldsymbol{q}\right\rangle \boldsymbol{I}}\right)\boldsymbol{\delta}. \quad (11.1.1)$$

Here $\nabla f(\boldsymbol{q})$ 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 (9.0.3), it is easy to check that

$$\nabla f(\boldsymbol{q};\widehat{\boldsymbol{Y}}) = \frac{1}{p}\sum_{k=1}^{p}\tanh\left(\frac{\boldsymbol{q}^*\widehat{\boldsymbol{y}}_k}{\mu}\right)\widehat{\boldsymbol{y}}_k, \tag{11.1.2}$$

$$\nabla^2 f(\boldsymbol{q};\widehat{\boldsymbol{Y}}) = \frac{1}{p}\sum_{k=1}^{p}\frac{1}{\mu}\left[1-\tanh^2\left(\frac{\boldsymbol{q}^*\widehat{\boldsymbol{y}}_k}{\mu}\right)\right]\widehat{\boldsymbol{y}}_k\widehat{\boldsymbol{y}}_k^*. \tag{11.1.3}$$

The quadratic approximation also naturally gives rise to the Riemannian gradient and Riemannian Hessian defined on $T_{\boldsymbol{q}}\mathbb{S}^{n-1}$ as

$$\operatorname{grad} f(\boldsymbol{q};\widehat{\boldsymbol{Y}}) = \mathcal{P}_{T_{\boldsymbol{q}}\mathbb{S}^{n-1}}\nabla f(\boldsymbol{q};\widehat{\boldsymbol{Y}}), \tag{11.1.4}$$

$$\operatorname{Hess} f(\boldsymbol{q};\widehat{\boldsymbol{Y}}) = \quad \mathcal{P}_{T_{\boldsymbol{q}}\mathbb{S}^{n-1}}\left(\nabla^2 f(\boldsymbol{q};\widehat{\boldsymbol{Y}}) - \left\langle \nabla f(\boldsymbol{q};\widehat{\boldsymbol{Y}}), \boldsymbol{q}\right\rangle \boldsymbol{I}\right)\mathcal{P}_{T_{\boldsymbol{q}}\mathbb{S}^{n-1}}. \tag{11.1.5}$$

Thus, the above quadratic approximation can be rewritten compactly as

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

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

$$\operatorname{grad} f(\boldsymbol{q};\widehat{\boldsymbol{Y}}) + \operatorname{Hess} f(\boldsymbol{q};\widehat{\boldsymbol{Y}})\boldsymbol{\delta}_\star = \boldsymbol{0}. \tag{11.1.6}$$

If $\operatorname{Hess} f(\boldsymbol{q};\widehat{\boldsymbol{Y}})$ is positive semidefinite and has "full rank" $n-1$ (hence "nondegenerate"[3]), the unique solution $\boldsymbol{\delta}_\star$ is

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

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

[3]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]):

$$\begin{aligned}\mathcal{P}_{\gamma}^{\tau\leftarrow 0} &= \left(\boldsymbol{I}-\frac{\boldsymbol{\delta}\boldsymbol{\delta}^*}{\|\boldsymbol{\delta}\|^2}\right)-\boldsymbol{q}\sin\left(\tau\|\boldsymbol{\delta}\|\right)\frac{\boldsymbol{\delta}^*}{\|\boldsymbol{\delta}\|}+\frac{\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|}\cos\left(\tau\|\boldsymbol{\delta}\|\right)\frac{\boldsymbol{\delta}^*}{\|\boldsymbol{\delta}\|}\\ &= \boldsymbol{I}+\left(\cos(\tau\|\boldsymbol{\delta}\|)-1\right)\frac{\boldsymbol{\delta}\boldsymbol{\delta}^*}{\|\boldsymbol{\delta}\|^2}-\sin\left(\tau\|\boldsymbol{\delta}\|\right)\frac{\boldsymbol{q}\boldsymbol{\delta}^*}{\|\boldsymbol{\delta}\|}.\end{aligned} \tag{11.1.7}$$

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 $\mathcal{P}_{\gamma}^{\tau\leftarrow 0}$.

11.1.2 The Riemannian Trust-Region Algorithm over the Sphere

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

For our f defined over $\mathbb{S}^{n-1}$, given the previous iterate $\boldsymbol{q}^{(r-1)}$, the TRM produces the next movement by generating a solution $\widehat{\boldsymbol{\delta}}$ to

$$\text{minimize}_{\boldsymbol{\delta}\in T_{\boldsymbol{q}^{(r-1)}}\mathbb{S}^{n-1},\ \|\boldsymbol{\delta}\|\le\Delta} \quad \widehat{f}\left(\boldsymbol{\delta};\boldsymbol{q}^{(r-1)}\right), \tag{11.1.8}$$

where $\widehat{f}\left(\boldsymbol{\delta};\boldsymbol{q}^{(r-1)}\right)$ is the local quadratic approximation defined in (11.1.1). The solution $\widehat{\boldsymbol{\delta}}$ is then pulled back to $\mathbb{S}^{n-1}$ from $T_{\boldsymbol{q}}\mathbb{S}^{n-1}$. If we choose the exponential map to pull back the movement $\widehat{\boldsymbol{\delta}}$,[4] the next iterate then reads

$$\boldsymbol{q}^{(r)} = \boldsymbol{q}^{(r-1)}\cos\|\widehat{\boldsymbol{\delta}}\| + \frac{\widehat{\boldsymbol{\delta}}}{\|\widehat{\boldsymbol{\delta}}\|}\sin\|\widehat{\boldsymbol{\delta}}\|. \tag{11.1.9}$$

To solve the subproblem (11.1.8) numerically, we can take any matrix $\boldsymbol{U} \in \mathbb{R}^{n\times(n-1)}$ whose columns form an orthonormal basis for $T_{\boldsymbol{q}^{(r-1)}}\mathbb{S}^{n-1}$, and produce a solution $\widehat{\boldsymbol{\xi}}$ to

$$\text{minimize}_{\|\boldsymbol{\xi}\|\le\Delta} \quad \widehat{f}\left(\boldsymbol{U}\boldsymbol{\xi};\boldsymbol{q}^{(r-1)}\right). \tag{11.1.10}$$

Solution to (11.1.8) can then be recovered as $\widehat{\boldsymbol{\delta}} = \boldsymbol{U}\widehat{\boldsymbol{\xi}}$.

The problem (11.1.10) is an instance of the classic *trust region subproblem*, i.e., minimizing a quadratic function subject to a single quadratic constraint.Albeit potentially nonconvex, this notable subproblem can be solved in polynomial time by several numerical methods [MS83, CGT00, RW97, YZ03, FW04, HK14]. Approximate solution of the subproblem suffices to guarantee convergence in theory, and lessens the storage

[4] 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].

and computational burden in practice. We will deploy the approximate version in simulations. For simplicity, however, our subsequent analysis assumes the subproblem is solved *exactly*. We next briefly describe how one can deploy the semidefinite programming (SDP) approach [RW97, YZ03, FW04, HK14] to solve the subproblem exactly. This choice is due to the well-known effectiveness and robustness of the SDP approach on this problem. We introduce

$$\tilde{\boldsymbol{\xi}} = [\boldsymbol{\xi}^*, 1]^*\,,\ \boldsymbol{\Theta} = \tilde{\boldsymbol{\xi}}\tilde{\boldsymbol{\xi}}^*,\ \boldsymbol{M} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{b} \\ \boldsymbol{b}^* & 0 \end{bmatrix}, \tag{11.1.11}$$

where $\boldsymbol{A} = \boldsymbol{U}^* \operatorname{Hess} f(\boldsymbol{q}^{(r-1)}; \widehat{\boldsymbol{Y}})\boldsymbol{U}$ and $\boldsymbol{b} = \boldsymbol{U}^* \operatorname{grad} \nabla f(\boldsymbol{q}^{(r-1)}; \widehat{\boldsymbol{Y}})$. The resulting SDP to solve is

$$\text{minimize}_{\,\boldsymbol{\Theta}}\ \langle \boldsymbol{M}, \boldsymbol{\Theta} \rangle\,, \quad \text{subject to } \operatorname{tr}(\boldsymbol{\Theta}) \le \Delta^2 + 1,\ \langle \boldsymbol{E}_{n+1}, \boldsymbol{\Theta} \rangle = 1,\ \boldsymbol{\Theta} \succeq \boldsymbol{0}, \tag{11.1.12}$$

where $\boldsymbol{E}_{n+1} = \boldsymbol{e}_{n+1}\boldsymbol{e}_{n+1}^*$. Once the problem (11.1.12) is solved to its optimum $\boldsymbol{\Theta}_\star$, one can provably recover the minimizer $\boldsymbol{\xi}_\star$ of (11.1.10) by computing the SVD of $\boldsymbol{\Theta}_\star = \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{V}}^*$, and extract as a subvector the first $n-1$ coordinates of the principal eigenvector $\widetilde{\boldsymbol{u}}_1$ (see Appendix B of [BV04]).

Using general convergence results on Riemannian TRM (see, e.g., Chapter 7 of [AMS09]), it is not difficult to prove that the gradient sequence $\operatorname{grad} f(q^{(r)}; \widehat{\boldsymbol{Y}})$ produced by TRM converges to zero (i.e., global convergence), or the sequence converges (at quadratic rate) to a local minimizer if the initialization is already close a local minimizer (i.e., local convergence). In this section, we show that under our probabilistic assumptions, these results can be substantially strengthened. In particular, the algorithm is guaranteed to produce an accurate approximation to a local minimizer of the objective function, in a number of iterations that is polynomial in the problem size, from arbitrary initializations. The arguments in the Chapter 10 showed that w.h.p. every local minimizer of f produces a close approximation to a row of $\boldsymbol{X}_0$. Taken together, this implies that the algorithm efficiently produces a close approximation to one row of $\boldsymbol{X}_0$.

Thorough the analysis, we assume the trust-region subproblem is exactly solved and the step size parameter Δ is fixed. Our next two theorems summarize the convergence results for orthogonal and complete dictionaries, respectively.

Theorem 11.1 (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 $\mu < c_a \min\left\{\theta n^{-1}, n^{-5/4}\right\}$, whenever*

$$p \ge \frac{C}{\mu^2\theta^2} n^3 \log \frac{n}{\mu\theta},$$

with probability at least $1 - c_b p^{-6}$, *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*

$$\Delta \leq \frac{c_c c_\star^3 \theta^3 \mu^2}{n^{7/2} \log^{7/2}(np)} \tag{11.1.13}$$

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.,* $\|\widehat{\boldsymbol{q}} - \boldsymbol{q}_\star\| \leq \varepsilon$*) in at most*

$$\max\left\{\frac{c_d n^6 \log^3(np)}{c_\star^3 \theta^3 \mu^4}, \frac{c_e n}{c_\star^2 \theta^2 \Delta^2}\right\} f(\boldsymbol{q}^{(0)}) + \log\log \frac{c_f c_\star \theta \mu}{\varepsilon n^{3/2} \log^{3/2}(np)} \tag{11.1.14}$$

iterations. Here $c_\star$ *is as defined in Theorem 10.1, and* c_a *through* c_f *are all positive constants.*

Theorem 11.2 (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* $\mu < c_a \min\left\{\theta n^{-1}, n^{-5/4}\right\}$, *whenever*

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

with probability at least $1 - c_b p^{-6}$, *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*

$$\Delta \leq \frac{c_c c_\star^3 \theta^3 \mu^2}{n^{7/2} \log^{7/2}(np)} \tag{11.1.15}$$

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.,* $\|\widehat{\boldsymbol{q}} - \boldsymbol{q}_\star\| \leq \varepsilon$*) in at most*

$$\max\left\{\frac{c_d n^6 \log^3(np)}{c_\star^3 \theta^3 \mu^4}, \frac{c_e n}{c_\star^2 \theta^2 \Delta^2}\right\} f(\boldsymbol{q}^{(0)}) + \log\log \frac{c_f c_\star \theta \mu}{\varepsilon n^{3/2} \log^{3/2}(np)} \tag{11.1.16}$$

iterations. Here $c_\star$ *is as in Theorem 10.1, and* c_a *through* c_f *are all positive constants.*

Our convergence result shows that for any target accuracy $\varepsilon > 0$ the algorithm terminates within polynomially many steps. Specifically, the first summand in (11.1.14) or (11.1.16) is the number of steps the sequence takes to enter the strongly convex region and be "reasonably" close to a local minimizer. All subsequent trust-region subproblems are then unconstrained (proved below) – the constraint is inactive at optimal point, and hence the steps behave like Newton steps. The second summand reflects the typical quadratic local convergence of the Newton steps.

Our estimate of the number of steps is pessimistic: the running time is a relatively high-degree polynomial in p and n. We will discuss practical implementation details that help speed up in Section 5.1. 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.

11.2 Complete Algorithm Pipeline and Main Results

For orthogonal dictionaries, from Theorem 10.1 and Corollary 10.2, 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 a certain $\alpha \neq 0$ and $i \in [n]$; in Theorem 11.1, 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 produce the target $\boldsymbol{q}_\star$. 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 refer the readers to Section III of [SQW15c] for detailed proofs. The same procedure can be used to recover complete dictionaries, though the analysis is slightly more complicated; again, we refer the readers to Section III of [SQW15c] for detailed proofs. 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 10.1 (resp. Theorem 10.3) and Theorem 11.1 (resp. Theorem 11.2), starting from any $\boldsymbol{q} \in \mathbb{S}^{n-1}$, 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 a 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:

$$\text{minimize}_{\boldsymbol{q}} \left\| \boldsymbol{q}^* \widehat{\boldsymbol{Y}} \right\|_1, \quad \text{subject to} \quad \langle \boldsymbol{r}, \boldsymbol{q} \rangle = 1, \tag{11.2.1}$$

   with $\boldsymbol{r} = \widehat{\boldsymbol{q}}$. We show 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 (11.2.1) 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 $\boldsymbol{X}_0$ ($1 \leq \ell \leq n-2$) have been recovered, say, by unit vectors $\boldsymbol{q}_\star^1, \ldots, \boldsymbol{q}_\star^\ell$, one takes an orthonormal basis $\boldsymbol{U}$ for $[\operatorname{span}(\boldsymbol{q}_\star^1, \ldots, \boldsymbol{q}_\star^\ell)]^\perp$, 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 (11.2.1) 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 $\boldsymbol{A}_0 = \boldsymbol{Y}\boldsymbol{X}_0^* \left(\boldsymbol{X}_0\boldsymbol{X}_0^*\right)^{-1}$.

Our recovery result can be summarized as follows.

Theorem 11.3 (Main theorem - recovering orthogonal dictionaries) *Assume the dictionary $\boldsymbol{A}_0$ is orthogonal and we take $\widehat{\boldsymbol{Y}} = \boldsymbol{Y}$. Suppose $\theta \in (0, 1/3)$, $\mu_\star < c_a \min\left\{\theta n^{-1}, n^{-5/4}\right\}$, and $p \geq Cn^3 \log \frac{n}{\mu_\star \theta} / \left(\mu_\star^2 \theta^2\right)$. The above algorithmic pipeline with parameter setting*

$$\Delta_\star = \frac{c_b c_\star^3 \theta^3 \mu_\star^2}{n^{7/2} \log^{7/2}(np)}, \tag{11.2.2}$$

recovers the dictionary $\boldsymbol{A}_0$ and $\boldsymbol{X}_0$ in polynomial time, with failure probability bounded by $c_c p^{-6}$. Here $c_\star$ is as defined in Theorem 10.1, and c_a through c_c, and C are all positive constants.

By working with the preconditioned data samples $\widehat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}} \doteq \sqrt{\theta p}\left(\boldsymbol{Y}\boldsymbol{Y}^*\right)^{-1/2}\boldsymbol{Y}$,[5] we can use the same procedure as described above to recover complete dictionaries.

Theorem 11.4 (Main theorem - recovering complete dictionaries) *Assume the dictionary $\boldsymbol{A}_0$ is complete with a condition number $\kappa(\boldsymbol{A}_0)$ and we take $\widehat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}}$. Suppose $\theta \in (0, 1/3)$, $\mu_\star < c_a \min\left\{\theta n^{-1}, n^{-5/4}\right\}$, and $p \geq \frac{C}{c_\star^2 \theta^2} \max\left\{\frac{n^4}{\mu^4}, \frac{n^5}{\mu^2}\right\} \kappa^8(\boldsymbol{A}_0) \log^4\left(\frac{\kappa(\boldsymbol{A}_0) n}{\mu\theta}\right)$. The algorithmic pipeline with parameter setting*

$$\Delta_\star = \frac{c_d c_\star^3 \theta^3 \mu_\star^2}{n^{7/2} \log^{7/2}(np)} \tag{11.2.3}$$

recovers the dictionary $\boldsymbol{A}_0$ and $\boldsymbol{X}_0$ in polynomial time, with failure probability bounded by $c_b p^{-6}$. Here $c_\star$ is as defined in Theorem 10.1, and c_a, c_b are both positive constants.

We refer the readers to Section III of [SQW15c] for the detailed proofs of Theorem 11.3 and Theorem 11.4.

[5] 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.

Chapter 12

Numerical Simulations

12.1 Practical TRM Implementation

Fixing a small step size and solving the trust-region subproblem exactly eases the analysis, but also renders 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 here. Recent theoretical results on the practical version include [CGT12, BAC16].

Here we describe a practical implementation based on the *Manopt* toolbox [BMAS14][1]. 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 of $f(\boldsymbol{q})$), the tCG solver explicitly uses the negative curvature direction[2] as the initial search direction. This modification ensures the TRM solver always escape from saddle points/local maximizers with negative

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

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

directional curvature. Hence, the modified TRM algorithm based on Manopt is expected to have the same qualitative behavior as the idealized version we analyzed above, with better scalability. We will perform our numerical simulations using the modified TRM algorithm whenever necessary. Algorithm 3 together with Lemmas 9 and 10 and the surrounding discussion in the very recent work [BAC16] provides a detailed description of this practical version.

12.2 Simulated Data

To corroborate our theory, we experiment with dictionary recovery on simulated data.[3] For simplicity, we focus on recovering orthogonal dictionaries and we declare success once a single row of the coefficient matrix is recovered.

Since the problem is invariant to rotations, w.l.o.g. we set the dictionary as $\boldsymbol{A}_0 = \boldsymbol{I} \in \mathbb{R}^{n\times n}$. For any fixed sparsity k, 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 have similar behaviors. For our sparsity surrogate, we fix the smoothing parameter as $\mu = 10^{-2}$. Because the target points are the signed basis vector $\pm\boldsymbol{e}_i$'s (to recover rows of $\boldsymbol{X}_0$), for a solution $\widehat{\boldsymbol{q}}$ returned by the TRM algorithm, we define the reconstruction error (RE) to be

$$\texttt{RE} = \min_{i\in[n]} \left(\|\widehat{\boldsymbol{q}} - \boldsymbol{e}_i\| \,, \|\widehat{\boldsymbol{q}} + \boldsymbol{e}_i\| \right). \tag{12.2.1}$$

One trial is determined to be a success once $\texttt{RE} \le \mu$, 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).

We consider two settings: (1) fix $p = 5n^2 \log n$ and vary the dimension n and sparsity k; (2) fix the sparsity level as $\lceil 0.2 \cdot n \rceil$ and vary the dimension n and number of samples p. For each pair of (k, n) for (1), and each pair of (p, n) for (2), we repeat the simulations independently for $T = 5$ times. Fig. 12.1 shows the phase transition for the two settings. It seems that our TRM algorithm can work well into the linear region whenever $p \in O(n^2 \log n)$ (Fig. 12.1-Top), but p should have order greater than $\Omega(n)$ (Fig. 12.1-Bottom). The sample complexity from our theory is significantly suboptimal compared to this.

[3] The code is available online: https://github.com/sunju/dl_focm

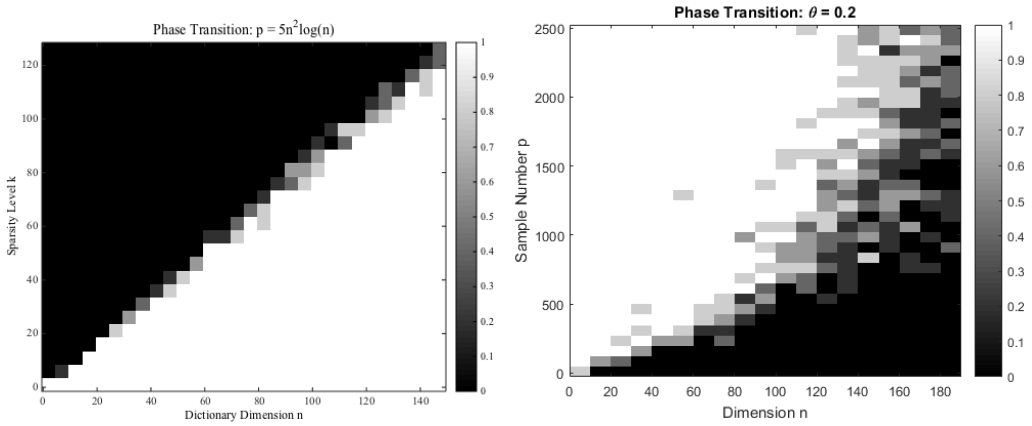


Figure 12.1: Phase transition for recovering a single sparse vector. **Top:** We fix $p = 5n^2 \log n$ and vary the dimension n and sparsity level k; **Bottom:** We fix the sparsity level as $\lceil 0.2 \cdot n \rceil$ and vary the dimension n and number of samples p. For each configuration, the experiment is independently repeated for five times. White indicates success, and black indicates failure.

12.3 Image Data Again

Our algorithmic framework has been derived based on the BG model on the coefficients. Real data may not admit sparse representations w.r.t. complete dictionaries, or even so, the coefficients may not obey the BG model. In this experiment, we explore how our algorithm performs in learning complete dictionaries for image patches, emulating our motivational experiment in Section 8.2 of Chapter 8. Thanks to research on image compression, we know patches of natural images tend to admit sparse representation, even w.r.t. simple orthogonal bases, such as Fourier basis or wavelets.

We take the three images that we used in the motivational experiment. For each image, we divide it into 8×8 non-overlapping patches, vectorize the patches, and then stack the vectorized patches into a data matrix $\boldsymbol{Y}$. $\boldsymbol{Y}$ is preconditioned as

$$\overline{\boldsymbol{Y}} = \left(\boldsymbol{Y}\boldsymbol{Y}^\top\right)^{-1/2}\boldsymbol{Y},$$

and the resulting $\overline{\boldsymbol{Y}}$ is fed to the dictionary learning pipeline described in Section 11.2. The smoothing parameter μ is fixed to 10^{-2}. Fig. 12.2 contains the learned dictionaries: the dictionaries generally contain localized, directional features that resemble subset of wavelets and generalizations. These are very reasonable representing elements for natural images. Thus, the BG coefficient model may be a sensible, simple model for natural images.

Another piece of strong evidence in support of the above claim is as follows. For each image, we repeat

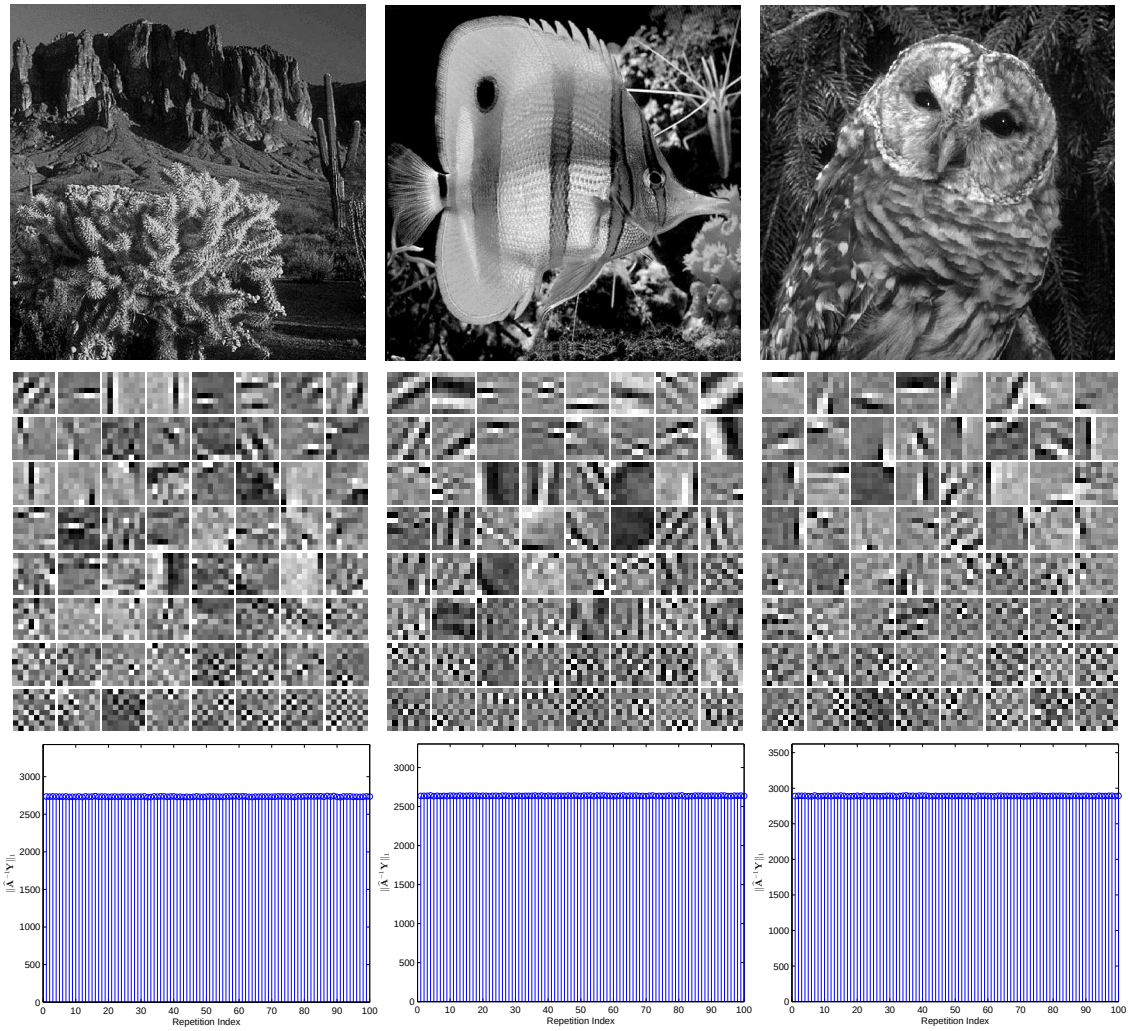


Figure 12.2: Results of learning complete dictionaries from image patches, using the algorithmic pipeline in Section 11.2. **Top:** Images we used for the experiment. These are the three images in Chapter 8. The way we formed the data matrix $\boldsymbol{Y}$ is exactly the same as in that experiment. **Middle:** The 64 dictionary elements we learned. **Bottom:** Let $\widehat{\boldsymbol{A}}$ be the final dictionary matrix at convergence. This row shows the value $\|\widehat{\boldsymbol{A}}^{-1}\boldsymbol{Y}\|_1$ across one hundred independent runs. The values are almost the same, with a relative difference less than 10^{-3}.

the learning pipeline for one hundred times, with independent initializations across the runs. Let $\widehat{\boldsymbol{A}}$ be the final learned dictionary for each run, we plot the value of $\|\widehat{\boldsymbol{A}}^{-1}\boldsymbol{Y}\|_1$ across the one hundred independent runs. Strikingly, the values are virtually the same, with a relative difference of 10^{-3}! This is predicted by our theory, under the BG model. If the model is unreasonable for natural images, the preconditioning, benign function landscape, LP rounding, and the deflation process that hinge on this model would have completely fallen down.

For this image experiment, $n = 64$ and $p = 4096$. A single run of the learning pipeline, including

solving 64 instances of the optimization over the sphere (with varying dimensions) and solving 64 instances of the LP rounding (using CVX), lasts about 20 minutes on a mid-range modern laptop. So with careful implementation we discussed above, the learning pipeline is actually not far from practical.

Chapter 13

Discussion

The dependency of p on n and other parameters could be suboptimal due to several factors: (1) The ℓ^1 proxy. Derivatives of the $\log\cosh$ function we adopted entail the $\tanh$ function, which is not amenable to effective approximation and affects the sample complexity; (2) Space of geometric characterization. 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) Dealing with the complete case. Treating the complete case directly, rather than using (pessimistic) bounds to treat it as a perturbation of the orthogonal case, is very likely to improve the sample complexity. 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 the 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 implicit functions. [1] For the coefficient model, as we alluded to in Section 8.4, our analysis and results likely can be carried through to coefficients with statistical dependence and physical constraints.

[1] This 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 challenges.

The connection to ICA we discussed in Section 8.4 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, whereby random initializations seem to work well, and the analytic techniques developed in that line are mostly fragmented and highly specialized. The analytic and algorithmic framework we developed here holds promise to providing a coherent account of these problems, see [SQW15d]. In particular, we have intentionally separated the geometric characterization and algorithm development, hoping to making both parts modular. It is interesting to see how far we can streamline the geometric characterization. Moreover, the separation allows development of more provable and practical algorithms, say in the direction of [GHJY15].

Part IV

Generalized Phase Retrieval

Can we recover a complex signal from its Fourier magnitudes? More generally, given a set of m measurements, $y_k = |\boldsymbol{a}_k^* \boldsymbol{x}|$ for $k = 1, \ldots, m$, 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 heuristics often work remarkably well for GPR in practice, but lack clear theoretical explanations. In this paper, 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 *numerous* 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 directional curvature around each saddle point. 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.

The remainder of this part is organized as follows. In Chapter 14 we motivate the generalized phase retrieval problem and overview main ingredients of our nonconvex approach. In Section 15, 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 Section 16 we present a modified trust-region method for solving GPR from an arbitrary initialization, which leads to our main computational guarantee. In Section 17 we study the empirical performance of our method for GPR. Section 18, concludes the main body with a discussion of open problems.

All the technical detailed are omitted in this part, we refer the readers to our paper [SQW16] for more detailed analysis.

Chapter 14

Introduction

14.1 Generalized Phase Retrieval and a Nonconvex Formulation

This chapter concerns the problem of recovering an n-dimensional complex vector $\boldsymbol{x}$ from the magnitudes $y_k = |\boldsymbol{a}_k^* \boldsymbol{x}|$ of its projections onto a collection of known complex vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m \in \mathbb{C}^n$. 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 $y_k = |\boldsymbol{a}_k^* \boldsymbol{x}|$ $(k = 1, \ldots, m)$, 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, 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, ZCL16, ZL16, WGE16, KÖ16, GX16, BE16, Wal16]. The surprising effectiveness of nonconvex heuristics on GPR remains largely mysterious. In this part of the thesis, we take a step towards bridging this gap.

We focus on a natural least-squares formulation[1] – discussed systematically in [SEC+15, JEH15] and first studied theoretically in [CLS15b, WWS15],

$$\text{minimize}_{\boldsymbol{z}\in\mathbb{C}^n}\ f(\boldsymbol{z}) \doteq \frac{1}{2m}\sum_{k=1}^{m}\left(y_k^2 - |\boldsymbol{a}_k^*\boldsymbol{z}|^2\right)^2. \tag{14.1.1}$$

We assume the $\boldsymbol{a}_k$'s are independent identically distributed (i.i.d.) complex Gaussian:

$$\boldsymbol{a}_k = \frac{1}{\sqrt{2}}\left(X_k + \mathrm{i}Y_k\right),\ \text{with } X_k, Y_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n) \text{ independent.} \tag{14.1.2}$$

$f(\boldsymbol{z})$ is a fourth-order polynomial in $\boldsymbol{z}$,[2] 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 (14.1.1) and related formulations rely on careful initialization in the vicinity of the ground truth [NJS13, CLS15b, CC15, WWS15, ZCL16, ZL16, WGE16, KÖ16, GX16, BE16, Wal16].

14.2 A Curious Experiment

We apply gradient descent to $f(\boldsymbol{z})$, starting from a *random initialization* $\boldsymbol{z}^{(0)}$:

$$\boldsymbol{z}^{(r+1)} = \boldsymbol{z}^{(r)} - \mu\nabla_{\boldsymbol{z}} f(\boldsymbol{z}^{(r)}),$$

where the step size μ is fixed for simplicity[3]. The result is quite striking (Figure 14.1): 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!

[1] Another least-squares formulation, $\text{minimize}_{\boldsymbol{z}}\ \frac{1}{2m}\sum_{k=1}^{m}(y_k - |\boldsymbol{a}_k^*\boldsymbol{z}|)^2$, was first studied in the seminal works [Fie82, GS72]. An obvious advantage of the $f(\boldsymbol{z})$ studied here is that it is differentiable in the sense of Wirtinger calculus introduced later.

[2] Strictly speaking, $f(\boldsymbol{z})$ is not a complex polynomial in $\boldsymbol{z}$ over the complex field; complex polynomials are necessarily complex differentiable. However, $f(\boldsymbol{z})$ is a fourth order real polynomial in real and complex parts of $\boldsymbol{z}$.

[3] Mathematically, $f(\boldsymbol{z})$ is not complex differentiable; here the gradient is defined based on the Wirtinger calculus [KD09]; see also [CLS15b]. This notion of gradient is a natural choice when optimizing real-valued functions of complex variables.

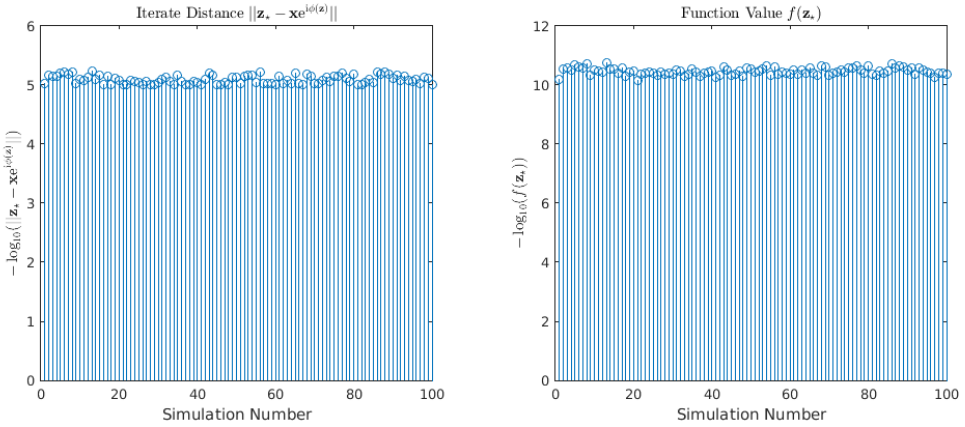


Figure 14.1: Gradient descent with random initialization seems to always return a global solution for (14.1.1)! Here $n = 100$, $m = 5n \log n$, step size $\mu = 0.05$, and stopping criterion is $\|\nabla_{\boldsymbol{z}} f(\boldsymbol{z})\| \leq 10^{-5}$. We fix the set of random measurements and the 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.

This contrasts with the typical "mental picture" of nonconvex objectives as possessing many spurious local minimizers.

14.3 A Geometric Analysis

The numerical surprise described above is not completely isolated. Simple heuristic methods have been observed to work surprisingly well for practical PR [Fie82, GS72, SEC+15, JEH15]. In this part of the thesis, we take a step towards explaining this phenomenon. We show that *although the function* (14.1.1) *is nonconvex, when m is reasonably large, it actually has benign global geometry which allows it to be globally optimized by efficient iterative methods, regardless of the initialization*.

This geometric structure is evident for real GPR (i.e., real signals with real random measurements) in $\mathbb{R}^2$. Figure 14.2 plots the function landscape of $f(\boldsymbol{z})$ for this case with large m (i.e., $\mathbb{E}_{\boldsymbol{a}}[f(\boldsymbol{z})]$ approximately). Notice that (i) the only local minimizers are exactly $\pm\boldsymbol{x}$ – they are also global minimizers;[4] (ii) there are saddle points (and a local maximizer), but around them there is a 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.)[5], for GPR in $\mathbb{C}^n$,

[4]Note that the global sign cannot be recovered.

[5]The probability is with respect to drawing of $\boldsymbol{a}_k$'s.

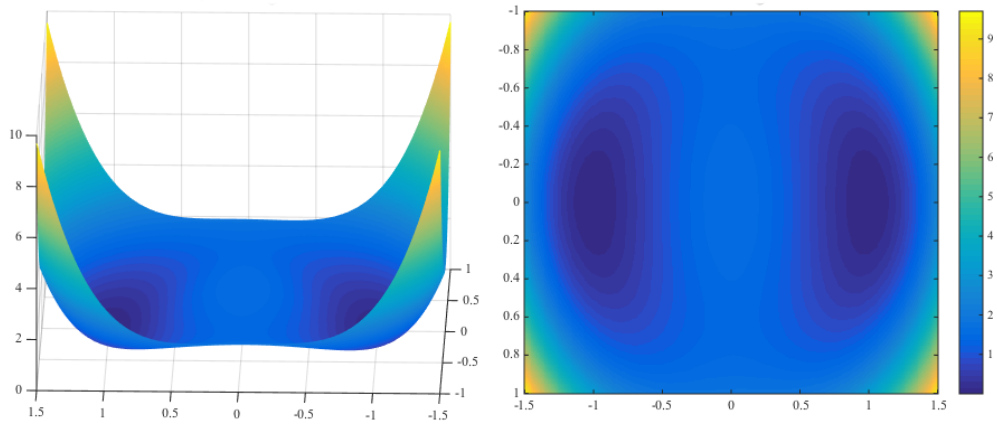


Figure 14.2: Function landscape of (14.1.1) for $\boldsymbol{x} = [1; 0]$ and $m \to \infty$. The only local and also global minimizers are $\pm\boldsymbol{x}$. There are two saddle points near $\pm[0; 1/\sqrt{2}]$, around each there is a negative curvature direction along $\pm\boldsymbol{x}$. (Left) 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.

when m is reasonably large (Theorem 15.1). In particular, we show that when $m \geq Cn\log^3 n$, w.h.p., (i) the only local and also global minimizers to (14.1.1) are the target $\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi}$ 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 (a weaker version of this local restricted strong convexity was first established in [CLS15b]; see also [WWS15]).

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, PP16]), curvilinear search [Gol80] and trust-region methods [CGT00, NP06, SQW15d]. The key property that the methods must possess is the ability to escape saddle points at which the Hessian has a strictly negative eigenvalue[6]. We corroborate this claim by developing a second-order trust-region algorithm for this problem, and prove that (Theorem 16.1) (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.

In sum, our geometrical analysis produces the following result.

Informal Statement of Our Main Results *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*

[6]Such saddle points are called ridable saddles [SQW15d] or strict saddles [GHJY15]; see [AG16] for computational methods for escaping from higher-order saddles also.

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 absolute positive 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 in polynomial time by optimizing the nonconvex formulation (14.1.1). Similar conclusions have been obtained in [NJS13, CLS15b, CC15, WWS15, ZCL16, ZL16, WGE16, KÖ16, GX16, BE16, Wal16], also based on nonconvex optimization. One salient feature of our result is that the optimization method is "initialization free" - any initialization in the prescribed ball works. This follows directly from the benign global geometry of $f(\boldsymbol{z})$. In contrast, all prior nonconvex methods require careful initializations that are already near the unknown target $\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi}$, based on characterization of only local geometry. We believe our global geometrical analysis sheds light on mechanism of the above numerical surprise.

The second-order trust-region method, albeit polynomial-time, may not be the most practical algorithm for solving GPR. Deriving the most practical algorithms is not the main focus of this thesis. We mentioned above that any iterative method with saddle-escaping capability can be deployed to solve the nonconvex formulation; our geometrical analysis constitutes a solid basis for developing and analyzing much more practical algorithms for GPR.

14.4 Prior Arts and Connections

The survey papers [SEC+15, JEH15] provide comprehensive accounts of recent progress on GPR. In this section, we focus on provable efficient (particularly, nonconvex) methods for GPR, and draw connections to other work on provable nonconvex heuristics for practical problems.

Provable methods for GPR. Although heuristic 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.[7]

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 (14.1.1) under the same hypotheses as the thesis, 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.[8] More recent work in this line [ZCL16, ZL16, WGE16, KÖ16, GX16, BE16, Wal16] concerns error stability, alternative formulations, algorithms, and measurement models. Compared to the SDP-based methods, these methods are more scalable and closer to methods used in practice. All these analyses are based on local geometry in nature, and hence depend on the spectral initializer being sufficiently close to the target set. In contrast, we explicitly characterize the global function landscape of (14.1.1). Its benign global geometric structure allows several algorithmic choices (see Section 14.3) that need *no special initialization* and scale much better than the convex approaches.

Near the target set (i.e., $\mathcal{R}_3$ in Theorem 15.1), [CLS15b, CC15] established a local curvature property that is strictly weaker than our restricted strong convexity result. The former is sufficient for obtaining convergence results for first-order methods, while the latter is necessary for establishing convergence results for second-order method. Besides these, [Sol14] and [WWS15] also explicitly established local strong convexity near the target set for real GPR in $\mathbb{R}^n$; the Hessian-form characterization presented in [WWS15] is real-version counterpart to ours here.

(Global) Geometric analysis of other nonconvex problems. The approach taken here is similar in spirit to our recent geometric analysis of a nonconvex formulation for complete dictionary learning [SQW15a].

[7] Another line of research [BCE06, BBCE09, ABFM14] seeks to co-design the measurements and recovery algorithms based on frame- or graph-theoretic tools. While revising this work, new convex relaxations based on second-order cone programming have been proposed [GS16, BR16, HV16**?**].

[8] 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.

For that problem, we also identified a similar geometric structure that allows efficient global optimization without special initialization. There, by analyzing the geometry of a nonconvex formulation, we derived a provable efficient algorithm for recovering square invertible dictionaries when the coefficient matrix has a constant fraction of nonzero entries. Previous results required the dictionary matrix to have far fewer nonzero entries. [SQW15d] provides a high-level overview of the common geometric structure that arises in dictionary learning, GPR and several other problems. This approach has also been applied to other problems [GHJY15, BBV16, BVB16, SC16, Kaw16, BNS16, GLM16, PKCS16]. Despite these similarities, GPR raises several novel technical challenges: the objective is heavy-tailed, and minimizing the number of measurements is important[9].

Our work sits amid the recent surge of work on provable nonconvex heuristics 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, SA14], 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 global geometric analysis allows flexible algorithm design (i.e., separation of geometry and algorithms) and gives some clues as to the behavior of nonconvex heuristics used in practice, which often succeed without clever initialization.

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 $\boldsymbol{M} \succeq \boldsymbol{0}$ in $\mathbb{R}^{n\times n}$ from linear measurement of the form $z_k = \mathrm{tr}(\boldsymbol{A}_k\boldsymbol{M})$ with symmetric $\boldsymbol{A}_k$ for $k = 1, \ldots, m$. 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 $z_k = \mathrm{tr}(\boldsymbol{X}^*\boldsymbol{A}_k\boldsymbol{X})$. 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, BNS16].

[9] The same challenge is also faced by [CLS15b, CC15].

14.5 Notations and Wirtinger Calculus

Basic notations and facts. Throughout this part of the thesis, we will often use the canonical identification of $\mathbb{C}^n$ and $\mathbb{R}^{2n}$, which assign $\boldsymbol{z} \in \mathbb{C}^n$ to $[\Re(\boldsymbol{z}); \Im(\boldsymbol{z})] \in \mathbb{R}^{2n}$. This is so natural that we will not explicitly state the identification when no confusion is caused. We say two complex vectors are orthogonal in the geometric (real) sense if they are orthogonal after the canonical identification[10]. 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 $f(\boldsymbol{z}) = f(\boldsymbol{z}\mathrm{e}^{\mathrm{i}\phi})$ for all ϕ, and the set $\left\{\boldsymbol{z}\mathrm{e}^{\mathrm{i}\phi} : \phi \in [0, 2\pi)\right\}$ 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 $\mathcal{X} \doteq \left\{\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi} : \phi \in [0, 2\pi)\right\}$. Moreover, we define

$$\phi(\boldsymbol{z}) \doteq \operatorname*{arg\,min}_{\phi\in[0,2\pi)} \left\|\boldsymbol{z} - \boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi}\right\|, \quad \boldsymbol{h}(\boldsymbol{z}) \doteq \boldsymbol{z} - \boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})}, \quad \operatorname{dist}(\boldsymbol{z}, \mathcal{X}) \doteq \|\boldsymbol{h}(\boldsymbol{z})\|. \tag{14.5.1}$$

for any $\boldsymbol{z} \in \mathbb{C}^n$. It is not difficult to see that $\boldsymbol{z}^*\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})} = |\boldsymbol{x}^*\boldsymbol{z}|$. Moreover, $\boldsymbol{z}_T \doteq \mathrm{i}\boldsymbol{z}/\|\boldsymbol{z}\|$ and $-\boldsymbol{z}_T$ are the unit vectors tangent to the circle $\left\{\boldsymbol{z}\mathrm{e}^{\mathrm{i}\phi} : \phi \in [0, 2\pi)\right\}$ 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

$$\begin{aligned}
\frac{\partial g}{\partial \boldsymbol{z}} &\doteq \left.\frac{\partial g(\boldsymbol{z}, \overline{\boldsymbol{z}})}{\partial \boldsymbol{z}}\right|_{\overline{\boldsymbol{z}} \text{ constant}} = \left.\left[\frac{\partial g(\boldsymbol{z}, \overline{\boldsymbol{z}})}{\partial z_1}, \ldots, \frac{\partial g(\boldsymbol{z}, \overline{\boldsymbol{z}})}{\partial z_n}\right]\right|_{\overline{\boldsymbol{z}} \text{ constant}} \\
\frac{\partial g}{\partial \overline{\boldsymbol{z}}} &\doteq \left.\frac{\partial g(\boldsymbol{z}, \overline{\boldsymbol{z}})}{\partial \overline{\boldsymbol{z}}}\right|_{\boldsymbol{z} \text{ constant}} = \left.\left[\frac{\partial g(\boldsymbol{z}, \overline{\boldsymbol{z}})}{\partial \overline{z_1}}, \ldots, \frac{\partial g(\boldsymbol{z}, \overline{\boldsymbol{z}})}{\partial \overline{z_n}}\right]\right|_{\boldsymbol{z} \text{ constant}}.
\end{aligned}$$

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 $\frac{\partial g(\boldsymbol{z}, \overline{\boldsymbol{z}})}{\partial z_i}$,

[10] Two complex vectors $\boldsymbol{w}, \boldsymbol{v}$ are orthogonal in complex sense if $\boldsymbol{w}^*\boldsymbol{v} = 0$.

all the usual rules of calculus apply.[11]

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

$$\nabla g(\boldsymbol{z}) = \left[\frac{\partial g}{\partial \boldsymbol{z}}, \frac{\partial g}{\partial \overline{\boldsymbol{z}}}\right]^* \quad \nabla^2 g(\boldsymbol{z}) = \begin{bmatrix} \frac{\partial}{\partial \boldsymbol{z}}\left(\frac{\partial g}{\partial \boldsymbol{z}}\right)^* & \frac{\partial}{\partial \overline{\boldsymbol{z}}}\left(\frac{\partial g}{\partial \boldsymbol{z}}\right)^* \\ \frac{\partial}{\partial \boldsymbol{z}}\left(\frac{\partial g}{\partial \overline{\boldsymbol{z}}}\right)^* & \frac{\partial}{\partial \overline{\boldsymbol{z}}}\left(\frac{\partial g}{\partial \overline{\boldsymbol{z}}}\right)^* \end{bmatrix}, \tag{14.5.2}$$

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 $(\boldsymbol{\delta}, \bar{\boldsymbol{\delta}})$.

For our particular function $f(\boldsymbol{z}) : \mathbb{C}^n \mapsto \mathbb{R}$ defined in (14.1.1), direct calculation gives

$$\nabla f(\boldsymbol{z}) = \frac{1}{m} \sum_{k=1}^{m} \begin{bmatrix} \left(\left|\boldsymbol{a}_k^* \boldsymbol{z}\right|^2 - y_k^2\right) \left(\boldsymbol{a}_k \boldsymbol{a}_k^*\right) \boldsymbol{z} \\ \left(\left|\boldsymbol{a}_k^* \boldsymbol{z}\right|^2 - y_k^2\right) \left(\boldsymbol{a}_k \boldsymbol{a}_k^*\right)^\top \overline{\boldsymbol{z}} \end{bmatrix}, \tag{14.5.3}$$

$$\nabla^2 f(\boldsymbol{z}) = \frac{1}{m} \sum_{k=1}^{m} \begin{bmatrix} \left(2\left|\boldsymbol{a}_k^* \boldsymbol{z}\right|^2 - y_k^2\right) \boldsymbol{a}_k \boldsymbol{a}_k^* & \left(\boldsymbol{a}_k^* \boldsymbol{z}\right)^2 \boldsymbol{a}_k \boldsymbol{a}_k^\top \\ \left(\boldsymbol{z}^* \boldsymbol{a}_k\right)^2 \overline{\boldsymbol{a}_k} \boldsymbol{a}_k^* & \left(2\left|\boldsymbol{a}_k^* \boldsymbol{z}\right|^2 - y_k^2\right) \overline{\boldsymbol{a}_k} \boldsymbol{a}_k^\top \end{bmatrix}. \tag{14.5.4}$$

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.

[11] 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, $\frac{\partial g}{\partial \overline{\boldsymbol{z}}} \doteq \frac{1}{2}\left(\frac{\partial g}{\partial \boldsymbol{u}} + \mathrm{i}\frac{\partial g}{\partial \boldsymbol{v}}\right)$.

Chapter 15

High Dimensional Geometry of the Objective Function

The low-dimensional example described in the introduction (Figure 14.2) 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 $m \to \infty$. In this chapter, We characterize this large-sample geometry. 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 $\boldsymbol{x}$.

More specifically, 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 n – degrees of freedom of $\boldsymbol{x}$. 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. Our main geometric result is as follows:

Theorem 15.1 (Main Geometric Results) *There exist positive absolute constants C, c, such that when $m \geq Cn\log^3 n$, it holds with probability at least $1 - cm^{-1}$ that $f(\boldsymbol{z})$ has no spurious local minimizers and the only local/global minimizers are exactly the target set $\mathcal{X}$. More precisely, with the same probability,*

$$\frac{1}{\|\boldsymbol{x}\|^2}\begin{bmatrix}\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})}\\ \overline{\boldsymbol{x}}\mathrm{e}^{-\mathrm{i}\phi(\boldsymbol{z})}\end{bmatrix}^* \nabla^2 f(\boldsymbol{z}) \begin{bmatrix}\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})}\\ \overline{\boldsymbol{x}}\mathrm{e}^{-\mathrm{i}\phi(\boldsymbol{z})}\end{bmatrix} \leq -\frac{1}{100}\|\boldsymbol{x}\|^2, \qquad \forall\, \boldsymbol{z} \in \mathcal{R}_1,$$

(Negative Curvature)

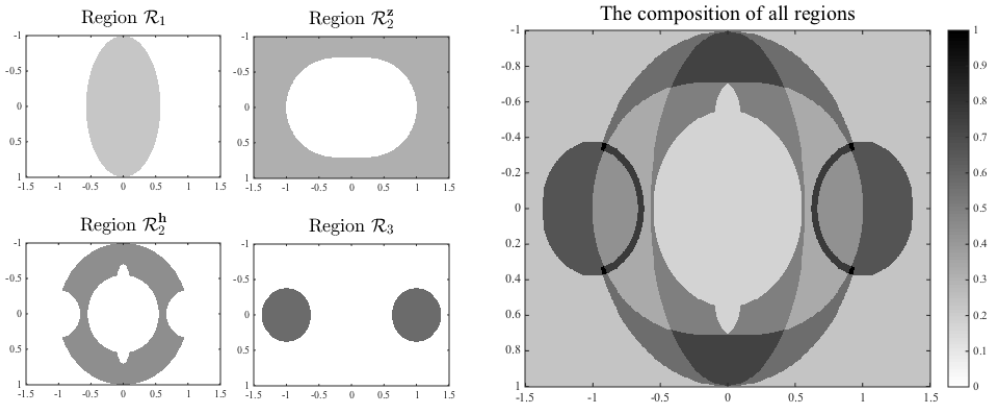


Figure 15.1: Schematic illustration of partitioning regions for Theorem 15.1. This plot corresponds to Figure 14.2, i.e., the target signal is $\boldsymbol{x} = [1; 0]$ and measurements are real Gaussians, such that the function is defined in $\mathbb{R}^2$. Here $\mathcal{R}_2^{\boldsymbol{z}} \cup \mathcal{R}_2^{\boldsymbol{h}}$ is $\mathcal{R}_2$; we will need the further sub-division of $\mathcal{R}_2$ in the proof.

$$\|\nabla_{\boldsymbol{z}} f(\boldsymbol{z})\| \;\geq\; \frac{1}{1000}\|\boldsymbol{x}\|^2 \|\boldsymbol{z}\|, \qquad \forall\, \boldsymbol{z} \in \mathcal{R}_2,$$

(Large Gradient)

$$\begin{bmatrix} \boldsymbol{g}(\boldsymbol{z}) \\ \overline{\boldsymbol{g}(\boldsymbol{z})} \end{bmatrix}^* \nabla^2 f(\boldsymbol{z}) \begin{bmatrix} \boldsymbol{g}(\boldsymbol{z}) \\ \overline{\boldsymbol{g}(\boldsymbol{z})} \end{bmatrix} \;\geq\; \frac{1}{4}\|\boldsymbol{x}\|^2, \qquad \forall\, \boldsymbol{z} \in \mathcal{R}_3,$$

(Restricted Strong Convexity)

where, assuming $\boldsymbol{h}(\boldsymbol{z})$ *as defined in* (14.5.1),

$$\boldsymbol{g}(\boldsymbol{z}) \doteq \begin{cases} \boldsymbol{h}(\boldsymbol{z})/\|\boldsymbol{h}(\boldsymbol{z})\| & \text{if } \operatorname{dist}(\boldsymbol{z}, \mathcal{X}) \neq 0, \\ \boldsymbol{h} \in \mathcal{S} \doteq \{\boldsymbol{h} : \Im(\boldsymbol{h}^* \boldsymbol{z}) = 0, \|\boldsymbol{h}\| = 1\} & \text{if } \boldsymbol{z} \in \mathcal{X}. \end{cases}$$

Here the regions $\mathcal{R}_1,\ \mathcal{R}_2,\ \mathcal{R}_3$ *are defined as*

$$\mathcal{R}_1 \;\doteq\; \left\{ \boldsymbol{z} : \begin{bmatrix} \boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})} \\ \overline{\boldsymbol{x}}\mathrm{e}^{-\mathrm{i}\phi(\boldsymbol{z})} \end{bmatrix}^* \mathbb{E}\left[\nabla^2 f(\boldsymbol{z})\right] \begin{bmatrix} \boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})} \\ \overline{\boldsymbol{x}}\mathrm{e}^{-\mathrm{i}\phi(\boldsymbol{z})} \end{bmatrix} \leq -\frac{1}{100}\|\boldsymbol{x}\|^2 \|\boldsymbol{z}\|^2 - \frac{1}{50}\|\boldsymbol{x}\|^4 \right\}, \tag{15.0.1}$$

$$\mathcal{R}_3 \;\doteq\; \left\{ \boldsymbol{z} : \operatorname{dist}(\boldsymbol{z}, \mathcal{X}) \leq \frac{1}{\sqrt{7}}\|\boldsymbol{x}\| \right\}, \tag{15.0.2}$$

$$\mathcal{R}_2 \;\doteq\; (\mathcal{R}_1 \cup \mathcal{R}_3)^c. \tag{15.0.3}$$

We refer the readers to [SQW16] for the detailed proofs of the theorem. Figure 15.1 visualizes the different regions described in Theorem 15.1, 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 **Ju:** $\mathcal{R}_2$ is far from a critical point. The rest of the space consists of points near critical points, and is covered by $\mathcal{R}_1 \cup \mathcal{R}_3$. For any $\boldsymbol{z}$ in $\mathcal{R}_1$, the quantity

$$\frac{1}{\|\boldsymbol{x}\|^2} \begin{bmatrix} \boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})} \\ \overline{\boldsymbol{x}}\mathrm{e}^{-\mathrm{i}\phi(\boldsymbol{z})} \end{bmatrix}^* \nabla^2 f(\boldsymbol{z}) \begin{bmatrix} \boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})} \\ \overline{\boldsymbol{x}}\mathrm{e}^{-\mathrm{i}\phi(\boldsymbol{z})} \end{bmatrix}$$

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 $\nabla_{\boldsymbol{z}} f(\boldsymbol{z}) = \boldsymbol{0}$. 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 $\mathcal{X}$ 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 $\left\{\boldsymbol{z}\mathrm{e}^{\mathrm{i}\phi} : \phi \in [0, 2\pi)\right\}$. In particular, the result applied to $\boldsymbol{z} \in \mathcal{X}$ implies that on $\mathcal{X}$, $f(\boldsymbol{z})$ is strongly convex in any direction orthogonal to $\mathcal{X}$ (i.e., any "radial" direction w.r.t. $\mathcal{X}$). This observation, together with the fact that the Hessian is Lipschitz, implies that there is a neighborhood $N(\mathcal{X})$ of $\mathcal{X}$, such that for all $\boldsymbol{z} \in N(\mathcal{X})$, $\boldsymbol{v}^* \nabla^2 f(\boldsymbol{z}) \boldsymbol{v} > 0$ for *every* $\boldsymbol{v}$ that is orthogonal to the trivial direction $\mathrm{i}\boldsymbol{z}$, 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. The geometric characterization of the whole space 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 Section 11.1).

In sum, our objective $f(\boldsymbol{z})$ has the benign geometry that each $\boldsymbol{z} \in \mathbb{C}^n$ has either large gradient or negative directional curvature, or lies in the vicinity of local minimizers around which the function is locally restrictedly strongly convex. Functions with this property lie in the ridable-saddle function class [GHJY15, SQW15d]. 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 obtain a local minimizer asymptotically. Theorem 15.1 shows that for our problem, every local minimizer is global, and so for our problem, these algorithms obtain a global minimizer asymptotically. Moreover, with appropriate quantitative assumptions on the geometric structure as we obtained (i.e., either gradient is *sufficiently* large, or the direction curvature is *sufficiently* negative, or local directional convexity is *sufficiently* strong),

these candidate methods actually find a global minimizer in polynomial time.

Chapter 16

Optimization by Trust-Region Method

Based on the geometric characterization in Chapter 15, we describe a second-order trust-region algorithm that produces a close approximation (i.e., up to numerical precision) to a global minimizer of (14.1.1) 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.

16.1 A Modified Trust-Region Algorithm

The basic idea of the trust-region method is simple: we generate a sequence of iterates $\boldsymbol{z}^{(0)}, \boldsymbol{z}^{(1)}, \ldots$, 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

$$\text{minimize}_{\boldsymbol{\delta} \in \mathbb{C}^n} \ \widehat{f}(\boldsymbol{\delta}; \boldsymbol{z}^{(r)}), \quad \text{subject to} \quad \Im\left(\boldsymbol{\delta}^* \boldsymbol{z}^{(r)}\right) = 0, \quad \|\boldsymbol{\delta}\| \leq \Delta, \tag{16.1.1}$$

to obtain the step $\boldsymbol{\delta}$. In (16.1.1), 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.

The modified trust-region subproblem is easily seen to be equivalent to the 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 $\mathbb{R}^{2n}$ (the canonical identification of $\mathbb{C}^n$ and $\mathbb{R}^{2n}$ applies whenever needed). 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 $\boldsymbol{U}(\boldsymbol{z}^{(r)})$)

$$\text{minimize}_{\boldsymbol{\xi} \in \mathbb{R}^{2n-1}} \ \widehat{f}(\boldsymbol{U}\boldsymbol{\xi}; \boldsymbol{z}^{(r)}), \quad \text{subject to} \quad \|\boldsymbol{\xi}\| \le \Delta. \tag{16.1.2}$$

Let us define

$$\boldsymbol{g}(\boldsymbol{z}^{(r)}) \doteq \begin{bmatrix} \boldsymbol{U} \\ \overline{\boldsymbol{U}} \end{bmatrix}^* \nabla f(\boldsymbol{z}^{(r)}), \quad \boldsymbol{H}(\boldsymbol{z}^{(r)}) \doteq \begin{bmatrix} \boldsymbol{U} \\ \overline{\boldsymbol{U}} \end{bmatrix}^* \nabla^2 f(\boldsymbol{z}^{(r)}) \begin{bmatrix} \boldsymbol{U} \\ \overline{\boldsymbol{U}} \end{bmatrix}. \tag{16.1.3}$$

Then, the quadratic approximation of $f(\boldsymbol{z})$ around $\boldsymbol{z}^{(r)}$ can be rewritten as

$$\widehat{f}(\boldsymbol{\xi}; \boldsymbol{z}^{(r)}) = f(\boldsymbol{z}^{(r)}) + \boldsymbol{\xi}^\top \boldsymbol{g}(\boldsymbol{z}^{(r)}) + \frac{1}{2} \boldsymbol{\xi}^\top \boldsymbol{H}(\boldsymbol{z}^{(r)}) \boldsymbol{\xi}. \tag{16.1.4}$$

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. Thus, the problem (16.1.2) is in fact an instance of the classical trust-region subproblem w.r.t. real variable $\boldsymbol{\xi}$. A minimizer to (16.1.1) can be obtained from a minimizer of (16.1.2) $\boldsymbol{\xi}_\star$ as $\boldsymbol{\delta}_\star = \boldsymbol{U}\boldsymbol{\xi}_\star$.

So, any method which can solve the classical trust-region subproblem can be directly applied to the modified problem (16.1.1). Although the resulting problem can be nonconvex (as $\boldsymbol{H}(\boldsymbol{z}^{(r)})$ in (16.1.4) can be indefinite), it can be solved in polynomial time, by root-finding [MS83, CGT00] or SDP relaxation [RW97, FW04]. Our convergence guarantees assume an exact solution of this problem.

16.2 Convergence of the Trust-region Method

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, $\frac{1}{m}\sum_{k=1}^m |\boldsymbol{a}_k^* \boldsymbol{x}|^2 \ge \frac{1}{9}\|\boldsymbol{x}\|^2$ with probability at least $1 - \exp(-cm)$. Thus, with the same probability, the

quantity

$$R_0 \doteq 3\left(\frac{1}{m}\sum_{k=1}^{m}|\boldsymbol{a}_k^*\boldsymbol{x}|^2\right)^{1/2}$$

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 the fact that $\max_{k\in[m]}\|\boldsymbol{a}_k\|^4 \le 10n^2\log^2 m$ with probability at least $1-c_am^{-n}$, we have that the following estimate

$$\begin{aligned}
&\inf_{\boldsymbol{z},\boldsymbol{z}':\ \|\boldsymbol{z}\|\le R_0,\ \|\boldsymbol{z}'\|\ge R_1} f(\boldsymbol{z}')-f(\boldsymbol{z})\\
&\quad=\inf_{\boldsymbol{z},\boldsymbol{z}':\ \|\boldsymbol{z}\|\le R_0,\ \|\boldsymbol{z}'\|\ge R_1}\frac{1}{m}\sum_{k=1}^{m}\Big[|\boldsymbol{a}_k^*\boldsymbol{z}'|^4-|\boldsymbol{a}_k^*\boldsymbol{z}|^4-2\,|\boldsymbol{a}_k^*\boldsymbol{z}'|^2\,|\boldsymbol{a}_k^*\boldsymbol{x}|^2+2\,|\boldsymbol{a}_k^*\boldsymbol{z}|^2\,|\boldsymbol{a}_k^*\boldsymbol{x}|^2\Big]\\
&\quad\ge\inf_{\boldsymbol{z},\boldsymbol{z}':\ \|\boldsymbol{z}\|\le R_0,\ \|\boldsymbol{z}'\|\ge R_1}\frac{199}{200}\|\boldsymbol{z}'\|^4-10n^2\log^2 m\,\|\boldsymbol{z}\|^4-\frac{201}{200}\left(\|\boldsymbol{z}'\|^2\|\boldsymbol{x}\|^2+|\boldsymbol{x}^*\boldsymbol{z}'|^2\right)\\
&\quad\ge\inf_{\boldsymbol{z}':\|\boldsymbol{z}'\|\ge R_1}\frac{199}{200}\|\boldsymbol{z}'\|^4-10n^2\log^2 mR_0^4-\frac{201}{100}\|\boldsymbol{z}'\|^2R_0^2
\end{aligned}$$

holds with probability at least $1-c_bm^{-1}$, provided $m\ge Cn\log n$ for a sufficiently large C. It can be checked that when

$$R_1=3\sqrt{n\log m}R_0, \tag{16.2.1}$$

we have

$$\inf_{\boldsymbol{z}':\|\boldsymbol{z}'\|\ge R_1}\frac{199}{200}\|\boldsymbol{z}'\|^4-10n^2\log^2 mR_0^4-\frac{201}{100}\|\boldsymbol{z}'\|^2R_0^2\ge 40n^2\log^2 mR_0^4.$$

Thus, we conclude that when $m\ge Cn\log n$, w.h.p., the sublevel set $\{\boldsymbol{z}: f(\boldsymbol{z})\le f(\boldsymbol{z}^{(0)})\}$ is contained in the set

$$\Gamma\doteq\mathbb{CB}^n(R_1). \tag{16.2.2}$$

TRM Convergence Throughout, we assume $m\ge Cn\log^3 n$ for a sufficiently large constant C, so that all the events of interest hold w.h.p.. The convergence guarantee of the trust-region method can be summarized as follows.

> **Theorem 16.1 (TRM Convergence)** *Suppose $m\ge Cn\log^3 n$ for a sufficiently large constant C. Then with probability at least $1-c_am^{-1}$, the trust-region algorithm with an* arbitrary initialization $\boldsymbol{z}^{(0)}\in\mathbb{CB}^n(R_0)$,

where $R_0 = 3(\frac{1}{m}\sum_{k=1}^{m} y_k^2)^{1/2}$*, will return a solution that is* ε*-close to the target set* $\mathcal{X}$ *in*

$$\frac{c_b}{\Delta^2 \|\boldsymbol{x}\|^2} f(\boldsymbol{z}^{(0)}) + \log\log\left(\frac{c_c \|\boldsymbol{x}\|}{\varepsilon}\right) \tag{16.2.3}$$

steps, provided that

$$\Delta \leq c_d (n^{7/2} \log^{7/2} m)^{-1} \|\boldsymbol{x}\| . \tag{16.2.4}$$

Here c_a *through* c_d *are positive absolute constants.*

Our initialization is an arbitrary point $\boldsymbol{z}^{(0)} \in \mathbb{CB}^n(R_0) \subseteq \Gamma$. We analyze effect of a trust-region step from any iterate $\boldsymbol{z}^{(r)} \in \Gamma$. Based on these arguments, we 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 $\boldsymbol{z} + \mathbb{CB}^n(\Delta)$. 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 $\widehat{f}(\boldsymbol{\delta}^\star, \boldsymbol{z}) - f(\boldsymbol{z}) < -\varepsilon$, 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. We refer the readers to our paper [SQW16] for more detailed analysis.

Chapter 17

Numerical Simulations

Our convergence analysis for the TRM is based on two idealizations: (i) the trust-region subproblem is solved exactly; and (ii) the step-size is fixed to be sufficiently small. These simplifications 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, by backtracking. It is relatively straightforward to modify our analysis to account for inexact subproblem solvers; for sake of brevity, we do not pursue this here[1].

In this section, 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 Manopt toolbox [BMAS14][2]. Manopt is a user-friendly Matlab toolbox that implements several sophisticated solvers for tackling optimization problems on 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. 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,

[1] The proof ideas are contained in Chap 6 of [CGT00]; see also [AMS09]. Intuitively, such result is possible because reasonably good approximate solutions to the TRM subproblem make qualitatively similar progress as the exact solution. Recent work [CGT12, BAC16] has established worst-case polynomial iteration complexity (under reasonable assumptions on the geometric parameters of the functions, of course) of TRM to converge to point verifying the second-order optimality conditions. Their results allow inexact trust-region subproblem solvers, as well as adaptive step sizes. Based on our geometric result, we could have directly called their results, producing slightly worse iteration complexity bounds. It is not hard to adapt their proof taking advantage of the stronger geometric property we established and produce tighter results.

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

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 $f(\boldsymbol{z})$), the tCG solver explicitly uses the negative curvature direction[3] as the initial search direction. This modification[4] ensures the TRM solver always escapes saddle points/local maximizers with directional negative curvature. Hence, the modified TRM algorithm based on Manopt is expected to have the same qualitative behavior as the idealized version we analyzed.

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

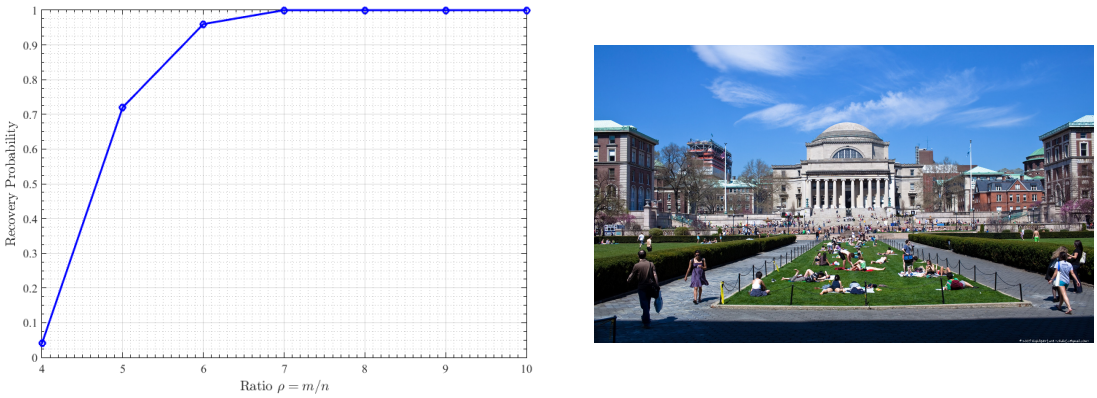


Figure 17.1: (Left) Recovery performance for GPR when optimizing (14.1.1) with the TRM. With $n = 1000$ and m varying, we consider a fixed problem instance for each m, and run the TRM algorithm 25 times from independently random initializations. The empirical recovery probability is a test of whether the benign geometric structure holds. (Right) A small "artistic" Columbia University campus image we use for comparing TRM and gradient descent.

signal $\boldsymbol{x}$, and a fixed set of complex Gaussian vectors. We run the TRM algorithm 25 times for each problem instance, with independent random initializations. Successfully recovery is declared if at termination the optimization variable $\boldsymbol{z}_\infty$ satisfies

$$\varepsilon_{\mathrm{Rel}} \doteq \|\boldsymbol{z}_\infty - \boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z}_\infty)}\| / \|\boldsymbol{x}\| \leq 10^{-3}.$$

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 17.1 shows the recovery performance. It seems that $m = 7n$ samples may be sufficient to ensure the geometric property holds.[5] On the other hand, $m = 6n$ is not sufficient,

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

[4] Similar modification is also adopted in the TRM algorithmic framework in the recent work [BAC16] (Algorithm 3).

[5] This prescription should be taken with a grain of salt, as here we have only tested a single fixed n.

whereas in theory it is known $4n$ samples are enough to guarantee measurement injectivity for complex signals [BCE06].[6]

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 17.1 (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.[7]

[6]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.

[7]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 18

Discussion

In this work, we provide a complete geometric characterization of the nonconvex formulation (14.1.1) 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 (14.1.1), without special initializations. We close this part of thesis by discussing possible extensions and relevant open problems.

Sample complexity and measurement schemes. Our result (Theorem 15.1 and Theorem 16.1) 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 for all signals; see also [CLS15b, WGE16, ZL16]). It is interesting to see if the gaps can be closed. Our current analysis pertains to 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 (**Ju: for reduced real version**) produces very positive result; see Figure 18.1.

Figure 18.1: Function landscape of (14.1.1) for $\boldsymbol{x} = [1; 0]$ and $m \to \infty$ for the real-value-masked discrete cosine transform measurements (i.e., real-valued version of the coded diffraction model [CLS15a]). The mask takes i.i.d. values from $\{1, 0, -1\}$; each entry takes 1 or -1 with probability $1/4$ respectively, and takes 0 with probability $1/2$. The landscape is qualitatively similar to that for the Gaussian model (Figure 14.2).

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 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, SA14, 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, SQW15d] (see also [GHJY15, AG16], and [Kaw16, SC16, BNS16, GLM16, PKCS16, BBV16, BVB16]) seem promising in this regard. It is interesting to consider whether analogous geometric structure exists for other practical problems.

Part V

Convolutional Phase Retrieval

We study the *convolutional phase retrieval* problem, which considers recovering an unknown signal $\boldsymbol{x} \in \mathbb{C}^n$ from m measurements consisting of the magnitude of its cyclic convolution with a known kernel $\boldsymbol{a} \in \mathbb{C}^m$. This model is motivated by applications such as channel estimation, optics, and underwater acoustic communication, where the signal of interest is acted on by a given channel/filter, and phase information is difficult or impossible to acquire. We show that when $\boldsymbol{a}$ is random and the sample number m is sufficiently large, with high probability $\boldsymbol{x}$ can be efficiently recovered up to a global phase using a combination of spectral initialization and generalized gradient descent. The main challenge is coping with dependencies in the measurement operator. We overcome this challenge by using ideas from decoupling theory, suprema of chaos processes and the restricted isometry property of random circulant matrices, and recent analysis for alternating minimization methods.

This part of the thesis is based on our paper [QZEW17], and it is organized as follows. In Chapter 19 we introduce and motivate the convolutional phase retrieval problem. In Chapter 20, we introduce the basic formulation of the problem and the algorithm. In Chapter 21, we present the main results and proof sketch, detailed analysis is postponed to Chapter 24. In Chapter 22, we corroborate our analysis with numerical experiments. We discuss the potential impacts of our work in Chapter 23. Finally, all the basic probability tools that are used in this part are postponed to Appendix B.

Chapter 19

Introduction

We study the problem of recovering an unknown signal $\boldsymbol{x} \in \mathbb{C}^n$ from measurements $\boldsymbol{y} = |\boldsymbol{a} \circledast \boldsymbol{x}|$, which consist of the magnitude of its convolution with a given filter $\boldsymbol{a} \in \mathbb{C}^m$,

$$\text{find} \quad \boldsymbol{z}, \quad \text{s.t.} \quad \boldsymbol{y} = |\boldsymbol{a} \circledast \boldsymbol{z}|\,, \tag{19.0.1}$$

where $\circledast$ denotes cyclic convolution modulo m. Let $\boldsymbol{C_a} \in \mathbb{C}^{m\times m}$ be a circulant matrix generated by $\boldsymbol{a}$, and let $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ be a matrix formed by the first n columns of $\boldsymbol{C_a}$. Then the *convolutional phase retrieval* problem can be rewritten in the common matrix-vector form

$$\text{find} \quad \boldsymbol{z}, \quad \text{s.t.} \quad \boldsymbol{y} = |\boldsymbol{A}\boldsymbol{z}|\,. \tag{19.0.2}$$

This problem is motivated by applications in areas such as *channel estimation* [WBJ15], *noncoherent optical communication* [GK76], and *underwater acoustic communication* [SCP94]. For example, in millimeter-wave (mm-wave) wireless communications for 5G networks [SGD+15], one important problem is to reconstruct the angle of arrival (AoA) of a signal from measurements, which are taken by the convolution of signal AoA and the antenna pattern. Because of technical difficulties the phase measurements are either very noisy and unreliable, or expensive to acquire, it is preferred to only take measurements of signal magnitude and the phase information is lost.

Most known results on the exact solution of phase retrieval problems [CSV13, Sol14, CC15, WGE16, WdM15, Wal16] pertain to *generic random matrices*, where the entries of $\boldsymbol{A}$ are independent subgaussian random variables. However, in practice it is almost impossible to implement purely random measurement matrices. In many applications, the measurement is much more structured – the convolutional model stud-

ied here is one such structured measurement operator. Moreover, structured measurements often admit more efficient numerical methods: by using the *fast Fourier transform* for matrix-vector products, the benign structure of the convolutional model (19.0.1) allows to design methods with $\mathcal{O}(m)$ memory and $\mathcal{O}(m \log m)$ computation cost per iteration. In contrast, for generic measurements, the cost is around $\mathcal{O}(mn)$.

In this work, we study the convolutional phase retrieval problem (19.0.1) under the assumption that the kernel $\boldsymbol{a} = [a_1, \cdots, a_m]^\top$ is random, with entries i.i.d. complex Gaussian,

$$\boldsymbol{a} = \boldsymbol{u} + \mathrm{i}\boldsymbol{v}, \quad \boldsymbol{u}, \boldsymbol{v} \sim_{\text{i.i.d.}} \mathcal{N}\left(\boldsymbol{0}, \tfrac{1}{2}\boldsymbol{I}\right). \tag{19.0.3}$$

Compared to the generic random measurements, the random convolution model we study here is far more structured: it is parameterized by only $\mathcal{O}(m)$ independent complex normal random variables, whereas the generic model involves $\mathcal{O}(mn)$ random variables. This extra structure poses significant challenges for analysis: the rows and columns of the sensing matrix $\boldsymbol{A}$ are probabilistically dependent, and classical probability tools (based on concentration of functions of independent random vectors) do not apply.

We propose and analyze a local gradient descent type method, minimizing a weighted, *nonconvex* and *nonsmooth* objective

$$\min_{\boldsymbol{z} \in \mathbb{C}^n} f(\boldsymbol{z}) = \frac{1}{2m} \left\| \boldsymbol{b}^{1/2} \odot (\boldsymbol{y} - |\boldsymbol{A}\boldsymbol{z}|) \right\|^2, \tag{19.0.4}$$

where $\odot$ denotes the Hadamard product. $\boldsymbol{b} \in \mathbb{R}^m_{++}$ is a weighting vector, which is introduced mainly for analysis purposes. The choice of $\boldsymbol{b}$ is discussed in Section 21. Our result can be informally summarized as follows.

Theorem 19.1 (Informal) *When $m \geq \Omega(n \operatorname{poly}\log n)$, with high probability, spectral initialization [NJS13, CLS15b] produces an initialization $\boldsymbol{z}^{(0)}$ that is $\mathcal{O}(1/\operatorname{poly}\log n)$ close to the optimum. Moreover, when $m \geq \Omega\left(\frac{\|\boldsymbol{C}_{\boldsymbol{x}}\|^2}{\|\boldsymbol{x}\|^2} n \operatorname{poly}\log n\right)$, with high probability, a certain gradient descent method based on (19.0.4) converges linearly from this initialization to the optimal set $\mathcal{X} = \left\{\boldsymbol{x}e^{\mathrm{i}\phi} \mid \phi \in [0, 2\pi)\right\}$ of points that differ from the true signal $\boldsymbol{x}$ only by a global phase.*

Here, $\boldsymbol{C}_{\boldsymbol{x}} \in \mathbb{C}^{m \times m}$ denotes the circulant matrix corresponding to cyclic convolution with a length m zero padding of $\boldsymbol{x}$, and $\operatorname{poly}\log n$ denotes a polynomial in $\log n$. Compared to the results of generalized phase retrieval, the sample complexity m here also depends on $\|\boldsymbol{C}_{\boldsymbol{x}}\|$, which is quite different. The operator norm $\|\boldsymbol{C}_{\boldsymbol{x}}\|$ is inhomogeneous over $\mathbb{CS}^{n-1}$: for a typical[1] $\boldsymbol{x} \in \mathbb{CS}^{n-1}$, $\|\boldsymbol{C}_{\boldsymbol{x}}\|$ is of the order $\mathcal{O}(\log n)$ and the

[1] e.g., $\boldsymbol{x}$ is drawn uniformly at random from $\mathbb{CS}^{n-1}$.

sample complexity matches that of the generalized phase retrieval up to $\log$ factors; the "bad" case is when $\boldsymbol{x}$ is *sparse* in the Fourier domain: $\|\boldsymbol{C}_{\boldsymbol{x}}\| \sim \mathcal{O}(\sqrt{n})$ and m can be as large as $\mathcal{O}(n^2 \operatorname{poly} \log n)$. Based on the result from the work [CL14], it raises the possibility that our dependence on the spectral spikiness of the target $\boldsymbol{x}$ could be unnecessary (although we don't see any easy way to carry our analysis through without this dependence). Further investigation is left for the future work.

Our proof is based on ideas from *decoupling theory* [DlPG99], the *suprema of chaos processes* and *restricted isometry property* of random circulant matrices [Rau10, KMR14], and inspired by a new iterative analysis of alternating minimization methods [Wal16]. Our analysis draws connections between the convergence properties of gradient descent and the classical alternating direction method. This allows us to avoid the need to argue that high-degree polynomials in the structured random matrix $\boldsymbol{A}$ concentrate uniformly, as would be required by a straightforward translation of existing analysis to this new setting. Instead, we control the bulk effect of phase errors uniformly in a neighborhood around the ground truth. This requires us to develop new decoupling and concentration tools for controlling nonlinear phase functions of circulant random matrices, which could be potentially useful for analyzing other random circulant convolution problems, such as blind deconvolution [ZLK+17] and convolutional dictionary learning [HHW15].

19.1 Literature Review

Prior art in phase retrieval The challenge of developing efficient, guaranteed methods for phase retrieval has attracted substantial interest over the past several decades [SEC+15, JEH15]. The problem is motivated by applications such as X-ray crystallography [Mil90, Rob93], microscopy [MIJ+02], astronomy [DF87], diffraction and array imaging [BDP+07, CMP11], and optics [Wal63]. The most classical method is the *error reduction algorithm* derived by Gerchberg and Saxton [GS72], also known as the alternating direction method. This approach has been further improved by the *hybrid input-output* (HIO) algorithm [Fie82]. For oversampled Fourier measurements, it often works surprisingly well in practice, while its global convergence properties still largely remains as a mystery.

For the *generalized phase retrieval problem* for which the sensing matrix $\boldsymbol{A}$ is random, the problem is better-studied: in many cases, when the number of measurements is large enough, the target solution can be exactly recovered by using either convex or nonconvex optimization methods. The first theoretical guarantees for global recovery of generalized phase retrieval are based on convex optimization – the so-called *Phaselift/Phasemax* methods [CSV13, CESV13, WdM15]. These methods lift the problem to a higher dimen-

sion and solve a semi-definite programming (SDP) problem. However, the high computational cost of SDP limits their practicality. Quite recently, [BR16, GS16, HV16] reveal that the problem can also be solved in the natural parameter space via linear programming.

Recently, a promising research direction for generalized phase retrieval is based on nonconvex optimization. The first result of this type is due to [NJS13], Netrapalli et al. showed that the alternating minimization method provably converges to the truth, when initialized using a spectral method and provided with fresh samples at each iteration. Later on, Candès et al. [CLS15b] showed that with the same initialization, gradient descent for the nonconvex least squares objective,

$$\min_{\boldsymbol{z}\in\mathbb{C}^n} f_1(\boldsymbol{z}) = \frac{1}{2m}\left\|\boldsymbol{y}^2 - |\boldsymbol{A}\boldsymbol{z}|^2\right\|^2, \tag{19.1.1}$$

provably recovers the ground truth, with near-optimal sample complexity $m \geq \Omega(n\log n)$. The subsequent work [CC15, ZL16, WGE16] further reduced the sample complexity to $m \geq \Omega(n)$ by using different nonconvex objectives and truncation techniques. In particular, recent work by [ZL16, WGE16] studied a nonsmooth objective that is similar to ours (19.0.4) with weighting $\boldsymbol{b} = \boldsymbol{1}$. Compared to the SDP-based methods, these methods are more scalable and closer to the methods used in practice. Moreover, Sun et. al. [SQW16] reveal that the nonconvex objective (19.1.1) actually has a benign *global geometry*: with high probability, it has no bad critical points with $m \geq \Omega(n\log^3 n)$ samples[2]. Such a result enables initialization-free nonconvex recovery[3].

Structured random measurements The study of structured random measurements in signal processing has quite a long history [KR14]. For compressed sensing [CRT06a], the work [CRT06b, CT06, EK12] studied random Fourier measurements, and later [Rau10, KMR14] proved similar results for partial random convolution measurements. However, the study of structured random measurements for phase retrieval is still quite limited. In particular, [GKK13] and [CLS15a] studied t-designs and coded diffraction patterns (i.e., random masked Fourier measurements) using semidefinite programming. Recent work studied nonconvex optimization using coded diffraction patterns [CLS15b] and STFT measurements [BE16], both of which minimize a nonconvex objective similar to (19.1.1). These different measurement models are motivated by different applications. For instance, the coded diffraction is designed for imaging applications such as X-

[2][Sol17] further tightened the sample complexity to $m \geq \Omega(n\log n)$ by using more advanced probability tools.

[3]For convolutional phase retrieval, it would be nicer to characterize the global geometry of the problem as in [GHJY15, SQW15d, SQW16, SQW15a]. However, the inhomogeneity of $\|\boldsymbol{C}_{\boldsymbol{x}}\|$ over $\mathbb{CS}^{n-1}$ causes tremendous difficulties for concentration with $m \geq \Omega(n\,\mathrm{poly}\log n)$ samples.

ray diffraction imaging, the STFT can be applied to frequency resolved optical gating [TKD+96] and some speech processing tasks [LO79]. Both of the results show iterative contraction in a region that is at most $\mathcal{O}(1/\sqrt{n})$-close to the optimum. The radius of the region is either not large enough for initialization to reach, or extra technique like resampling is needed for initialization. In comparison, the contraction region we show for the random convolutional model is larger $\mathcal{O}(1/\mathrm{polylog}(n))$, which is achievable in the initialization stage via the spectral method. For a more detailed review of this subject, we refer the readers to Section 4 of [KR14].

In addition, the convolutional measurement can also be reviewed as a single masked coded diffraction partterns, as we have $\boldsymbol{a} \circledast \boldsymbol{x} = F^{-1}(\widehat{\boldsymbol{a}} \odot \widehat{\boldsymbol{x}})$, where $\widehat{\boldsymbol{a}}$ is the Fourier transform of $\boldsymbol{a}$ and $\widehat{\boldsymbol{x}}$ is the oversampled Fourier transform of $\boldsymbol{x}$. The sample complexity $m \geq \Omega(n \log^4 n)$ in [CLS15b] suggests that the dependence of our sample complexity on $\|\boldsymbol{C}_{\boldsymbol{x}}\|$ for convolutional phase retrieval might not be necessary and can be improved. On the other hand, our results suggest that the contraction region is larger than $\mathcal{O}(1/\sqrt{n})$ for coded diffraction patterns, that resampling for initialization might not be necessary.

19.2 Notations

We use $\boldsymbol{C}_{\boldsymbol{a}} \in \mathbb{C}^{m\times m}$ to denote a circulant matrix generated from $\boldsymbol{a}$, i.e.,

$$\boldsymbol{C}_{\boldsymbol{a}} = \begin{bmatrix} a_1 & a_m & \cdots & a_3 & a_2 \\ a_2 & a_1 & a_m & & a_3 \\ \vdots & a_2 & a_1 & \ddots & \vdots \\ a_{m-1} & & \ddots & \ddots & a_m \\ a_m & a_{m-1} & \cdots & a_2 & a_1 \end{bmatrix} = \begin{bmatrix} s_0[\boldsymbol{a}] & s_1[\boldsymbol{a}] & \cdots & s_{m-1}[\boldsymbol{a}] \end{bmatrix}, \tag{19.2.1}$$

where $s_\ell[\cdot]$ $(0 \leq \ell \leq m-1)$ denotes a circulant shift by ℓ samples. We use $g_1 \perp\!\!\!\perp g_2$ to denote the independence of two random variables $g_1,\ g_2$. For a random variable X, its L^p norm is defined as

$$\|X\|_{L^p} = \mathbb{E}\left[|X|^p\right]^{1/p},$$

for any positive integer $p \geq 1$. For a smooth function $f \in \mathcal{C}^1$, its L^∞ norm is defined as

$$\|f\|_{L^\infty} = \sup_{t\in \mathrm{dom}(f)} |f(t)|.$$

For an arbitrary set Ω, we use $|\Omega|$ to denote the cardinality of Ω, and use $\mathrm{supp}(\Omega)$ to denote the support set of Ω, i.e., the subset containing elements which are not mapped to zero. If $|\Omega| = \ell$, we use $\boldsymbol{R}_\Omega : \mathbb{R}^m \mapsto \mathbb{R}^\ell$ to denote a mapping that maps a vector into its coordinates restricted to the set Ω. We use $\mathbf{1}_\Omega$ to denote the indicator function of the set Ω, i.e.,

$$[\mathbf{1}_\Omega]_j = \begin{cases} 1 & \text{if } j \in \Omega, \\ 0 & \text{otherwise.} \end{cases}$$

where $[\cdot]_j$ denotes the jth coordinate of vector. Let $\boldsymbol{F}_n \in \mathbb{C}^{n\times n}$ denote a unnormalized $n \times n$ Fourier matrix with $\|\boldsymbol{F}_n\| = \sqrt{n}$, and let $\boldsymbol{F}_n^m \in \mathbb{C}^{m\times n}$ $(m \geq n)$ to be an oversampled Fourier matrix. Throughout this part of the thesis, we assume the ground truth signal to be $\boldsymbol{x} \in \mathbb{C}^n$. Because the problem can only be solved up to a global phase shift, we define the optimal solution set as $\mathcal{X} = \left\{\boldsymbol{x}e^{\mathrm{i}\phi} \mid \phi \in [0, 2\pi)\right\}$, and correspondingly define the distance from a point $\boldsymbol{z} \in \mathbb{C}^n$ to the set $\mathcal{X}$ as

$$\mathrm{dist}(\boldsymbol{z}, \mathcal{X}) \doteq \inf_{\phi\in[0,2\pi)} \left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\phi}\right\|.$$

Chapter 20

Algorithm

In this work, we develop an approach to convolutional phase retrieval based on local nonconvex optimization. Our proposed algorithm has two components: (1) a careful initialization using the spectral method; (2) local refinement by (generalized) gradient descent. We introduce the two steps in reverse order.

20.1 Minimization of a nonconvex and nonsmooth objective

We consider minimizing a weighted *nonconvex* and *nonsmooth* objective

$$f(\boldsymbol{z}) = \frac{1}{2m} \left\| \boldsymbol{b}^{1/2} \odot (\boldsymbol{y} - |\boldsymbol{A}\boldsymbol{z}|) \right\|^2. \tag{20.1.1}$$

The introduction of the positive weights $\boldsymbol{b}$ facilitates our analysis, by enabling us to compare certain functions of the dependent random matrix $\boldsymbol{A}$ to functions involving more independent random variables. We will substantiate this claim in the next section.

Although the function (19.0.4) is not complex-differentiable, for reasons explained in [Sol14] and Section 1 of [SQW16], we adopt the *Wirtinger calculus* instead[KD09], which can be thought of as a compact way of organizing the real partial derivatives. It should also be noted that the absolute value $|\cdot|$ is nonsmooth at 0, and hence the function $f(\cdot)$ is not differentiable everywhere even in the real sense. Similar to [WGE16], for any complex number $u \in \mathbb{C}$, if we define its phase $\phi(u)$ by

$$\exp\left(\mathrm{i}\phi(u)\right) \doteq \begin{cases} u/|u| & \text{if } |u| \neq 0, \\ 1 & \text{otherwise,} \end{cases}$$

the (generalized) Wirtinger gradient of (19.0.4) is

$$\frac{\partial}{\partial \boldsymbol{z}} f(\boldsymbol{z}) \;=\; \frac{1}{m}\boldsymbol{A}^* \operatorname{diag}(\boldsymbol{b}) \left[\boldsymbol{A}\boldsymbol{z} - \boldsymbol{y} \odot \exp\left(\mathrm{i}\phi(\boldsymbol{A}\boldsymbol{z})\right)\right]. \tag{20.1.2}$$

Starting from some initialization $\boldsymbol{z}^{(0)}$, we minimize the objective (20.1.1) by gradient descent

$$\boldsymbol{z}^{(r+1)} \;=\; \boldsymbol{z}^{(r)} - \tau \frac{\partial}{\partial \boldsymbol{z}} f(\boldsymbol{z}^{(r)}), \tag{20.1.3}$$

where $\tau > 0$ is the stepsize. Indeed, $\frac{\partial}{\partial \boldsymbol{z}} f(\boldsymbol{z})$ can be interpreted as the gradient of $f(\boldsymbol{z})$ as in the real case; this method is also referred to as *amplitude flow* [WGE16].

20.2 Initialization via spectral method

Algorithm 2 Spectral Initialization

Input: Observations $\{y_k\}_{k=1}^m$.
Output: The initial guess $\boldsymbol{z}^{(0)}$.

1: Estimate the norm of $\boldsymbol{x}$ by

$$\lambda = \sqrt{\frac{1}{m}\sum_{k=1}^{m} y_k^2}$$

2: Compute the leading eigenvector $\widetilde{\boldsymbol{z}}^{(0)} \in \mathbb{CS}^{n-1}$ of the matrix,

$$\boldsymbol{Y} = \frac{1}{m}\sum_{k=1}^{m} y_k^2 \boldsymbol{a}_k \boldsymbol{a}_k^* = \frac{1}{m}\boldsymbol{A}^* \operatorname{diag}\left(\boldsymbol{y}^2\right) \boldsymbol{A},$$

3: Set $\boldsymbol{z}^{(0)} = \lambda \widetilde{\boldsymbol{z}}^{(0)}$.

Similar to [NJS13, Sol14], we compute the initialization $\boldsymbol{z}^{(0)}$ via a spectral method, detailed in Algorithm 2. More specifically, $\boldsymbol{z}^{(0)}$ is a scaled version of the leading eigenvector of the following matrix

$$\boldsymbol{Y} \;=\; \frac{1}{m}\sum_{k=1}^{m} y_k^2 \boldsymbol{a}_k \boldsymbol{a}_k^* \;=\; \frac{1}{m}\boldsymbol{A}^* \operatorname{diag}\left(\boldsymbol{y}^2\right) \boldsymbol{A}, \tag{20.2.1}$$

which is constructed from the knowledge of the sensing vectors and observations. The leading eigenvector of $\boldsymbol{Y}$ can be efficiently computed via the power method. Note that $\mathbb{E}[\boldsymbol{Y}] = \|\boldsymbol{x}\|^2 \boldsymbol{I} + \boldsymbol{x}\boldsymbol{x}^*$, so the leading eigenvector of $\mathbb{E}[\boldsymbol{Y}]$ is proportional to the target solution $\boldsymbol{x}$. Under the random convolutional model of $\boldsymbol{A}$, by using probability tools from [KR14], we show that $\boldsymbol{v}^*\boldsymbol{Y}\boldsymbol{v}$ concentrates to its expectation $\boldsymbol{v}^*\mathbb{E}[\boldsymbol{Y}]\boldsymbol{v}$ for all $\boldsymbol{v} \in \mathbb{CS}^{n-1}$ whenever $m \geq \Omega(n \operatorname{poly}\log n)$, ensuring that the initialization $\boldsymbol{z}^{(0)}$ is close to the optimal set $\mathcal{X}$.[1]

[1]Several variants of this initialization approach have been introduced in the literature. They slightly improve the sample complexity

for generalized phase retrieval with i.i.d. measurements. Those methods include the truncated spectral method [CC15], null initialization [CFL] and orthogonality-promoting initialization [WGE16]. For the simplicity of analysis, here we only consider Algorithm 2 for the convolutional model.

Chapter 21

Main Result and Analysis

In this chapter, we introduce our main theoretical result, and sketch the basic ideas behind the analysis.

21.1 Main Result

Our main theoretical result shows that with high probability, the algorithm described in the previous section succeeds.

Theorem 21.1 (Main Result) *Whenever $m \geq C_0 n \log^{31} n$, Algorithm 2 produces an initialization $\boldsymbol{z}^{(0)}$ that satisfies*

$$\operatorname{dist}\left(\boldsymbol{z}^{(0)}, \mathcal{X}\right) \;\leq\; c_0 \log^{-6} n \, \|\boldsymbol{x}\|$$

with probability at least $1 - c_1 m^{-c_2}$. Suppose $\boldsymbol{b} = \zeta_{\sigma^2}(\boldsymbol{y})$, where

$$\zeta_{\sigma^2}(t) = 1 - 2\pi\sigma^2 \xi_{\sigma^2}(t), \qquad \xi_{\sigma^2}(t) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{|t|^2}{2\sigma^2}\right), \tag{21.1.1}$$

with $\sigma^2 > 1/2$. Starting from $\boldsymbol{z}^{(0)}$, with $\sigma^2 = 0.51$ and stepsize $\tau = 2.02$, whenever $m \geq C_1 \frac{\|\boldsymbol{C}_{\boldsymbol{x}}\|^2}{\|\boldsymbol{x}\|^2} \max\left\{\log^{17} n, n \log^4 n\right\}$, with probability at least $1 - c_3 m^{-c_4}$ for all iterates $\boldsymbol{z}^{(r)}$ ($r \geq 1$) defined in (24.2.2), we have

$$\operatorname{dist}\left(\boldsymbol{z}^{(r)}, \mathcal{X}\right) \;\leq\; (1-\varrho)^r \operatorname{dist}\left(\boldsymbol{z}^{(0)}, \mathcal{X}\right), \tag{21.1.2}$$

for some numerical constant $\varrho \in (0, 1)$. Here, c_0, c_1, c_2, c_3, c_4 and C_0, C_1 are positive numerical constants.

Remark: Our result shows that by initializing the problem $\mathcal{O}(1/\mathrm{polylog}(n))$-close to the optimum via spectral method, the gradient descent (24.2.2) converges linearly to the optimal solution. As we can see, the sample complexity here also depends on $\|\boldsymbol{C}_{\boldsymbol{x}}\|$, which is quite different from the i.i.d. case. For a typical $\boldsymbol{x} \in \mathbb{CS}^{n-1}$ (e.g., $\boldsymbol{x}$ is drawn uniformly random from $\mathbb{CS}^{n-1}$), $\|\boldsymbol{C}_{\boldsymbol{x}}\|$ remains as $\mathcal{O}(\log n)$, the sample complexity $m \geq \Omega\left(n \operatorname{poly}\log n\right)$ matches the i.i.d. case up to log factors. However, $\|\boldsymbol{C}_{\boldsymbol{x}}\|$ is nonhomogeneous over $\boldsymbol{x} \in \mathbb{CS}^{n-1}$: if $\boldsymbol{x}$ is sparse in the Fourier domain (e.g., $\boldsymbol{x} = \frac{1}{\sqrt{n}}\mathbf{1}$), the sample complexity can be as large as $m \geq \Omega\left(n^2 \operatorname{poly}\log n\right)$. Such a behavior is also demonstrated in the experiments of Section 22. We believe the (very large!) number of logarithms in our result is an artifact of our analysis, rather than a limitation of the method. We expect to reduce the sample complexity to $m \geq \Omega\left(\frac{\|\boldsymbol{C}_{\boldsymbol{x}}\|^2}{\|\boldsymbol{x}\|^2} n \log^6 n\right)$ by a tighter analysis, which is left for future work. The choices of the weighting $\boldsymbol{b} \in \mathbb{R}^m$ in (21.1.1), $\sigma^2 = 0.51$, and the stepsize $\tau = 2.02$ are purely for the purpose of analysis. In practice, the algorithm converges with $\boldsymbol{b} = \mathbf{1}$ and a choice of small stepsize τ, or by using backtracking linesearch for the stepsize τ.

21.2 A Sketch of Analysis

In this subsection, we briefly highlight some major challenges and novel ideas behind the analysis. All the detailed proofs are postponed to Section 24. The core idea behind the analysis is to show that the iterate contracts once we initialize close enough to the optimum. In the following, we first describe the basic ideas of proving iterative contraction, which critically depends on bounding a certain nonlinear function of a random circulant matrix. Second, we sketch the core ideas how to bound such a complicated term via the decoupling technique.

21.2.1 Proof sketch of iterative contraction

Our iterative analysis is inspired by the recent analysis of *alternating direction method* (ADM) [Wal16]. In the following, we draw connections between the gradient descent method (24.2.2) and ADM, and sketch the basic ideas of convergence analysis.

ADM iteration. ADM is a classical method for solving phase retrieval problems [GS72, NJS13, Wal16], which can be considered as a heuristic method for solving the following nonconvex problem

$$\min_{\boldsymbol{z}\in\mathbb{C}^n, |\boldsymbol{u}|=\mathbf{1}} \frac{1}{2}\left\|\boldsymbol{A}\boldsymbol{z} - \boldsymbol{y}\odot\boldsymbol{u}\right\|^2.$$

At every iterate $\widehat{\boldsymbol{z}}^{(r)}$, ADM proceeds in two steps:

$$\begin{aligned}\boldsymbol{c}^{(r+1)} &= \boldsymbol{y} \odot \exp\left(\boldsymbol{A}\widehat{\boldsymbol{z}}^{(r)}\right),\\ \widehat{\boldsymbol{z}}^{(r+1)} &= \arg\min_{\boldsymbol{z}} \frac{1}{2}\left\|\boldsymbol{A}\boldsymbol{z} - \boldsymbol{c}^{(r+1)}\right\|^2,\end{aligned}$$

which leads to the following update

$$\widehat{\boldsymbol{z}}^{(r+1)} \;=\; \boldsymbol{A}^\dagger\left(\boldsymbol{y} \odot \exp\left(\boldsymbol{A}\widehat{\boldsymbol{z}}^{(r)}\right)\right),$$

where $\boldsymbol{A}^\dagger = \left(\boldsymbol{A}^*\boldsymbol{A}\right)^{-1}\boldsymbol{A}^*$ is the pseudo-inverse of $\boldsymbol{A}$. Let $\widehat{\theta}_r = \arg\min_{\overline{\theta}\in[0,2\pi)}\left\|\widehat{\boldsymbol{z}}^{(r)} - \boldsymbol{x}e^{\mathrm{i}\overline{\theta}}\right\|$. The distance between $\widehat{\boldsymbol{z}}^{(r+1)}$ and $\mathcal{X}$ is bounded by

$$\operatorname{dist}\left(\widehat{\boldsymbol{z}}^{(r+1)}, \mathcal{X}\right) = \left\|\widehat{\boldsymbol{z}}^{(r+1)} - \boldsymbol{x}e^{\mathrm{i}\widehat{\theta}_{r+1}}\right\| \le \left\|\boldsymbol{A}^\dagger\right\| \left\|\boldsymbol{A}\boldsymbol{x}e^{\mathrm{i}\widehat{\theta}_r} - \left(\boldsymbol{y}\odot\exp\left(\boldsymbol{A}\widehat{\boldsymbol{z}}^{(r)}\right)\right)\right\|. \tag{21.2.1}$$

Gradient descent with $\boldsymbol{b} = \boldsymbol{1}$. For simplicity and illustration purposes, let us first consider the gradient descent update (24.2.2) with $\boldsymbol{b} = \boldsymbol{1}$. Let $\theta_r = \arg\min_{\overline{\theta}\in[0,2\pi)}\left\|\boldsymbol{z}^{(r)} - \boldsymbol{x}e^{\mathrm{i}\overline{\theta}}\right\|$, with stepsize $\tau = 1$. The distance between the iterate $\boldsymbol{z}^{(r+1)}$ and the optimal set $\mathcal{X}$ is bounded by

$$\begin{aligned}\operatorname{dist}\left(\boldsymbol{z}^{(r+1)}, \mathcal{X}\right) = \left\|\boldsymbol{z}^{(r+1)} - \boldsymbol{x}e^{\mathrm{i}\theta_{r+1}}\right\| &\le \left\|\boldsymbol{I} - \frac{1}{m}\boldsymbol{A}^*\boldsymbol{A}\right\| \left\|\boldsymbol{z}^{(r)} - \boldsymbol{x}e^{\mathrm{i}\theta_r}\right\| \\ &\quad + \frac{1}{m}\left\|\boldsymbol{A}\right\| \left\|\boldsymbol{A}\boldsymbol{x}e^{\mathrm{i}\theta_r} - \boldsymbol{y}\odot\exp\left(\mathrm{i}\phi(\boldsymbol{A}\boldsymbol{z}^{(r)})\right)\right\|.\end{aligned} \tag{21.2.2}$$

Towards iterative contraction. By the measure concentration, it can be shown that

$$\left\|\boldsymbol{I} - \frac{1}{m}\boldsymbol{A}^*\boldsymbol{A}\right\| = o(1), \quad \left\|\boldsymbol{A}\right\| \approx \sqrt{m}, \quad \left\|\boldsymbol{A}^\dagger\right\| \approx 1/\sqrt{m}, \tag{21.2.3}$$

holds with high probability whenever $m \ge \Omega\left(n \operatorname{poly}\log n\right)$. Therefore, to show iterative contraction of both ADM and gradient descent methods, based on (21.2.1) and (21.2.2), it is sufficient to show that

$$\left\|\boldsymbol{A}\boldsymbol{x}e^{\mathrm{i}\theta} - \boldsymbol{y}\odot\exp\left(\mathrm{i}\phi(\boldsymbol{A}\boldsymbol{z})\right)\right\| \le (1-\eta)\sqrt{m}\left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\|, \tag{21.2.4}$$

for some constant $\eta \in (0,1)$ sufficiently small, where $\theta = \arg\min_{\overline{\theta}\in[0,2\pi)}\left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\overline{\theta}}\right\|$ such that $e^{\mathrm{i}\theta} = \boldsymbol{x}^*\boldsymbol{z}/\left|\boldsymbol{x}^*\boldsymbol{z}\right|$. By borrowing ideas of controlling (21.2.4) for the ADM method [Wal16], this observation provides a new way of analyzing the gradient descent method. As an attempt to show (21.2.4) for the random circulant matrix $\boldsymbol{A}$, we invoke Lemma B.1, which controls the error in a first order approximation to $\exp(\mathrm{i}\phi(\cdot))$. Let

us decompose

$$\boldsymbol{z} = \alpha \boldsymbol{x} + \beta \boldsymbol{w},$$

where $\boldsymbol{w} \in \mathbb{CS}^{n-1}$ with $\boldsymbol{w} \perp \boldsymbol{x}$, and $\alpha, \beta \in \mathbb{C}$. Notice that $\phi(\alpha) = \theta$, then by Lemma B.1, for any $\rho \in (0,1)$ we have

$$\begin{aligned} \left\| \boldsymbol{A}\boldsymbol{x} e^{\mathrm{i}\theta} - \boldsymbol{y} \odot \exp\left(\mathrm{i}\phi(\boldsymbol{A}\boldsymbol{z})\right) \right\| &= \left\| |\boldsymbol{A}\boldsymbol{x}| \odot \left[\exp\left(\mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{x}\right)\right) - \exp\left(\mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{x} + \frac{\beta}{\alpha}\boldsymbol{A}\boldsymbol{w}\right)\right) \right] \right\| \\ &\le \underbrace{\left\| |\boldsymbol{A}\boldsymbol{x}| \odot \mathbb{1}_{\left|\frac{\beta}{\alpha}\right||\boldsymbol{A}\boldsymbol{w}| \ge \rho |\boldsymbol{A}\boldsymbol{x}|} \right\|}_{\mathcal{T}_1} + \frac{1}{1-\rho}\left|\frac{\beta}{\alpha}\right| \underbrace{\left\| \Im\left((\boldsymbol{A}\boldsymbol{w}) \odot \exp\left(-\mathrm{i}\phi(\boldsymbol{A}\boldsymbol{x})\right) \right) \right\|}_{\mathcal{T}_2}. \end{aligned}$$

The first term $\mathcal{T}_1$ can be bounded using the *restricted isometry property* of a random circulant matrix [KMR14], together with some auxiliary analysis. The detailed analysis is provided in Section 24.4. The second term $\mathcal{T}_2$ involves a nonlinear function $\exp\left(-\mathrm{i}\phi(\boldsymbol{A}\boldsymbol{x})\right)$ of the random circulant matrix $\boldsymbol{A}$. Controlling this nonlinear, highly dependent random process for all $\boldsymbol{w}$ is a nontrivial task. In the next subsection, we explain why bounding $\mathcal{T}_2$ is technically challenging, and we sketch the key ideas about how to control a smoothed variant of $\mathcal{T}_2$, by using the weighting $\boldsymbol{b}$ introduced in (21.1.1). We also provide intuitions as to why the weighting $\boldsymbol{b}$ is helpful.

21.2.2 Controlling a smoothed variant of the phase term $\mathcal{T}_2$

As elaborated above, the major challenge of showing iterative contraction is bounding the suprema of the nonlinear, dependent random process $\mathcal{T}_2(\boldsymbol{w})$ over the set

$$\mathcal{S} \doteq \left\{ \boldsymbol{w} \in \mathbb{CS}^{n-1} \mid \boldsymbol{w} \perp \boldsymbol{x} \right\}.$$

By using the fact that $\Im(u) = \frac{1}{2\mathrm{i}}(u - \overline{u})$ for any $u \in \mathbb{C}$, we have

$$\sup_{\boldsymbol{w}\in\mathcal{S}} \mathcal{T}_2^2(\boldsymbol{w}) \;\le\; \tfrac{1}{2}\|\boldsymbol{A}\|^2 + \tfrac{1}{2}\sup_{\boldsymbol{w}\in\mathcal{S}} \left| \underbrace{\boldsymbol{w}^\top \boldsymbol{A}^\top \operatorname{diag}\left(\psi(\boldsymbol{A}\boldsymbol{x})\right) \boldsymbol{A}\boldsymbol{w}}_{\mathcal{L}(\boldsymbol{a},\boldsymbol{w})} \right|,$$

where we define $\psi(t) \doteq \exp\left(-2\mathrm{i}\phi(t)\right)$. As from (21.2.3), we know that $\|\boldsymbol{A}\| \approx \sqrt{m}$. Thus, to show (21.2.4), the major task left is to prove that

$$\sup_{\boldsymbol{w}\in\mathcal{S}} |\mathcal{L}(\boldsymbol{a},\boldsymbol{w})| < (1-\eta')m \tag{21.2.5}$$

for some constant $\eta' \in (0,1)$.

Why decoupling? Let $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1^* \\ \dots \\ \boldsymbol{a}_m^* \end{bmatrix}$, the term

$$\mathcal{L}(\boldsymbol{a}, \boldsymbol{w}) \;=\; \boldsymbol{w}^\top \boldsymbol{A}^\top \operatorname{diag}\left(\psi(\boldsymbol{A}\boldsymbol{x})\right) \boldsymbol{A}\boldsymbol{w} \;=\; \sum_{k=1}^{m} \underbrace{\psi(\boldsymbol{a}_k^* \boldsymbol{x}) \boldsymbol{w}^\top \overline{\boldsymbol{a}}_k \overline{\boldsymbol{a}}_k^\top \boldsymbol{w}}_{\text{dependence across } k}$$

is a summation of dependent random variables. To address this problem, we deploy ideas from *decoupling* [DlPG99]. Informally, decoupling allows us to compare moments of random functions to functions of more independent random variables, which are usually easier to analyze. The book [DlPG99] provides a beautiful introduction to this area. In our problem, notice that the random vector $\boldsymbol{a}$ occurs twice in the definition of $\mathcal{L}(\boldsymbol{a}, \boldsymbol{w})$ – one in the phase term $\psi(\boldsymbol{A}\boldsymbol{x}) = \exp(-2\mathrm{i}\phi(\boldsymbol{A}\boldsymbol{x}))$, and another in the quadratic term. The general spirit of decoupling is to seek to replace one of these copies of $\boldsymbol{a}$ with an *independent* copy $\boldsymbol{a}'$ of the same random vector, yielding a random process with fewer dependencies. Here, we seek to replace $\mathcal{L}(\boldsymbol{a}, \boldsymbol{w})$ with

$$\mathcal{Q}_{dec}^{\mathcal{L}}(\boldsymbol{a}, \boldsymbol{a}', \boldsymbol{w}) = \boldsymbol{w}^\top \boldsymbol{A}^\top \operatorname{diag}\left(\psi(\boldsymbol{A}'\boldsymbol{x})\right) \boldsymbol{A}\boldsymbol{w}. \tag{21.2.6}$$

The utility of this new, decoupled form $\mathcal{Q}_{dec}^{\mathcal{L}}(\boldsymbol{a}, \boldsymbol{a}', \boldsymbol{w})$ of $\mathcal{L}(\boldsymbol{a}, \boldsymbol{w})$ is that it introduces extra randomness — $\mathcal{Q}_{dec}^{\mathcal{L}}(\boldsymbol{a}, \boldsymbol{a}', \boldsymbol{w})$ is now a *chaos* process of $\boldsymbol{a}$ conditioned on $\boldsymbol{a}'$. This makes analyzing $\sup_{\boldsymbol{w} \in \mathcal{S}} \mathcal{Q}_{dec}^{\mathcal{L}}(\boldsymbol{a}, \boldsymbol{a}', \boldsymbol{w})$ amenable to existing analysis of *suprema of chaos processes* for random circulant matrices [KR14]. However, achieving the decoupling requires additional work; the most general existing results on decoupling pertain to *tetrahedral polynomials*, which are polynomials with no monomials involving any power larger than one of any random variable. By appropriately tracking cross terms, these results can also be applied to more general (non-tetrahedral) polynomials in Gaussian random variables [Kwa87]. However, our random process $\mathcal{L}(\boldsymbol{a}, \boldsymbol{w})$ involves a nonlinear phase term $\psi(\boldsymbol{A}\boldsymbol{w})$ which is not a polynomial, and hence is not amenable to a direct appeal to existing results.

Decoupling is "recoupling". Existing results [Kwa87] for decoupling polynomials of Gaussian random variables are derived from two simple facts:

1. orthogonal projections of Gaussian variables are independent;

2. Jensen's inequality.

Indeed, for the random vector $\boldsymbol{a} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$, let us introduce an independent copy $\boldsymbol{\delta} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$. Write

$$\boldsymbol{g}^1 = \boldsymbol{a} + \boldsymbol{\delta}, \qquad \boldsymbol{g}^2 = \boldsymbol{a} - \boldsymbol{\delta}.$$

Because of Fact 1, $\boldsymbol{g}^1$ and $\boldsymbol{g}^2$ are two *independent* $\mathcal{CN}(\boldsymbol{0}, 2\boldsymbol{I})$ vectors. Now, by taking conditional expectation with respect to $\boldsymbol{\delta}$, we have

$$\mathbb{E}_{\boldsymbol{\delta}}\left[\mathcal{Q}^{\mathcal{L}}_{dec}(\boldsymbol{g}^1, \boldsymbol{g}^2, \boldsymbol{w})\right] \quad = \quad \mathbb{E}_{\boldsymbol{\delta}}\left[\mathcal{Q}^{\mathcal{L}}_{dec}(\boldsymbol{a}+\boldsymbol{\delta}, \boldsymbol{a}-\boldsymbol{\delta}, \boldsymbol{w})\right] \quad \doteq \quad \widehat{\mathcal{L}}(\boldsymbol{a}, \boldsymbol{w}). \tag{21.2.7}$$

Thus, we can see that the key idea of decoupling $\mathcal{L}(\boldsymbol{a}, \boldsymbol{w})$ into $\mathcal{Q}^{\mathcal{L}}_{dec}(\boldsymbol{a}, \boldsymbol{a}', \boldsymbol{w})$, is essentially "recoupling" $\mathcal{Q}^{\mathcal{L}}_{dec}(\boldsymbol{g}^1, \boldsymbol{g}^2, \boldsymbol{w})$ via conditional expectation – the "recoupled" term $\widehat{\mathcal{L}}$ can be reviewed as an approximation of $\mathcal{L}(\boldsymbol{a}, \boldsymbol{w})$. Notice that by Fact 2, Jensen's inequality, for any convex function φ,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{a}}\left[\sup_{\boldsymbol{w}\in\mathcal{S}} \varphi\left(\widehat{\mathcal{L}}(\boldsymbol{a}, \boldsymbol{w})\right)\right] &= \mathbb{E}_{\boldsymbol{a}}\left[\sup_{\boldsymbol{w}\in\mathcal{S}} \varphi\left(\mathbb{E}_{\boldsymbol{\delta}}\left[\mathcal{Q}^{\mathcal{L}}_{dec}(\boldsymbol{a}+\boldsymbol{\delta}, \boldsymbol{a}-\boldsymbol{\delta}, \boldsymbol{w})\right]\right)\right] \\
&\le \mathbb{E}_{\boldsymbol{a},\boldsymbol{\delta}}\left[\sup_{\boldsymbol{w}\in\mathcal{S}} \varphi\left(\mathcal{Q}^{\mathcal{L}}_{dec}(\boldsymbol{a}+\boldsymbol{\delta}, \boldsymbol{a}-\boldsymbol{\delta}, \boldsymbol{w})\right)\right] \\
&= \mathbb{E}_{\boldsymbol{g}^1,\boldsymbol{g}^2}\left[\sup_{\boldsymbol{w}\in\mathcal{S}} \varphi\left(\mathcal{Q}^{\mathcal{L}}_{dec}(\boldsymbol{g}^1, \boldsymbol{g}^2, \boldsymbol{w})\right)\right].
\end{aligned}$$

Thus, by choosing φ appropriately, i.e., as $\varphi(t) = |t|^p$, we can control all the moments of $\sup_{\boldsymbol{w}\in\mathcal{S}} \widehat{\mathcal{L}}(\boldsymbol{a}, \boldsymbol{w})$ via

$$\left\|\sup_{\boldsymbol{w}\in\mathcal{S}} \left|\widehat{\mathcal{L}}(\boldsymbol{a}, \boldsymbol{w})\right|\right\|_{L^p} \le \left\|\sup_{\boldsymbol{w}\in\mathcal{S}} \left|\mathcal{Q}^{\mathcal{L}}_{dec}(\boldsymbol{g}^1, \boldsymbol{g}^2, \boldsymbol{w})\right|\right\|_{L^p}. \tag{21.2.8}$$

This type of inequality is very useful because it relates the moments of $\sup_{\boldsymbol{w}\in\mathcal{S}} \left|\widehat{\mathcal{L}}(\boldsymbol{a}, \boldsymbol{w})\right|$ to that of $\sup_{\boldsymbol{w}\in\mathcal{S}} \left|\mathcal{Q}^{\mathcal{L}}_{dec}(\boldsymbol{a}, \boldsymbol{a}', \boldsymbol{w})\right|$. As discussed previously, $\mathcal{Q}^{\mathcal{L}}_{dec}$ is a chaos process of $\boldsymbol{g}^1$ by conditioning on $\boldsymbol{g}^2$. Its moments can be bounded using existing results [KMR14].

If $\mathcal{L}$ was a tetrahedral polynomial, we have $\widehat{\mathcal{L}} = \mathcal{L}$, i.e., the approximation is exact. As the tail bound of $\sup_{\boldsymbol{w}\in\mathcal{S}} |\mathcal{L}(\boldsymbol{a}, \boldsymbol{w})|$ can be controlled via its moments bounds [FR13, Chapter 7.2], this allows us to directly control the object $\mathcal{L}$ of interest. The reason that this control obtains is because the conditional expectation operator $\mathbb{E}_{\boldsymbol{\delta}}[\cdot \mid \boldsymbol{a}]$ "recouples" $\mathcal{Q}^{\mathcal{L}}_{dec}(\boldsymbol{a}, \boldsymbol{a}', \boldsymbol{w})$ back to the target $\mathcal{L}(\boldsymbol{a}, \boldsymbol{w})$. In slogan form, (Gaussian) *decoupling is recoupling*.

"Recoupling" is Gaussian smoothing. A distinctive feature of the term $\mathcal{L}(\boldsymbol{a}, \boldsymbol{w})$ in convolutional phase retrieval is that $\psi(\cdot)$ is a phase function and therefore $\mathcal{L}$ is *not* a polynomial. Hence, it may be challenging to posit a $\mathcal{Q}^{\mathcal{L}}_{dec}$ which "recouples" back to $\mathcal{L}$. In other words, in the existing form, we need to tolerate an *approximation error* as $\widehat{\mathcal{L}} \neq \mathcal{L}$. Although $\widehat{\mathcal{L}}$ is not exactly $\mathcal{L}$, we can still control $\sup_{\boldsymbol{w}\in\mathcal{S}} |\mathcal{L}(\boldsymbol{a}, \boldsymbol{w})|$ through its

approximation $\widehat{\mathcal{L}}$,

$$\sup_{\boldsymbol{w}\in\mathcal{S}}|\mathcal{L}(\boldsymbol{a},\boldsymbol{w})| \quad\leq\quad \sup_{\boldsymbol{w}\in\mathcal{S}}\left|\widehat{\mathcal{L}}(\boldsymbol{a},\boldsymbol{w})\right| \;+\; \sup_{\boldsymbol{w}\in\mathcal{S}}\left|\widehat{\mathcal{L}}(\boldsymbol{a},\boldsymbol{w})-\mathcal{L}(\boldsymbol{a},\boldsymbol{w})\right|. \tag{21.2.9}$$

As we discussed above, the term $\sup_{\boldsymbol{w}\in\mathcal{S}}\left|\widehat{\mathcal{L}}(\boldsymbol{a},\boldsymbol{w})\right|$ can be controlled by using decoupling and the moments bound in (21.2.8). Therefore, the inequality (21.2.9) is useful to derive a sufficiently tight bound for $\mathcal{L}(\boldsymbol{a},\boldsymbol{w})$ if $\widehat{\mathcal{L}}(\boldsymbol{a},\boldsymbol{w})$ is very close to $\mathcal{L}(\boldsymbol{a},\boldsymbol{w})$ uniformly, i.e., the approximation error is small. Now the question is: *for what $\mathcal{L}$ is it possible to find a "well-behaved" $\mathcal{Q}_{dec}^{\mathcal{L}}$ for which the approximation error is small?* To understand this question, recall that the mechanism that links $\mathcal{Q}_{dec}$ to $\widehat{\mathcal{L}}$ is the conditional expectation operator $\mathbb{E}_{\boldsymbol{\delta}}\left[\cdot \mid \boldsymbol{a}\right]$. For our case, from (21.2.7) orthogonality leads to

$$\widehat{\mathcal{L}}(\boldsymbol{a},\boldsymbol{w})=\boldsymbol{w}^{\top}\boldsymbol{A}^{\top}\operatorname{diag}\left(h(\boldsymbol{A}\boldsymbol{x})\right)\boldsymbol{A}\boldsymbol{w}, \qquad h(t)\;\doteq\;\mathbb{E}_{s\sim\mathcal{CN}(0,\|\boldsymbol{x}\|^2)}\left[\psi(t+s)\right]. \tag{21.2.10}$$

Thus, by using the results in (21.2.9) and (21.2.10), we can bound $\sup_{\boldsymbol{w}\in\mathcal{S}}|\mathcal{L}(\boldsymbol{a},\boldsymbol{w})|$ as

$$\sup_{\boldsymbol{w}\in\mathcal{S}}|\mathcal{L}(\boldsymbol{a},\boldsymbol{w})| \quad\leq\quad \sup_{\boldsymbol{w}\in\mathcal{S}}\left|\widehat{\mathcal{L}}(\boldsymbol{a},\boldsymbol{w})\right| \;+\; \underbrace{\|h-\psi\|_{L^\infty}}_{\text{approximation error}}\;\|\boldsymbol{A}\|^2. \tag{21.2.11}$$

Notice that the function h is not exactly ψ, but generated by convolving ψ with a multivariate Gaussian *pdf*: indeed, *recoupling is Gaussian smoothing*. The Fourier transform of a multivariate Gaussian is again a Gaussian; it decays quickly with frequency. So, in order to admit a small approximation error, the target ψ must be *smooth*. However, in our case, the function $\psi(t)=\exp(-2\mathrm{i}\phi(t))$ is discontinuous at $t=0$; it changes extremely rapidly in the vicinity of $t=0$, and hence its Fourier transform (appropriately defined) does not decay quickly at all. Therefore, the term $\mathcal{L}(\boldsymbol{a},\boldsymbol{w})$ is a poor target for approximation by using a smooth function $\widehat{\mathcal{L}}(\boldsymbol{a},\boldsymbol{w})=\mathbb{E}_{\boldsymbol{\delta}}[\mathcal{Q}_{dec}^{\mathcal{L}}(\boldsymbol{g}^1,\boldsymbol{g}^2,\boldsymbol{w})]$. From Fig. 21.1, the difference between h and ψ increases as $|t|\searrow 0$. The poor approximation error $\|\psi-f\|_{L^\infty}=1$ results in a trivial bound for $\sup_{\boldsymbol{w}\in\mathcal{S}}|\mathcal{L}(\boldsymbol{a},\boldsymbol{w})|$ instead of (21.2.5).

Decoupling and convolutional phase retrieval. To reduce the approximation error caused by the nonsmoothness of ψ at $t=0$, we smooth ψ. More specifically, we introduce a new weighted objective (20.1.1) with Gaussian weighting $\boldsymbol{b}=\zeta_{\sigma^2}(\boldsymbol{y})$ in (21.1.1), replacing the analyzing target $\mathcal{T}_2$ with

$$\widehat{\mathcal{T}}_2=\left\|\operatorname{diag}\left(\boldsymbol{b}^{1/2}\right)\Im\left((\boldsymbol{A}\boldsymbol{w})\odot\exp\left(-\mathrm{i}\phi(\boldsymbol{A}\boldsymbol{x})\right)\right)\right\|.$$

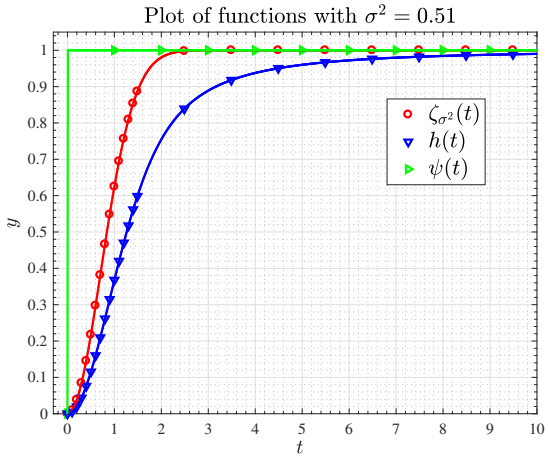


Figure 21.1: Plots of functions $h(t)$, $\psi(t)$ and $\zeta_{\sigma^2}(t)$ over the real line. The $\psi(t)$ function is discontinuous at 0, and cannot be uniformly approximated by $h(t)$. On the other hand, the function $h(t)$ serves as a good approximation of the weighting $\psi(t)$.

Consequently, we obtain a smoothed variant $\mathcal{L}_s(\boldsymbol{a}, \boldsymbol{w})$ of $\mathcal{L}(\boldsymbol{a}, \boldsymbol{w})$,

$$\mathcal{L}_s(\boldsymbol{a}, \boldsymbol{w}) = \boldsymbol{w}^\top \boldsymbol{A}^\top \operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{y}) \odot \psi(\boldsymbol{A}\boldsymbol{x})\right) \boldsymbol{A}\boldsymbol{w}.$$

Similar to (21.2.11), we obtain

$$\sup_{\boldsymbol{w}\in\mathcal{S}} |\mathcal{L}_s(\boldsymbol{a}, \boldsymbol{w})| \quad \leq \quad \sup_{\boldsymbol{w}\in\mathcal{S}} \left|\widehat{\mathcal{L}}(\boldsymbol{a}, \boldsymbol{w})\right| \; + \; \|h(t) - \zeta_{\sigma^2}(t)\psi(t)\|_{L^\infty} \|\boldsymbol{A}\|^2 .$$

Now the approximation error $\|h - \psi\|_{L^\infty}$ in (21.2.11) is replaced by $\|h(t) - \zeta_{\sigma^2}(t)\psi(t)\|_{L^\infty}$. As observed from Fig. 21.1, the function $\zeta_{\sigma^2}(t)$ smoothes $\psi(t)$ especially near the vicinity of $t = 0$, such that the new approximation error $\|f(t) - \zeta_{\sigma^2}(t)\psi(t)\|_{L^\infty}$ is significantly reduced. Thus, by using similar ideas above, we can provide a nontrivial bound

$$\sup_{\boldsymbol{w}\in\mathcal{S}} |\mathcal{L}_s(\boldsymbol{a}, \boldsymbol{w})| < (1 - \eta_s)m,$$

for some $\eta_s \in (0, 1)$, which is sufficient for showing iterative contraction. Finally, because of the weighting $\boldsymbol{b} = \zeta_{\sigma^2}(\boldsymbol{y})$, it should be noticed that the overall analysis needs to be slightly modified accordingly. For more detailed analysis, we refer the readers to Section 24.

Chapter 22

Numerical Results

In this section, we conduct some experiments on both synthetic and real dataset to demonstrate the effectiveness of the proposed method.

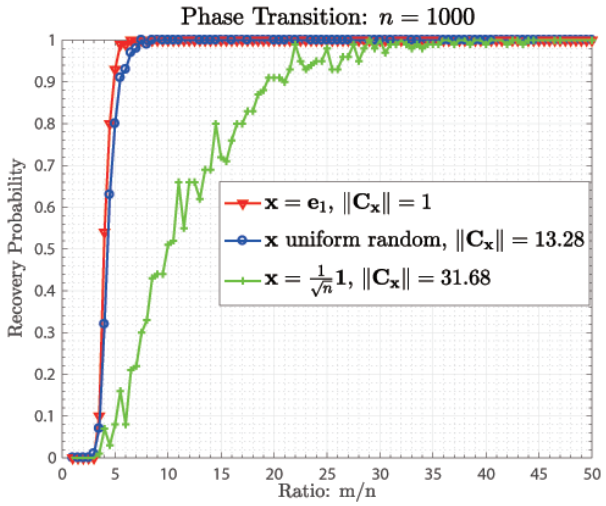


Figure 22.1: Phase transition for recovering the signal $\boldsymbol{x} \in \mathbb{CS}^{n-1}$ with different signal patterns and $\|\boldsymbol{C}_{\boldsymbol{x}}\|$.

Dependence of sample complexity on $\|\boldsymbol{C}_{\boldsymbol{x}}\|$. First, we investigate the dependence of the sample complexity m on $\|\boldsymbol{C}_{\boldsymbol{x}}\|$. We assume the ground truth $\boldsymbol{x} \in \mathbb{CS}^{n-1}$, and consider three cases:

- $\boldsymbol{x} = \boldsymbol{e}_1$ with $\boldsymbol{e}_1$ to be the standard basis vector, such that $\|\boldsymbol{C}_{\boldsymbol{x}}\| = 1$;
- $\boldsymbol{x}$ is uniformly random generated on the complex sphere $\mathbb{CS}^{n-1}$;
- $\boldsymbol{x} = \frac{1}{\sqrt{n}}\mathbf{1}$, such that $\|\boldsymbol{C}_{\boldsymbol{x}}\| = \sqrt{n}$.

For each case, we fix the signal length $n = 1000$ and vary the ratio m/n. For each ratio m/n, we randomly generate the kernel $\boldsymbol{a} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$ in (19.0.1) and repeat the experiment 100 times. We initialize the algorithm by the spectral method in Algorithm 2 and run the gradient descent (24.2.2). Given the algorithm output $\widehat{\boldsymbol{x}}$, we judge the success of recovery by

$$\inf_{\phi \in [0,2\pi)} \left\| \widehat{\boldsymbol{x}} - \boldsymbol{x} e^{\mathrm{i}\phi} \right\| \le \epsilon, \tag{22.0.1}$$

where $\epsilon = 10^{-5}$. From Fig. 22.1, for the case when $\|\boldsymbol{C}_{\boldsymbol{x}}\| = \mathcal{O}(1)$, the number of measurements needed is far less than our Theorem 21.1 suggests. Bridging the gap between the practice and theory is left for the future work.

Another observation is that the larger the $\|\boldsymbol{C}_{\boldsymbol{x}}\|$ is, the more samples we needed for the success of recovery. One possibility is that the sample complexity depends on $\|\boldsymbol{C}_{\boldsymbol{x}}\|$, another possibility is that the extra logarithmic factors in our analysis are truly necessary for worst case (here, spectral sparse) inputs.

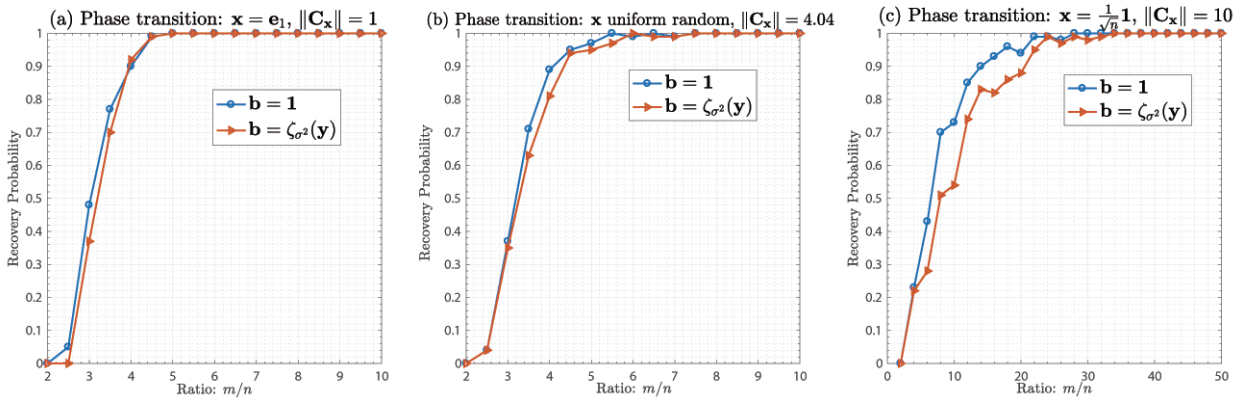


Figure 22.2: Phase transition for convolutional phase retrieval with weightings for $\boldsymbol{b}$.

Effects of weighting $\boldsymbol{b}$. Although the weighting $\boldsymbol{b}$ in (21.1.1) that we introduced in Theorem 21.1 is mainly for analysis, here we investigate the effectiveness in practice. We consider the same three cases for $\boldsymbol{x}$ as we did before. For each case, we fix the signal length $n = 100$ and vary the ratio m/n. For each ratio m/n, we randomly generate the kernel $\boldsymbol{a} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$ in (19.0.1) and repeat the experiment 100 times. We initialize the algorithm by the spectral method in Algorithm 2 and run the gradient descent (24.2.2) with weighting $\boldsymbol{b} = \boldsymbol{1}$ and $\boldsymbol{b}$ in (21.1.1), respectively. We judge success of recovery once the error (22.0.1) is smaller than 10^{-5}. From Fig. 22.2, we can see that the sample complexity is slightly larger for $\boldsymbol{b} = \zeta_{\sigma^2}(\boldsymbol{y})$, the benefits of weighting here is more for the ease of analysis.

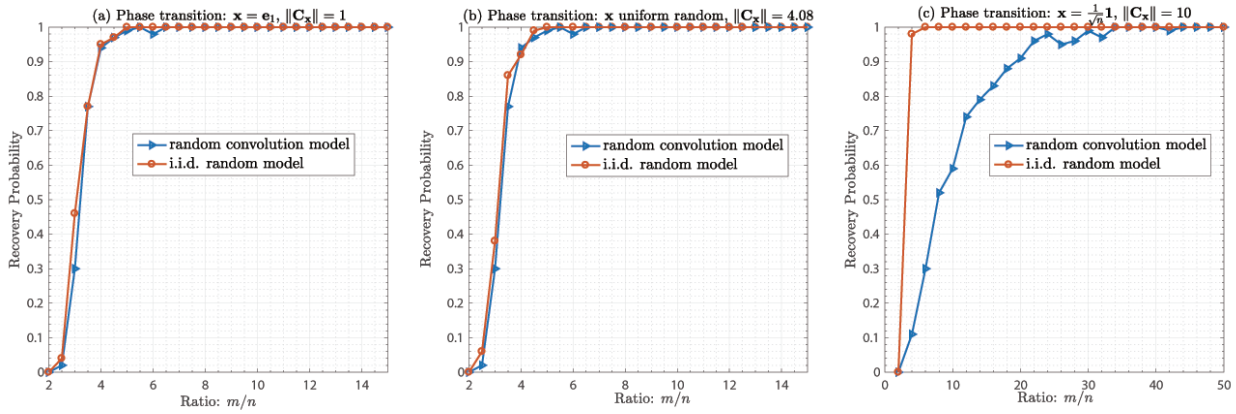


Figure 22.3: Phase transition of random convolution model vs. i.i.d. random model.

Comparison with generic random measurements. Another interesting question is that, in comparison with pure random model, how much more samples are needed for the random convolutional model in practice? We investigate this question numerically. We consider the same three cases for $\boldsymbol{x}$ as we did before, and consider two random measurement models

$$\boldsymbol{y}_1 = |\boldsymbol{a} \circledast \boldsymbol{x}|, \qquad \boldsymbol{y}_2 = |\boldsymbol{A}\boldsymbol{x}|,$$

where $\boldsymbol{a} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$, and $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1 \\ \cdots \\ \boldsymbol{a}_m \end{bmatrix}$ with $\boldsymbol{a}_k \sim_{i.i.d.} \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$. For each case, we fix the signal length $n = 100$ and vary the ratio m/n. We repeat the experiment 100 times. We initialize the algorithm by the spectral method in Algorithm 2 for both models, and run gradient descent (24.2.2). We judge success of recovery once the error (22.0.1) is smaller than 10^{-5}. From Fig. 22.3, we can see that when $\boldsymbol{x}$ is typical (e.g., $\boldsymbol{x} = \boldsymbol{e}_1$ or $\boldsymbol{x}$ is uniformly random generated from $\mathbb{CS}^{n-1}$), under the same settings, the samples needed for the two random models are almost the same. However, when $\boldsymbol{x}$ is Fourier sparse (e.g., $\boldsymbol{x} = \frac{1}{\sqrt{n}}\boldsymbol{1}$), more samples are required for the random convolution model.

Necessity of initializations. As has been shown in [SQW16, Sol17], for phase retrieval with generic measurement, when the sample complexity satisfies $m \geq \Omega(n \log n)$, with high probability the landscape of the nonconvex objective (19.1.1) is nice enough that it enables initialization free global optimization. This raises an interesting question that whether spectral initialization is still necessary for the random convolutional model. We consider similar setting as the previous experiment, where the ground truth $\boldsymbol{x} \in \mathbb{C}^n$ is drawn uniformly random from $\mathbb{CS}^{n-1}$. We fix the dimension $n = 1000$ and change the ratio m/n. For each ratio, we

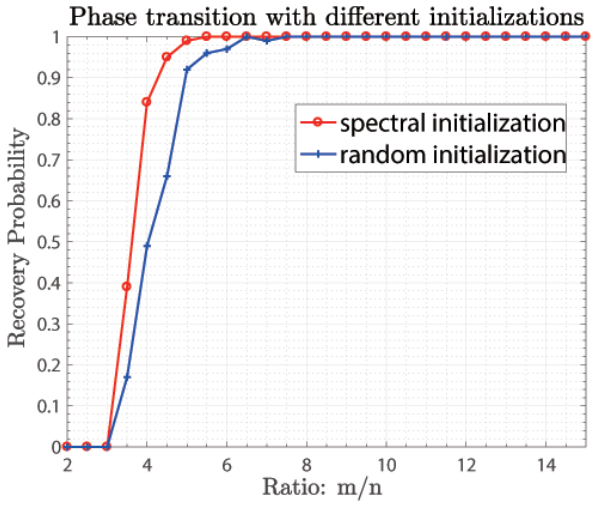


Figure 22.4: Phase transition for convolutional phase retrieval with different initialization schemes, where $\boldsymbol{x}$ is generated uniformly random from $\mathbb{CS}^{n-1}$.

randomly generate the kernel $\boldsymbol{a} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$ in (19.0.1) and repeat the experiment for 100 times. For each instance, we start the algorithm from random initialization and spectral initialization, respectively. We choose the stepsize via backtracking linesearch and terminate the experiment either when the number iteration is larger than 2×10^4 or the distance of the iterate to the solution is smaller than 1×10^{-5}. As we can see from Fig. 22.4, the sample number required for successful recovery with random initializations is only slightly more than that with the spectral initialization. This implies that the spectral initialization is not that critical for the random convolutional model, neither.

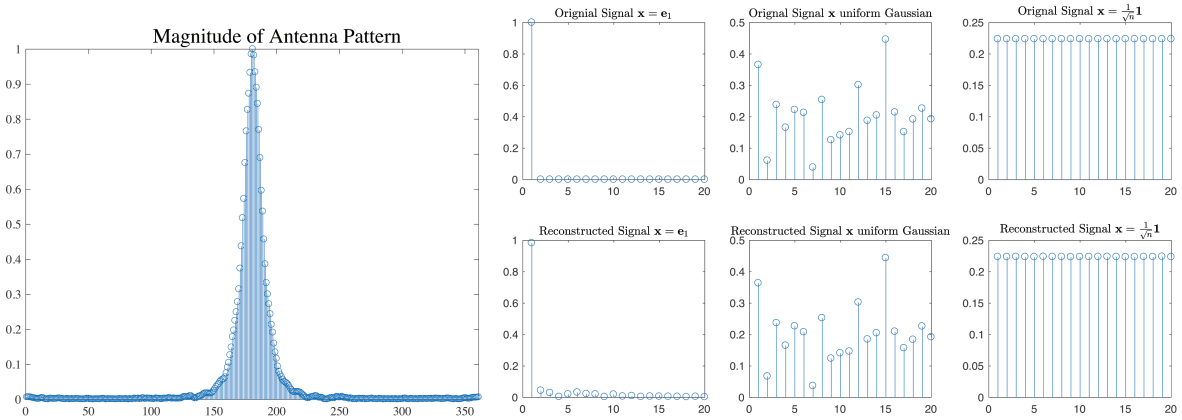


Figure 22.5: Experiment on real data.

Experiments on real antenna data for 5G communication. Finally, we demonstrate the effectiveness of the proposed method on a problem arising in 5G communication, as we mentioned in the introduction. Fig. 22.6 (left) shows an antenna pattern $\boldsymbol{a} \in \mathbb{C}^{361}$ obtained from Bell labs. We observe the modulus of the convolution of this pattern with the signal of interest. For three different types of signals with length $n = 20$, (1) $\boldsymbol{x} = \boldsymbol{e}_1$, (2) $\boldsymbol{x}$ is uniformly random generated from $\mathbb{CS}^{n-1}$, (3) $\boldsymbol{x} = \frac{1}{\sqrt{n}}\mathbf{1}$, our result in Fig. 22.6 shows that we can achieve almost perfect recovery.

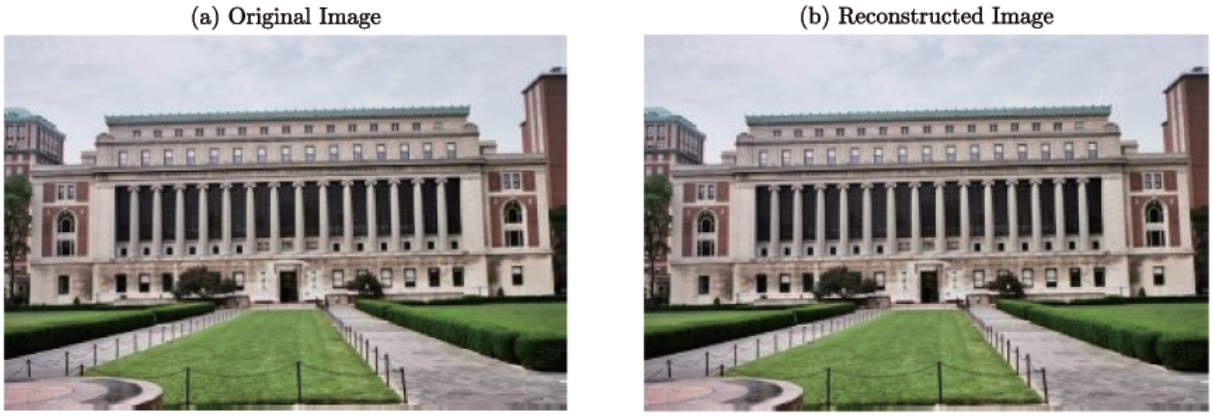


Figure 22.6: Experiment on real images.

Experiments on real image. Finally, we run the experiment on some real dataset to demonstrate the effectiveness and the efficiency of the proposed method. We choose an image of size 200×300 as in Fig. 22.6, we use $m = 5n \log n$ samples for reconstruction. The kernel $\boldsymbol{a} \in \mathbb{C}^m$ is randomly generated as complex Gaussian $\mathcal{CN}(\mathbf{0}, \boldsymbol{I})$. We run power method for 100 iterations for initialization, and stop the algorithm once the error is smaller than 1×10^{-4}. It takes $197.08s$ to reconstruct all the RGB channels. Experiment using general Gaussian measurements $\boldsymbol{A} \in \mathbb{C}^{m \times n}$ could easily run out of memory on a personal computer for problems of this size.

Chapter 23

Discussion

In this part of the thesis, we showed that via nonconvex optimization, the phase retrieval problem with random convolutional measurement can be solved to global optimum with $m \geq \Omega\left(\frac{\|\boldsymbol{C}_{\boldsymbol{x}}\|^2}{\|\boldsymbol{x}\|^2} n \operatorname{poly}\log n\right)$ samples. Our result raises several interesting questions that we discuss below.

Tightening sample complexity. Our estimate of the sample complexity is only tight up to logarithm factors: there is a substantial gap between our theory and practice for the dependence of the logarithm factors. We believe the high order dependence of the logarithm factors is an artifact of our analysis. In particular, our analysis in Section 24.4 is based on the result of RIP conditions for partial circulant random matrices, which is no way tight. We believe that by using advanced tools in probability, the sample complexity can be tightened to at least $m \geq \Omega\left(n \log^6 n\right)$, which is left for future work.

Geometric analysis and global result. Our convergence analysis is based on showing iterative contraction of gradient descent methods. However, it would be interesting if we could characterize the function landscape of nonconvex objectives as we did in [SQW16]. Such a result would provide a better explanation why the gradient descent method works, and help us design more efficient algorithms. The major difficulty we encountered is the lack of probability tools for analyzing the random convolutional model: because of the nonhomogeneity of $\|\boldsymbol{C}_{\boldsymbol{z}}\|$, it is hard to tightly uniformize quantities of random convolutional matrices over the complex sphere $\mathbb{CS}^{n-1}$: our preliminary analysis results in suboptimal bounds for sample complexity. We hope this work can invite more ideas for theoretical understandings of this problem.

Tools for analyzing other structured nonconvex problems. This work is part of a recent surge of research efforts on deriving provable and practical nonconvex algorithms to central problems in modern signal processing and machine learning [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, SA14, LWB13, LJ15, LLJB15, EW15, Bou16, JJKN15]. On the other hand, we believe the probability tools of decoupling and measure concentration we developed here laid a solid foundation for studying other nonconvex problem under the random convolutional model. Those problems include blind calibration [LS15, CJ16, LS16], blind deconvolution [LWDF09, ETS11, CM14b, ARR14, LLJB15, LLSW16, LTR16, LS17], and convolutional dictionary learning[BEL13, BL14, HHW15, HA], etc.

Application ideas. Finally, despite the cases we mentioned in the introduction, the application of convolutional phase retrieval seems ubiquitous in many signal processing problems but largely unexplored. We hope that the algorithm and theoretical guarantees we developed here could invite and inspire more application ideas.

Chapter 24

Proof of Technical Results

In this section, we provide the detailed proof of Theorem 21.1. The whole section is organized as follows. In Subsection 24.1, we show that the the initialization produced by Algorithm 2 is close to the optimum. In Subsection 24.2, we sketch the proof of our main result, i.e., Theorem 21.1, where some key proofing details is provided in Subsection 24.3. All the other supporting results are provided subsequently. We provide detailed proofs of two key supporting lemmas in Subsection 24.4 and Subsection 24.5, respectively. Finally, other supporting lemmas are postponed to the appendices: in Appendix B.1, we introduce the elementary tools and results that are useful throughout analysis; in Appendix B.2, we provide results of bounding the suprema of chaos processes for random circulant matrices. In Appendix B.3, we provide concentration results for suprema of some dependent random processes via decoupling.

24.1 Spectral Initialization

Proposition 24.1 *Suppose $\boldsymbol{z}_0$ is produced by Algorithm 2. Given a fixed scalar $\delta > 0$, whenever $m \geq C\delta^{-2} n \log^7 n$, we have*

$$\operatorname{dist}^2 (\boldsymbol{z}_0, \mathcal{X}) \leq \delta \|\boldsymbol{x}\|^2$$

with probability at least $1 - c_1 m^{-c_2}$. Here c_1, c_2 and C are some positive numerical constants.

The proof is similar to that of [Sol14], while we here are handling random circulant matrices. We sketch the main ideas of the proof below, some detailed analysis is retained to Appendix B.2 and Appendix B.3.

Proof Without loss of generality, we assume that $\|\boldsymbol{x}\| = 1$. Let $\widetilde{\boldsymbol{z}}_0$ be the leading eigenvector of

$$\boldsymbol{Y} = \frac{1}{m}\sum_{k=1}^{m} \left|\boldsymbol{a}_k^* \boldsymbol{x}\right|^2 \boldsymbol{a}_k \boldsymbol{a}_k^*$$

with $\|\widetilde{\boldsymbol{z}}_0\| = 1$, and let σ_1 be the corresponding eigenvalue. We have

$$\operatorname{dist}(\boldsymbol{z}_0, \mathcal{X}) \le \|\boldsymbol{z}_0 - \widetilde{\boldsymbol{z}}_0\| + \operatorname{dist}\left(\widetilde{\boldsymbol{z}}_0, \mathcal{X}\right).$$

First, since $\boldsymbol{z}_0 = \lambda \widetilde{\boldsymbol{z}}_0$, we have

$$\|\boldsymbol{z}_0 - \widetilde{\boldsymbol{z}}_0\| = |\lambda - 1|.$$

By Theorem B.12 in Appendix B.2, for any $\varepsilon > 0$, whenever $m \ge C\varepsilon^{-2} n \log^4 n$, we know that

$$|\lambda - 1| \le \left|\lambda^2 - 1\right| = \left|\frac{1}{m}\sum_{k=1}^{m} \left|\boldsymbol{a}_k^* \boldsymbol{x}\right|^2 - 1\right| \le \varepsilon/2 \tag{24.1.1}$$

with probability at least $1 - 2m^{-c\log^3 n}$, where $c, C > 0$ are some numerical constants. On the other hand, we have

$$\operatorname{dist}^2(\widetilde{\boldsymbol{z}}_0, \mathcal{X}) = \arg\min_{\theta} \left\|\widetilde{\boldsymbol{z}}_0 - \boldsymbol{x} e^{\mathrm{i}\theta}\right\|^2 = 2 - 2\left|\boldsymbol{x}^* \widetilde{\boldsymbol{z}}_0\right|.$$

Theorem B.21 in Appendix B.3 implies that for any $\delta > 0$, whenever $m \ge C'\delta^{-2} n \log^7 n$

$$\left\|\boldsymbol{Y} - \left(\boldsymbol{x}\boldsymbol{x}^* + \|\boldsymbol{x}\|^2 \boldsymbol{I}\right)\right\| \le \delta,$$

with probability at least $1 - 2m^{-c_1}$. Here $c_1 > 0$ is some numerical constant. It further implies that

$$\left|\widetilde{\boldsymbol{z}}_0^* \boldsymbol{Y} \widetilde{\boldsymbol{z}}_0 - \left|\widetilde{\boldsymbol{z}}_0^* \boldsymbol{x}\right|^2 - 1\right| \le \delta,$$

so that

$$\left|\widetilde{\boldsymbol{z}}_0^* \boldsymbol{x}\right|^2 \ge \sigma_1 - 1 - \delta,$$

where σ_1 is the top singular value of $\boldsymbol{Y}$. Since σ_1 is the top singular value, we have

$$\sigma_1 \ge \boldsymbol{x}^* \boldsymbol{Y} \boldsymbol{x} = \boldsymbol{x}^* (\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{x}^* - \|\boldsymbol{x}\|^2 \boldsymbol{I}) \boldsymbol{x} + 2 \ge 2 - \delta.$$

Thus, for $\delta > 0$ sufficiently small, we obtain

$$\operatorname{dist}^2(\widetilde{\boldsymbol{z}}_0, \mathcal{X}) \le 2 - 2\sqrt{1 - 2\delta} \le 2\delta. \tag{24.1.2}$$

Choose $\delta = \varepsilon^2/8$, and combining the results in (24.1.1) and (24.1.2), we obtain

$$\operatorname{dist}(\boldsymbol{z}_0, \mathcal{X}) \leq \|\boldsymbol{z}_0 - \widetilde{\boldsymbol{z}}_0\| + \operatorname{dist}(\widetilde{\boldsymbol{z}}_0, \mathcal{X}) \leq \varepsilon,$$

holds with high probability. ■

24.2 Proof of Main Result

In this section, we proof Theorem 21.1 in the main manuscript, where we restate the result as below.

Theorem 24.2 (Main Result) *Whenever $m \geq C_0 n \log^{31} n$, Algorithm 2 produces an initialization $\boldsymbol{z}^{(0)}$ that satisfies*

$$\operatorname{dist}\left(\boldsymbol{z}^{(0)}, \mathcal{X}\right) \;\leq\; c_0 \log^{-6} n \, \|\boldsymbol{x}\|$$

with probability at least $1 - c_1 m^{-c_2}$. Suppose $\boldsymbol{b} = \zeta_{\sigma^2}(\boldsymbol{y})$, where

$$\zeta_{\sigma^2}(t) = 1 - 2\pi\sigma^2 \xi_{\sigma^2}(t), \qquad \xi_{\sigma^2}(t) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{|t|^2}{2\sigma^2}\right), \tag{24.2.1}$$

with $\sigma^2 > 1/2$. Starting from $\boldsymbol{z}^{(0)}$, with $\sigma^2 = 0.51$ and stepsize $\tau = 2.02$, whenever $m \geq C_1 \frac{\|\boldsymbol{C}_{\boldsymbol{x}}\|^2}{\|\boldsymbol{x}\|^2} \max\left\{\log^{17} n, n \log^4 n\right\}$, with probability at least $1 - c_3 m^{-c_4}$ for all iterate $\boldsymbol{z}^{(r)} (r \geq 1)$ defined in (24.2.2), we have

$$\operatorname{dist}\left(\boldsymbol{z}^{(r)}, \mathcal{X}\right) \;\leq\; (1-\varrho)^r \operatorname{dist}\left(\boldsymbol{z}^{(0)}, \mathcal{X}\right),$$

holds for some small numerical constant $\varrho \in (0, 1)$. Here, c_0, c_1, c_2, c_3, c_4 and C_0, C_1 are positive numerical constants.

Our proof critically depends on the following result, where we show that with high probability for every $\boldsymbol{z} \in \mathbb{C}^n$ close enough to the optimal set $\mathcal{X}$, the iterate

$$\widehat{\boldsymbol{z}} = \boldsymbol{z} - \tau \frac{\partial}{\partial \boldsymbol{z}} f(\boldsymbol{z}), \tag{24.2.2}$$

$$\frac{\partial}{\partial \boldsymbol{z}} f(\boldsymbol{z}) = \frac{1}{m} \boldsymbol{A}^* \operatorname{diag}(\boldsymbol{b}) \left[\boldsymbol{A}\boldsymbol{z} - \boldsymbol{y} \odot \exp\left(\mathrm{i}\phi(\boldsymbol{A}\boldsymbol{z})\right)\right]. \tag{24.2.3}$$

is a contraction.

Proposition 24.3 (Iterative Contraction) *Let $\sigma^2 = 0.51$ and the stepsize $\tau = 2.02$. There exists some positive constants c_1, c_2, c_3 and C, whenever $m \geq C\frac{\|\boldsymbol{C}_{\boldsymbol{x}}\|^2}{\|\boldsymbol{x}\|^2}\max\left\{\log^{17} n, n\log^4 n\right\}$, with probability at least $1-c_1 m^{-c_2}$ for every $\boldsymbol{z} \in \mathbb{C}^n$ such that $\operatorname{dist}(\boldsymbol{z}, \mathcal{X}) \leq c_3 \log^{-6} n \|\boldsymbol{x}\|$, we have*

$$\operatorname{dist}\left(\boldsymbol{z} - \tau\frac{\partial}{\partial \boldsymbol{z}} f(\boldsymbol{z}), \mathcal{X}\right) \leq (1-\varrho)\operatorname{dist}(\boldsymbol{z}, \mathcal{X})$$

holds for some small constant $\varrho \in (0,1)$. Here, $\frac{\partial}{\partial \boldsymbol{z}} f(\boldsymbol{z})$ is defined in (24.2.3).

We sketch the main idea of the proof below. More detailed analysis is postponed to Section 24.3, Section 24.4 and Section 24.5.

Proof [Proof of Proposition 24.3] Without loss of generality, throughout the analysis we assume that $\|\boldsymbol{x}\| = 1$. By (24.2.2) and (24.2.3), and with the choice of stepsize $\tau = 2\sigma^2 + 1$, we have

$$\begin{aligned}\widehat{\boldsymbol{z}} &= \boldsymbol{z} - \frac{2\sigma^2+1}{m}\boldsymbol{A}^* \operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{y})\right)\left[\boldsymbol{A}\boldsymbol{z} - \boldsymbol{y}\odot\exp\left(\mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{z}\right)\right)\right] \\ &= \boldsymbol{z} - \boldsymbol{M}\boldsymbol{z} + \frac{2\sigma^2+1}{m}\boldsymbol{A}^* \operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{y})\right)\left[\boldsymbol{y}\odot\exp\left(\mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{z}\right)\right)\right],\end{aligned}$$

where we define

$$\boldsymbol{M}(\boldsymbol{a}) = \frac{2\sigma^2+1}{m}\boldsymbol{A}^* \operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{y})\right)\boldsymbol{A}. \tag{24.2.4}$$

For any $\boldsymbol{z} \in \mathbb{C}^n$, let us decompose $\boldsymbol{z}$ as

$$\boldsymbol{z} = \alpha\boldsymbol{x} + \beta\boldsymbol{w}, \tag{24.2.5}$$

where $\alpha, \beta \in \mathbb{C}$, and $\boldsymbol{w} \in \mathbb{CS}^{n-1}$ with $\boldsymbol{w} \perp \boldsymbol{x}$, and $\alpha = |\alpha| e^{\mathrm{i}\phi(\alpha)}$ with the phase $\phi(\alpha)$ of α satisfies $e^{\mathrm{i}\phi(\alpha)} = \boldsymbol{x}^*\boldsymbol{z}/|\boldsymbol{x}^*\boldsymbol{z}|$. Therefore, if we let

$$\theta = \arg\min_{\overline{\theta}\in[0,2\pi)}\left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\overline{\theta}}\right\|, \tag{24.2.6}$$

then we also have $\phi(\alpha) = \theta$. Thus, by using the results above, we observe

$$\operatorname{dist}^2(\widehat{\boldsymbol{z}}, \mathcal{X}) = \min_{\overline{\theta}\in[0,2\pi)}\left\|\widehat{\boldsymbol{z}} - \boldsymbol{x}e^{\mathrm{i}\overline{\theta}}\right\|^2 \leq \left\|\widehat{\boldsymbol{z}} - e^{\mathrm{i}\theta}\boldsymbol{x}\right\|^2 \leq \left\|\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{d}\right\|^2 + \left\|\boldsymbol{P}_{\boldsymbol{x}}\boldsymbol{d}\right\|^2,$$

where we define

$$\boldsymbol{d}(\boldsymbol{z}) \doteq (\boldsymbol{I} - \boldsymbol{M})\left(\boldsymbol{z} - e^{\mathrm{i}\theta}\boldsymbol{x}\right) - e^{\mathrm{i}\theta}\boldsymbol{M}\boldsymbol{x} + \frac{2\sigma^2+1}{m}\boldsymbol{A}^* \operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{y})\right)\left[\boldsymbol{y}\odot\exp\left(\mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{z}\right)\right)\right]. \tag{24.2.7}$$

Let $\delta > 0$, by Lemma 24.7 and Lemma 24.8, whenever $m \geq C\|\boldsymbol{C}_{\boldsymbol{x}}\|^2 \max\left\{\log^{17} n, \delta^{-2} n\log^4 n\right\}$, with proba-

bility at least $1 - c_1 m^{-c_2}$ for all $\boldsymbol{z} \in \mathbb{C}^n$ such that $\left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\| \le c_3\delta^3 \log^{-6} n$, we have

$$\|\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{d}\| \le \left[\delta + (1+\delta)\left(\frac{4\delta}{\rho}\sqrt{2\sigma^2+1} + \frac{1}{1-\rho}\frac{1}{1-\delta}\sqrt{\frac{1+(2+\varepsilon)\delta+(1+\delta)\Delta_\infty(\varepsilon)}{2}}\right)\right]\left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\|$$
$$\|\boldsymbol{P}_{\boldsymbol{x}}\boldsymbol{d}\| \le \left(\frac{2\sigma^2}{1+2\sigma^2} + c_{\sigma^2}\delta\right)\left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\|$$

holds for any $\rho \in (0,1)$, where $\Delta_\infty(\varepsilon)$ be defined in (24.5.1) with $\varepsilon \in (0,1)$. Here, $c_1, c_2, c_3, c_{\sigma^2}$ and C are some positive numerical constants, where c_{σ^2} is only depending on σ^2. With $\varepsilon = 0.2$ and $\sigma^2 = 0.51$, Lemma 24.7 implies that $\Delta(\varepsilon) \le 0.404$. Thus, we have

$$\begin{aligned}\|\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{d}\| &\le \left[\delta + (1+\delta)\left(\frac{5.686\delta}{\rho} + \frac{1}{1-\rho}\frac{1}{1-\delta}\sqrt{\frac{1+2.2\delta+0.404(1+\delta)}{2}}\right)\right]\left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\| \\ \|\boldsymbol{P}_{\boldsymbol{x}}\boldsymbol{d}\| &\le (0.505 + c_{\sigma^2}\delta)\left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\|.\end{aligned}$$

By choosing the constants δ and ρ sufficiently small, direct calculation reveals that

$$\operatorname{dist}^2(\widehat{\boldsymbol{z}}, \mathcal{X}) \le \|\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{d}\|^2 + \|\boldsymbol{P}_{\boldsymbol{x}}\boldsymbol{d}\|^2 \le 0.96\left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\|^2 = 0.96\operatorname{dist}^2(\boldsymbol{z}, \mathcal{X}),$$

as desired. ∎

Now with Proposition 24.3 in hand, we are ready to prove Theorem 24.2.

Proof [Proof of Theorem 3.1] We prove the theorem by recursion. Let us assume that the properties in Proposition 24.3 holds, which happens on an event $\mathcal{E}$ with probability at least $1 - c_1 m^{-c_2}$ for some numerical constants $c_1, c_2 > 0$. By Proposition 24.1 in Appendix 24.1, for any numerical constant $\delta > 0$, whenever $m \ge C\delta^{-12} n \log^{31} n$, the initialization $\boldsymbol{z}^{(0)}$ produced by Algorithm 2 satisfies

$$\operatorname{dist}\left(\boldsymbol{z}^{(0)}, \mathcal{X}\right) \le c_3\delta^3 \log^{-6} n \|\boldsymbol{x}\|,$$

with probability at least $1 - c_4 m^{-c_5}$, for some constants. Here, c_3, c_4, c_5 and $C > 0$ are some numerical constants. Therefore, conditioned on the event $\mathcal{E}$, we know that

$$\operatorname{dist}\left(\boldsymbol{z}^{(1)}, \mathcal{X}\right) = \operatorname{dist}\left(\boldsymbol{z}^{(0)} - \tau\frac{\partial}{\partial \boldsymbol{z}}f(\boldsymbol{z}), \mathcal{X}\right) \le (1-\varrho)\operatorname{dist}\left(\boldsymbol{z}^{(0)}, \mathcal{X}\right)$$

holds for some small constant $\varrho \in (0,1)$. This proves (21.1.2) for the first iteration $\boldsymbol{z}^{(1)}$. Notice that the inequality above also implies that $\operatorname{dist}\left(\boldsymbol{z}^{(1)}, \mathcal{X}\right) \le c_3\delta^3 \log^{-6} n \|\boldsymbol{x}\|$. Therefore, by reapplying the same reasoning, we can prove (21.1.2) for the iterations $r = 2, 3, \cdots$. ∎

24.3 Bounding $\|\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{d}(\boldsymbol{z})\|$ and $\|\boldsymbol{P}_{\boldsymbol{x}}\boldsymbol{d}(\boldsymbol{z})\|$

Let $\boldsymbol{d}(\boldsymbol{z})$ be defined as in (24.2.7) and assume that $\|\boldsymbol{x}\| = 1$. In this section, we provide bounds for $\|\boldsymbol{P}_{\boldsymbol{x}}\boldsymbol{d}\|$ and $\|\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{d}\|$ under the condition that $\boldsymbol{z}$ and $\boldsymbol{x}$ are close. Before presenting the main results, let us first introduce some useful preliminary lemmas. First, based on the decomposition of $\boldsymbol{z}$ in (24.2.5) and the definition of θ in (24.2.6), we can show the following result.

Lemma 24.4 *Let* $\theta = \arg\min_{\bar{\theta}\in[0,2\pi)} \left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\bar{\theta}}\right\|$ *and suppose* $\operatorname{dist}(\boldsymbol{z},\boldsymbol{x}) = \left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\| \le \epsilon$ *for some* $\epsilon \in (0,1)$*, then we have*

$$\left|\frac{\beta}{\alpha}\right| \le \frac{1}{1-\epsilon}\left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\|$$

Proof Given the facts in (24.2.5) and (24.2.6) that $\boldsymbol{z} = \alpha\boldsymbol{x} + \beta\boldsymbol{w}$ with $\boldsymbol{w} \in \mathbb{CS}^{n-1}$ and $\boldsymbol{w} \perp \boldsymbol{x}$, and $\phi(\alpha) = \theta$, we have

$$\left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\|^2 = (|\alpha| - 1)^2 + |\beta|^2 .$$

This implies that

$$|\beta| \le \left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\|,\ |\alpha| \ge 1 - \left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\| \implies \left|\frac{\beta}{\alpha}\right| \le \frac{\left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\|}{1 - \left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\|} \le \frac{1}{1-\epsilon}\left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\|,$$

as desired. ∎

Our proof is also critically depends on the concentration of $\boldsymbol{M}(\boldsymbol{a})$ in Theorem B.24 in Appendix B.3, and the following Lemmas. Please refer to Section24.4 and Section 24.5 for the detailed proofs.

Lemma 24.5 *For any given scalar* $\delta \in (0,1)$*, let* $\gamma = c_0\delta^3\log^{-6} n$*, whenever* $m \ge C\max\left\{\|\boldsymbol{C}_{\boldsymbol{x}}\|^2\log^{17} n, \delta^{-2}n\log^4 n\right\}$*, with probability at least* $1 - c_1 m^{-c_2}$ *for all* $\boldsymbol{w}$ *with* $\|\boldsymbol{w}\| \le \gamma\|\boldsymbol{x}\|$*, the inequality*

$$\left\|\boldsymbol{A}\boldsymbol{x}\odot\boldsymbol{1}_{|\boldsymbol{A}\boldsymbol{w}|\ge|\boldsymbol{A}\boldsymbol{x}|}\right\| \le \delta\sqrt{m}\,\|\boldsymbol{w}\|$$

holds. Here, c_0, c_1, c_2 *and* C *are some positive numerical constants.*

Lemma 24.6 *For any scalar* $\delta \in (0,1)$*, whenever* $m \ge C\|\boldsymbol{C}_{\boldsymbol{x}}\|^2\delta^{-2}n\log^4 n$*, with probability at least* $1 -$

$cm^{-c' \log^3 n}$ for all $\boldsymbol{w} \in \mathbb{C}^n$ with $\boldsymbol{w} \perp \boldsymbol{x}$, we have

$$\left\| \sqrt{\frac{2\sigma^2+1}{m}} \operatorname{diag}\left(\zeta_{\sigma^2}^{1/2}(\boldsymbol{y})\right) \Im\left(\boldsymbol{A}\boldsymbol{w} \odot \exp\left(-\mathrm{i}\phi(\boldsymbol{A}\boldsymbol{x})\right)\right) \right\|^2 \le \frac{1+(2+\varepsilon)\delta+(1+\delta)\Delta_\infty(\varepsilon)}{2} \|\boldsymbol{w}\|^2$$

holds. Here c, c' are some numerical constants. In particular, when $\sigma^2 = 0.51$ and $\varepsilon = 0.2$, we have $\Delta(\varepsilon) \le 0.404$. With the same probability for all $\boldsymbol{w} \in \mathbb{C}^n$ with $\boldsymbol{w} \perp \boldsymbol{x}$, we have

$$\left\| \sqrt{\frac{2\sigma^2+1}{m}} \operatorname{diag}\left(\zeta_{\sigma^2}^{1/2}(\boldsymbol{y})\right) \Im\left(\boldsymbol{A}\boldsymbol{w} \odot \exp\left(-\mathrm{i}\phi(\boldsymbol{A}\boldsymbol{x})\right)\right) \right\|^2 \le \frac{1+2.2\delta+0.404(1+\delta)}{2} \|\boldsymbol{w}\|^2 .$$

24.3.0.1 Bounding the "$\boldsymbol{x}$-perpendicular" term $\|\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{d}\|$

Lemma 24.7 *Let $\boldsymbol{d}$ be defined in (24.2.7), and suppose $\sigma^2 > 1/2$ be a constant. For any $\delta > 0$, whenever $m \ge C \|\boldsymbol{C}_{\boldsymbol{x}}\|^2 \max\left\{\log^{17} n, \delta^{-2} n \log^4 n\right\}$, with probability at least $1 - c_1 m^{-c_2}$ for all $\boldsymbol{z} \in \mathbb{C}^n$ such that $\left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\| \le c_3 \delta^3 \log^{-6} n$, we have*

$$\|\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{d}\| \le \left[\delta + (1+\delta)\left(\frac{4\delta}{\rho}\sqrt{2\sigma^2+1} + \frac{1}{1-\rho}\frac{1}{1-\delta}\sqrt{\frac{1+2\delta+(1+\delta)\Delta_\infty(\varepsilon)}{2}}\right)\right] \left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\| .$$

Here, $\Delta_\infty(\varepsilon)$ are defined in (24.5.1) for any scalar $\varepsilon \in (0,1)$, and c_1, c_2, c_3 and C are some numerical constants. In particular, when $\varepsilon = 0.2$ and $\sigma^2 = 0.51$, we have $\Delta_\infty(\varepsilon) \le 0.404$.

The analysis of bounding $\|\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{d}\|$ is similar to that of [Wal16].

Proof By the definition (24.2.7) of $\boldsymbol{d}(\boldsymbol{z})$, notice that

$$\begin{aligned}\|\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{d}\| \le& \left\|\boldsymbol{P}_{\boldsymbol{x}^\perp}\left\{\frac{2\sigma^2+1}{m}\boldsymbol{A}^* \operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{y})\right)\left[\boldsymbol{y} \odot \exp\left(\mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{z}\right)\right)\right] - e^{\mathrm{i}\theta}\boldsymbol{M}\boldsymbol{x}\right\}\right\| \\ &+ \|\boldsymbol{P}_{\boldsymbol{x}^\perp}(\boldsymbol{I}-\boldsymbol{M})\| \left\|\boldsymbol{z} - e^{\mathrm{i}\theta}\boldsymbol{x}\right\|\end{aligned}$$

For the second term, by Theorem B.24, for any $\delta > 0$, whenever $m \ge C_1 \delta^{-2} \|\boldsymbol{C}_{\boldsymbol{x}}\|^2 n \log^4 n$, we have

$$\|\boldsymbol{P}_{\boldsymbol{x}^\perp}(\boldsymbol{I}-\boldsymbol{M})\| \le \delta, \tag{24.3.1}$$

with probability at least $1 - c_1 m^{-c_2 \log^3 n}$. For the first term, we observe

$$\begin{aligned}&\left\|\boldsymbol{P}_{\boldsymbol{x}^\perp}\left\{\frac{2\sigma^2+1}{m}\boldsymbol{A}^* \operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{y})\right)\left[\boldsymbol{y} \odot \exp\left(\mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{z}\right)\right)\right] - e^{\mathrm{i}\theta}\boldsymbol{M}\boldsymbol{x}\right\}\right\| \\ =& \left\|\boldsymbol{P}_{\boldsymbol{x}^\perp}\left\{\frac{2\sigma^2+1}{m}\boldsymbol{A}^* \operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{y})\right)\left(|\boldsymbol{A}\boldsymbol{x}| \odot \left[\exp\left(\mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{z}\right)\right) - \exp\left(\mathrm{i}\theta + \mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{x}\right)\right)\right]\right)\right\}\right\| \\ \le& \left\|\sqrt{\frac{2\sigma^2+1}{m}}\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{A}^* \operatorname{diag}\left(\zeta_{\sigma^2}^{1/2}(\boldsymbol{y})\right)\right\| \times\end{aligned}$$

$$\left\| \sqrt{\frac{2\sigma^2+1}{m}} \operatorname{diag}\left(\zeta_{\sigma^2}^{1/2}(\boldsymbol{y})\right) \left(|\boldsymbol{A}\boldsymbol{x}| \odot \left[\exp\left(\mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{z}\right)\right) - \exp\left(\mathrm{i}\theta + \mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{x}\right)\right)\right]\right) \right\|.$$

By Theorem B.24 and Lemma B.30 in Appendix B.3, for any $\delta > 0$, whenever $m \geq C_1\delta^{-2} \left\|\boldsymbol{C}_{\boldsymbol{x}}\right\|^2 n \log^4 n$, we have

$$\begin{aligned}\left\| \sqrt{\frac{2\sigma^2+1}{m}} \boldsymbol{P}_{\boldsymbol{x}^\perp} \boldsymbol{A}^* \operatorname{diag}\left(\zeta_{\sigma^2}^{1/2}(\boldsymbol{y})\right) \right\| &\leq \|\boldsymbol{H}\|^{1/2} \leq \left(\|\mathbb{E}\left[\boldsymbol{H}\right]\| + \|\boldsymbol{H} - \mathbb{E}\left[\boldsymbol{H}\right]\|\right)^{1/2} \\ &\leq \left(1+\delta\right)^{1/2} \leq 1+\delta,\end{aligned}$$

with probability at least $1 - c_1 m^{-c_2 \log^3 n}$. And by Lemma B.1 and decomposition of $\boldsymbol{z}$ in (24.2.5) with $\phi(\alpha) = \theta$, we obtain

$$\begin{aligned}&\left\| \sqrt{\frac{2\sigma^2+1}{m}} \operatorname{diag}\left(\zeta_{\sigma^2}^{1/2}(\boldsymbol{y})\right) \left(|\boldsymbol{A}\boldsymbol{x}| \odot \left[\exp\left(\mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{z}\right)\right) - \exp\left(\mathrm{i}\theta + \mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{x}\right)\right)\right]\right) \right\| \\ =& \left\| \sqrt{\frac{2\sigma^2+1}{m}} \operatorname{diag}\left(\zeta_{\sigma^2}^{1/2}(\boldsymbol{y})\right) \left(|\boldsymbol{A}\boldsymbol{x}| \odot \left[\exp\left(\mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{x}\right)\right) - \exp\left(\mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{x} + \frac{\beta}{\alpha}\boldsymbol{A}\boldsymbol{w}\right)\right)\right]\right) \right\| \\ \leq& \frac{1}{1-\rho}\left|\frac{\beta}{\alpha}\right| \left\| \sqrt{\frac{2\sigma^2+1}{m}} \operatorname{diag}\left(\zeta_{\sigma^2}^{1/2}(\boldsymbol{y})\right) \Im\left(\boldsymbol{A}\boldsymbol{w} \odot \exp\left(-\mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{x}\right)\right)\right) \right\| \\ &+ 2\sqrt{\frac{2\sigma^2+1}{m}} \left\| |\boldsymbol{A}\boldsymbol{x}| \odot \mathbf{1}_{|\frac{\beta}{\alpha}||\boldsymbol{A}\boldsymbol{w}| \geq \rho |\boldsymbol{A}\boldsymbol{x}|} \right\|,\end{aligned}$$

for any $\rho \in (0,1)$. By Lemma 24.4, we know that $\rho^{-1}\left|\frac{\beta}{\alpha}\right| \leq \frac{2}{\rho}\left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\| < c_\rho \delta^3 \log^{-6} n$ holds under our assumption, where c_ρ is a constant depending on ρ. Thus, whenever $m \geq C_2 \max\left\{\left\|\boldsymbol{C}_{\boldsymbol{x}}\right\|^2 \log^{17} n, \delta^{-2} n \log^4 n\right\}$ for any $\delta \in (0,1)$, with probability at least $1 - c_1 m^{-c_2}$ for all $\boldsymbol{w} \in \mathbb{CS}^{n-1}$, Lemma 24.5 implies that

$$\left\| |\boldsymbol{A}\boldsymbol{x}| \odot \mathbf{1}_{|\frac{\beta}{\alpha}||\boldsymbol{A}\boldsymbol{w}| \geq \rho |\boldsymbol{A}\boldsymbol{x}|} \right\| \leq \frac{\delta}{\rho}\left|\frac{\beta}{\alpha}\right| \sqrt{m} \leq \frac{2\delta}{\rho}\sqrt{m}\left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\|.$$

Here, c_1, c_2 and C_2 are some positive numerical constants.

Moreover, for any $\delta \in (0,1)$, whenever $m \geq C_3 \left\|\boldsymbol{C}_{\boldsymbol{x}}\right\|^2 n \log^4 n$, with probability at least $1 - c_3 m^{-c_4 \log^3 n}$ for all $\boldsymbol{w} \in \mathbb{CS}^{n-1}$ with $\boldsymbol{w} \perp \boldsymbol{x}$, Lemma 24.6 implies that

$$\left\| \sqrt{\frac{2\sigma^2+1}{m}} \operatorname{diag}\left(\zeta_{\sigma^2}^{1/2}(\boldsymbol{y})\right) \Im\left(\boldsymbol{A}\boldsymbol{w} \odot \exp\left(-\mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{x}\right)\right)\right) \right\| \leq \sqrt{\frac{1 + 2\delta + (1+\delta)\Delta_\infty(\varepsilon)}{2}},$$

where $\Delta_\infty(\varepsilon)$ is defined in (24.5.1) for some $\varepsilon \in (0,1)$, and c_3, c_4 are some positive numerical constants. In addition, whenever $\left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\| \leq c_5 \delta^3 \log^{-6} n \left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\|$ for some constant $c_5 > 0$, Lemma 24.4 implies that

$$\left|\frac{\beta}{\alpha}\right| \leq \frac{1}{1 - c_5\delta^3 \log^{-6} n}\left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\| \leq \frac{1}{1-\delta}\left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\|,$$

for $\delta > 0$ sufficiently small. Thus, combining the results above, we have the bound

$$\|\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{d}\| \le \left[\delta + (1+\delta)\left(\frac{4\delta}{\rho}\sqrt{2\sigma^2+1} + \frac{1}{1-\rho}\frac{1}{1-\delta}\sqrt{\frac{1+2\delta+(1+\delta)\Delta_\infty(\varepsilon)}{2}}\right)\right]\left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\|$$

holds as desired. Finally, when $\sigma^2 = 0.51$ and $\varepsilon = 0.2$, the bound for $\Delta_\infty(\varepsilon)$ can be found in Lemma 24.15 in Section 24.5. ∎

24.3.0.2 Bounding the "$\boldsymbol{x}$-parallel" term $\|\boldsymbol{P}_{\boldsymbol{x}}\boldsymbol{d}\|$

Lemma 24.8 *Let $\boldsymbol{d}(\boldsymbol{z})$ be defined in (24.2.7), and let $\sigma^2 > 1/2$ be a constant. For any $\delta > 0$, whenever $m \ge C\|\boldsymbol{C}_{\boldsymbol{x}}\|^2 \max\left\{\log^{17} n, \delta^{-2} n \log^4 n\right\}$, with probability at least $1 - c_1 m^{-c_2}$ for all $\boldsymbol{z}$ such that $\left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\| \le c_3\delta^3 \log^{-6} n$, we have*

$$\|\boldsymbol{P}_{\boldsymbol{x}}\boldsymbol{d}\| \le \left(\frac{2\sigma^2}{1+2\sigma^2} + c_{\sigma^2}\delta\right)\left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\|.$$

Here, c_1, c_2, c_3 and C are some positive numerical constants, and $c_{\sigma^2} > 0$ is some numerical constant depending only on σ^2.

Proof Given the decomposition of $\boldsymbol{z}$ in (24.2.5) with $\boldsymbol{w} \perp \boldsymbol{x}$ and $\phi(\alpha) = \theta$, and by the definition of $\boldsymbol{d}(\boldsymbol{z})$ in (24.2.7), we observe

$$\begin{aligned}
\|\boldsymbol{P}_{\boldsymbol{x}}\boldsymbol{d}\| &= \left|\boldsymbol{x}^*\left\{(\boldsymbol{I}-\boldsymbol{M})\left(\boldsymbol{z} - e^{\mathrm{i}\theta}\boldsymbol{x}\right) - e^{\mathrm{i}\theta}\boldsymbol{M}\boldsymbol{x} + \frac{2\sigma^2+1}{m}\boldsymbol{A}^*\operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{y})\right)\left[\boldsymbol{y}\odot\exp\left(\mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{z}\right)\right)\right]\right\}\right| \\
&\le \left|\left(1 - \boldsymbol{x}^*\mathbb{E}\left[\boldsymbol{M}\right]\boldsymbol{x}\right)\left(|\alpha|-1\right)e^{\mathrm{i}\theta} - e^{\mathrm{i}\theta}\boldsymbol{x}^*\boldsymbol{M}\boldsymbol{x} + \frac{2\sigma^2+1}{m}\boldsymbol{x}^*\boldsymbol{A}^*\operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{y})\right)\left[\boldsymbol{y}\odot\exp\left(\mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{z}\right)\right)\right]\right| \\
&\quad + \|\boldsymbol{M} - \mathbb{E}\left[\boldsymbol{M}\right]\|\left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\| \\
&\le \underbrace{\left|\left(1 - \boldsymbol{x}^*\mathbb{E}\left[\boldsymbol{M}\right]\boldsymbol{x}\right)\right|\left||\alpha|-1\right|}_{\mathcal{T}_1} + \|\boldsymbol{M} - \mathbb{E}\left[\boldsymbol{M}\right]\|\left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\| \\
&\quad + \underbrace{\left|\frac{2\sigma^2+1}{m}\boldsymbol{x}^*\boldsymbol{A}^*\operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{y})\right)\left[\left(\boldsymbol{A}\boldsymbol{x}\right)\odot\left(\exp\left(\mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{z} - \mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{x}\right)\right) - e^{\mathrm{i}\theta}\mathbf{1}\right)\right)\right]\right|}_{\mathcal{T}_2},
\end{aligned}$$

where for the second inequality, we used Lemma B.30 such that $\boldsymbol{x}^*\left(\boldsymbol{I} - \mathbb{E}\left[\boldsymbol{M}\right]\right)\boldsymbol{w} = 0$. For the first term $\mathcal{T}_1$, notice that

$$\left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\| = \sqrt{\left||\alpha|-1\right|^2 + \|\beta\boldsymbol{w}\|^2} \ge \left||\alpha|-1\right|,$$

and by using the fact that $\mathbb{E}[\boldsymbol{M}] = \boldsymbol{I} + \frac{2\sigma^2}{1+2\sigma^2}\boldsymbol{x}\boldsymbol{x}^*$ in Lemma B.30, we have

$$\mathcal{T}_1 \;=\; \frac{2\sigma^2}{1+2\sigma^2}\left||\alpha|-1\right| \;\leq\; \frac{2\sigma^2}{1+2\sigma^2}\left\|\boldsymbol{z}-\boldsymbol{x}e^{\mathrm{i}\theta}\right\|.$$

For the term $\mathcal{T}_2$, using the fact that $\boldsymbol{z} = \alpha\boldsymbol{x} + \beta\boldsymbol{w}$ and $\theta = \phi(\alpha)$, and by Lemma B.2, notice that

$$\begin{aligned}
&\left|\exp\left(\mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{z}\right) - \mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{x}\right)\right) - e^{\mathrm{i}\theta}\boldsymbol{1} + \mathrm{i}e^{\mathrm{i}\theta}\Im\left(\frac{\beta\boldsymbol{A}\boldsymbol{w}}{\alpha\boldsymbol{A}\boldsymbol{x}}\right)\right| \\
=\;& \left|\exp\left(\mathrm{i}\phi\left(1+\frac{\beta\boldsymbol{A}\boldsymbol{w}}{\alpha\boldsymbol{A}\boldsymbol{x}}\right)\right) - \boldsymbol{1} + \mathrm{i}\Im\left(\frac{\beta\boldsymbol{A}\boldsymbol{w}}{\alpha\boldsymbol{A}\boldsymbol{x}}\right)\right| \leq 6\left|\frac{\beta}{\alpha}\right|^2\left|\frac{\boldsymbol{A}\boldsymbol{w}}{\boldsymbol{A}\boldsymbol{x}}\right|^2,
\end{aligned}$$

whenever $\left|\frac{\beta\boldsymbol{A}\boldsymbol{w}}{\alpha\boldsymbol{A}\boldsymbol{x}}\right| \leq 1/2$. Thus, by using the result above, we observe

$$\begin{aligned}
\mathcal{T}_2 \;\leq\;& 2\left|\frac{2\sigma^2+1}{m}\boldsymbol{x}^*\boldsymbol{A}^*\operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{y})\right)\left[\left(\boldsymbol{A}\boldsymbol{x}\right)\odot\boldsymbol{1}_{|\frac{\beta}{\alpha}||\boldsymbol{A}\boldsymbol{w}|\geq\frac{1}{2}|\boldsymbol{A}\boldsymbol{x}|}\right]\right| \\
&+\left|\frac{2\sigma^2+1}{m}\boldsymbol{x}^*\boldsymbol{A}^*\operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{y})\odot\left(\exp\left(\mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{z}\right)-\mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{x}\right)\right)-e^{\mathrm{i}\theta}\boldsymbol{1}\right)\odot\boldsymbol{1}_{|\frac{\beta}{\alpha}||\boldsymbol{A}\boldsymbol{w}|\leq\frac{1}{2}|\boldsymbol{A}\boldsymbol{x}|}\right)\boldsymbol{A}\boldsymbol{x}\right| \\
\leq\;& 2\left|\frac{2\sigma^2+1}{m}\boldsymbol{x}^*\boldsymbol{A}^*\operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{y})\right)\left[\left(\boldsymbol{A}\boldsymbol{x}\right)\odot\boldsymbol{1}_{|\frac{\beta}{\alpha}||\boldsymbol{A}\boldsymbol{w}|\geq\frac{1}{2}|\boldsymbol{A}\boldsymbol{x}|}\right]\right| \\
&+\left|\frac{2\sigma^2+1}{m}\boldsymbol{x}^*\boldsymbol{A}^*\operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{y})\right)\left[\left(\boldsymbol{A}\boldsymbol{x}\right)\odot\Im\left(\frac{\beta\boldsymbol{A}\boldsymbol{w}}{\alpha\boldsymbol{A}\boldsymbol{x}}\right)\right]\right| \\
&+6\left|\frac{\beta}{\alpha}\right|^2\left|\frac{2\sigma^2+1}{m}\boldsymbol{x}^*\boldsymbol{A}^*\operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{y})\odot\frac{|\boldsymbol{A}\boldsymbol{w}|^2}{|\boldsymbol{A}\boldsymbol{x}|^2}\right)\boldsymbol{A}\boldsymbol{x}\right| \\
\leq\;& 2\frac{2\sigma^2+1}{m}\left\|\boldsymbol{A}\right\|\left\|\boldsymbol{A}\boldsymbol{x}\odot\boldsymbol{1}_{|\frac{\beta}{\alpha}||\boldsymbol{A}\boldsymbol{w}|\leq\frac{1}{2}|\boldsymbol{A}\boldsymbol{x}|}\right\| + 6\left|\frac{\beta}{\alpha}\right|^2\left|\frac{2\sigma^2+1}{m}\boldsymbol{w}^*\boldsymbol{A}^*\operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{y})\right)\boldsymbol{A}\boldsymbol{w}\right| \\
&+\frac{1}{2}\left|\frac{\beta}{\alpha}\right|\left|\frac{2\sigma^2+1}{m}\boldsymbol{x}^*\boldsymbol{A}^*\operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{y})\right)\boldsymbol{A}\boldsymbol{w}\right| \\
&+\frac{1}{2}\left|\frac{\beta}{\alpha}\right|\left|\frac{2\sigma^2+1}{m}\boldsymbol{x}^*\boldsymbol{A}^*\operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{y})\right)\left[\exp\left(2\mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{x}\right)\right)\odot\overline{\boldsymbol{A}\boldsymbol{w}}\right]\right|
\end{aligned}$$

Given the fact that $\boldsymbol{x}\perp\boldsymbol{w}$, by Lemma B.30 again we have $\boldsymbol{x}^*\mathbb{E}[\boldsymbol{M}]\boldsymbol{w} = 0$. Thus

$$\left|\frac{2\sigma^2+1}{m}\boldsymbol{x}^*\boldsymbol{A}^*\operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{y})\right)\boldsymbol{A}\boldsymbol{w}\right| = \left|\boldsymbol{x}^*\boldsymbol{M}\boldsymbol{w} - \boldsymbol{x}^*\mathbb{E}[\boldsymbol{M}]\boldsymbol{w}\right| \leq \left\|\boldsymbol{M}-\mathbb{E}[\boldsymbol{M}]\right\|,$$

and similarly we have

$$\begin{aligned}
&\left|\frac{2\sigma^2+1}{m}\boldsymbol{x}^*\boldsymbol{A}^*\operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{y})\right)\left[\exp\left(2\mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{x}\right)\right)\odot\overline{\boldsymbol{A}\boldsymbol{w}}\right]\right| \\
=\;& \left|\frac{2\sigma^2+1}{m}\boldsymbol{w}^\top\boldsymbol{A}^\top\operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{y})\right)\left[\exp\left(-2\mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{x}\right)\right)\odot\boldsymbol{A}\boldsymbol{x}\right]\right| \\
=\;& \left|\frac{2\sigma^2+1}{m}\boldsymbol{w}^\top\boldsymbol{A}^\top\operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{y})\right)\overline{\boldsymbol{A}\boldsymbol{x}}\right| \\
=\;& \left|\frac{2\sigma^2+1}{m}\boldsymbol{x}^*\boldsymbol{A}^*\operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{y})\right)\boldsymbol{A}\boldsymbol{w}\right| \leq \left\|\boldsymbol{M}-\mathbb{E}[\boldsymbol{M}]\right\|.
\end{aligned}$$

Thus, suppose $\left\|\boldsymbol{z}-\boldsymbol{x}e^{\mathrm{i}\theta}\right\| \leq \frac{1}{2}$, by using Lemma 24.4 we know that $\left|\frac{\beta}{\alpha}\right| \leq 2\left\|\boldsymbol{z}-\boldsymbol{x}e^{\mathrm{i}\theta}\right\|$. Thus, combining the estimates above, we have

$$\begin{aligned}\mathcal{T}_3 \leq & 2\frac{2\sigma^2+1}{m}\|\boldsymbol{A}\|\left\|\boldsymbol{A}\boldsymbol{x}\odot\mathbf{1}_{\left|\frac{\beta}{\alpha}\right||\boldsymbol{A}\boldsymbol{w}|\leq\frac{1}{2}|\boldsymbol{A}\boldsymbol{x}|}\right\|+24\left\|\boldsymbol{z}-\boldsymbol{x}e^{\mathrm{i}\theta}\right\|^2\|\boldsymbol{M}\| \\ & +2\|\boldsymbol{M}-\mathbb{E}[\boldsymbol{M}]\|\left\|\boldsymbol{z}-\boldsymbol{x}e^{\mathrm{i}\theta}\right\|.\end{aligned}$$

Combining the estimates for $\mathcal{T}_1$ and $\mathcal{T}_2$, we have

$$\begin{aligned}\|\boldsymbol{P}_{\boldsymbol{x}}\boldsymbol{d}\| \leq & \frac{2\sigma^2}{1+2\sigma^2}\left\|\boldsymbol{z}-\boldsymbol{x}^{\mathrm{i}\theta}\right\|+3\|\boldsymbol{M}-\mathbb{E}[\boldsymbol{M}]\|\left\|\boldsymbol{z}-\boldsymbol{x}^{\mathrm{i}\theta}\right\| \\ & +2\frac{2\sigma^2+1}{m}\|\boldsymbol{A}\|\left\|\boldsymbol{A}\boldsymbol{x}\odot\mathbf{1}_{\left|\frac{\beta}{\alpha}\right||\boldsymbol{A}\boldsymbol{w}|\leq\frac{1}{2}|\boldsymbol{A}\boldsymbol{x}|}\right\|+24\|\boldsymbol{M}\|\left\|\boldsymbol{z}-\boldsymbol{x}e^{\mathrm{i}\theta}\right\|^2.\end{aligned}$$

By Theorem B.24, for any $\delta>0$, whenever $m \geq C_1\delta^{-2}\|\boldsymbol{C}_{\boldsymbol{x}}\|^2 n\log^4 n$, we have

$$\|\boldsymbol{M}-\mathbb{E}[\boldsymbol{M}]\| \leq \delta, \quad \|\boldsymbol{M}\| \leq \|\mathbb{E}[\boldsymbol{M}]\|+\delta = \frac{1+4\sigma^2}{1+2\sigma^2}+\delta$$

holds with probability at least $1-c_1 m^{-c_2\log^3 n}$. Here c_1, c_2, and C_1 are positive numerical constants. By Corollary B.13, for any $\delta\in(0,1)$, whenever $m \geq C_2\delta^{-2}n\log^4 n$, we have

$$\|\boldsymbol{A}\| \leq (1+\delta)\sqrt{m}$$

holds with probability at least $1-2m^{-c_3\log^3 n}$ for some constant $c_3>0$. If $\frac{1}{2}\left|\frac{\beta}{\alpha}\right| \leq \left\|\boldsymbol{z}-\boldsymbol{x}e^{\mathrm{i}\theta}\right\| \leq c_4\delta^3\log^{-6} n$, whenever $m \geq C_3\max\left\{\|\boldsymbol{C}_{\boldsymbol{x}}\|^2\log^{17} n, \delta^{-2}n\log^4 n\right\}$, Lemma 24.5 implies that

$$\left\||\boldsymbol{A}\boldsymbol{x}|\odot\mathbf{1}_{\left|\frac{\beta}{\alpha}\right||\boldsymbol{A}\boldsymbol{w}|\geq\frac{1}{2}|\boldsymbol{A}\boldsymbol{x}|}\right\| \leq 2\delta\left|\frac{\beta}{\alpha}\right|\sqrt{m} \leq 4\delta\sqrt{m}\left\|\boldsymbol{z}-\boldsymbol{x}e^{\mathrm{i}\theta}\right\|$$

holds for all $\boldsymbol{w}\in\mathbb{CS}^{n-1}$ with probability at least $1-c_5 m^{-c_6}$. Here, c_4, c_5, c_6 and C_2, C_3 are some numerical constants. Given $\left\|\boldsymbol{z}-\boldsymbol{x}e^{\mathrm{i}\theta}\right\| \leq \frac{c_4}{4}\delta^3\log^{-6} n$, combining the estimates above, we have

$$\begin{aligned}\|\boldsymbol{P}_{\boldsymbol{x}}\boldsymbol{d}\| & \leq \left[\frac{2\sigma^2}{1+2\sigma^2}+3\delta+8(1+\delta)\delta\left(2\sigma^2+1\right)+24c_4\left(\frac{1+4\sigma^2}{1+2\sigma^2}+\delta\right)\delta^3\log^{-6} n\right]\left\|\boldsymbol{z}-\boldsymbol{x}e^{\mathrm{i}\theta}\right\| \\ & \leq \left(\frac{2\sigma^2}{1+2\sigma^2}+c_{\sigma^2}\delta\right)\left\|\boldsymbol{z}-\boldsymbol{x}e^{\mathrm{i}\theta}\right\|\end{aligned}$$

for δ sufficiently small. Here, c_{σ^2} is some positive numerical constant depending only on σ^2. ■

24.4 Proof of Lemma 24.5

In this section, we assume $\|\boldsymbol{x}\| = 1$, and we prove the Lemma 24.5 in Subsection 24.3, which can be restated as follows.

Lemma 24.9 *For any given scalar $\delta \in (0,1)$, let $\gamma = c_0\delta^3 \log^{-6} n$, whenever $m \geq C \max\left\{\|\boldsymbol{C}_{\boldsymbol{x}}\|^2 \log^{17} n, \delta^{-2} n \log^4 n\right\}$, with probability at least $1 - c_1 m^{-c_2}$ for all $\boldsymbol{w}$ with $\|\boldsymbol{w}\| \leq \gamma \|\boldsymbol{x}\|$, the inequality*

$$\left\|\boldsymbol{A}\boldsymbol{x} \odot \mathbf{1}_{|\boldsymbol{A}\boldsymbol{w}| \geq |\boldsymbol{A}\boldsymbol{x}|}\right\| \leq \delta\sqrt{m}\,\|\boldsymbol{w}\| \tag{24.4.1}$$

holds. Here, c_0, c_1, c_2 and C are some positive numerical constants.

We prove this lemma using the results in Lemma 24.10 and Lemma 24.11.

Proof By Corollary B.13, for some small scalar $\varepsilon \in (0,1)$, whenever $m \geq Cn\log^4 n$, with probability at least $1 - m^{-c\log^3 n}$ for every $\boldsymbol{w}$ with $\|\boldsymbol{w}\| \leq \gamma\|\boldsymbol{x}\|$, we have

$$\|\boldsymbol{A}\boldsymbol{w}\| \leq (1+\varepsilon)\sqrt{m}\,\|\boldsymbol{w}\| \leq (1+\varepsilon)\gamma\sqrt{m}\,\|\boldsymbol{x}\| \leq \left(\frac{1+\varepsilon}{1-\varepsilon}\right)^{1/2} \gamma\,\|\boldsymbol{A}\boldsymbol{x}\| \leq 2\gamma\,\|\boldsymbol{A}\boldsymbol{x}\|\,.$$

Here, c, C are some numerical constants. Let us define a set

$$\mathcal{S} \doteq \{k \mid |\boldsymbol{a}_k^*\boldsymbol{w}| \geq |\boldsymbol{a}_k^*\boldsymbol{x}|\}\,.$$

By Lemma 24.10, with probability at least $1 - \exp\left(-\frac{\rho^4 m}{2\|\boldsymbol{C}_{\boldsymbol{x}}\|^2}\right)$, for every set $\mathcal{S}$ with $|\mathcal{S}| > \rho m$ (for some $\rho \in (0,1)$ to be chosen later), we have

$$\|(\boldsymbol{A}\boldsymbol{x}) \odot \mathbf{1}_{\mathcal{S}}\| > \frac{\rho^{3/2}}{32}\,\|\boldsymbol{A}\boldsymbol{x}\|$$

holds. Choose ρ such that $\gamma = \frac{\rho^{3/2}}{64}$, we have

$$\|\boldsymbol{A}\boldsymbol{w}\| \geq \|(\boldsymbol{A}\boldsymbol{w}) \odot \mathbf{1}_{\mathcal{S}}\| \geq \|(\boldsymbol{A}\boldsymbol{x}) \odot \mathbf{1}_{\mathcal{S}}\| > 2\gamma\,\|\boldsymbol{A}\boldsymbol{x}\|\,.$$

This contradicts with the fact that $\|\boldsymbol{A}\boldsymbol{w}\| \leq 2\gamma\,\|\boldsymbol{A}\boldsymbol{x}\|$. Therefore, whenever $\|\boldsymbol{w}\| \leq \gamma\|\boldsymbol{x}\|$, with high probability we have $|\mathcal{S}| \leq \rho m$ holds. Given any $\delta > 0$, choose $\gamma = c\delta^3 \log^{-6} n$ for some constant $c > 0$. Because $\gamma = \frac{\rho^{3/2}}{64}$, we know that $\rho = c'\delta^2/\log^4 n$. By Lemma 24.11, whenever $m \geq C\delta^{-2} n \log^4 n$, with probability at least $1 - 2m^{-c\log^2 n}$ for all $\boldsymbol{w} \in \mathbb{CS}^{n-1}$, we have

$$\left\||\boldsymbol{A}\boldsymbol{x}| \odot \mathbf{1}_{|\boldsymbol{A}\boldsymbol{w}| \geq |\boldsymbol{A}\boldsymbol{x}|}\right\| \leq \||\boldsymbol{A}\boldsymbol{w}| \odot \mathbf{1}_{\mathcal{S}}\| \leq \delta\sqrt{m}\,\|\boldsymbol{w}\|\,,$$

holds . Here $c, c',\ C$ are some numerical constants. Combining the results above, we complete the proof. ■

Lemma 24.10 *Let $\rho \in (0,1)$ be a positive scalar, with probability at least $1 - \exp\left(-\frac{\rho^4 m}{2\|\boldsymbol{C}_{\boldsymbol{x}}\|^2}\right)$, for every set $\mathcal{S} \in [m]$ with $|\mathcal{S}| \geq \rho m$, we have*

$$\||\boldsymbol{A}\boldsymbol{x}| \odot \mathbf{1}_{\mathcal{S}}\| > \frac{1}{32}\rho^{3/2} \|\boldsymbol{A}\boldsymbol{x}\| .$$

Proof Let $g_\rho(\boldsymbol{A}\boldsymbol{x})$ be defined as in Lemma 24.12, we know that

$$\|g_\rho(\boldsymbol{A}\boldsymbol{x})\|_1 \geq \left\|\mathbf{1}_{|\boldsymbol{A}\boldsymbol{x}| \leq \rho}\right\|_1$$

holds uniformly. Thus, for an independent copy $\boldsymbol{a}'$ of $\boldsymbol{a}$, we have

$$\begin{aligned}\left|\|g_\rho(\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{a})\|_1 - \|g_\rho(\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{a}')\|_1\right| \leq \|g_\rho(\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{a}) - g_\rho(\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{a}')\|_1 &\leq \frac{\sqrt{m}}{\rho} \|\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{a} - \boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{a}'\| \\ &\leq \frac{\sqrt{m}}{\rho} \|\boldsymbol{C}_{\boldsymbol{x}}\| \|\boldsymbol{a} - \boldsymbol{a}'\| .\end{aligned}$$

Therefore, we can see that $\|g_\rho(\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{a})\|_1$ is L-Lipschitz with respect to $\boldsymbol{a}$, with $L = \frac{\sqrt{m}}{\rho}\|\boldsymbol{C}_{\boldsymbol{x}}\|$. By Gaussian concentration inequality in Lemma B.3, we have

$$\mathbb{P}\left(\left|\|g_\rho(\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{a})\|_1 - \mathbb{E}\left[\|g_\rho(\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{a})\|_1\right]\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2L^2}\right). \tag{24.4.2}$$

By using the fact that $\sqrt{2}\,|\boldsymbol{a}_k^*\boldsymbol{x}|$ follows χ distribution, we have

$$\mathbb{E}\left[\|g_\rho(\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{a})\|_1\right] \leq \sum_{k=1}^m \mathbb{E}\left[\mathbb{1}_{|\boldsymbol{a}_k^*\boldsymbol{x}| \leq 2\rho}\right] = \sum_{k=1}^m \mathbb{P}\left(|\boldsymbol{a}_k^*\boldsymbol{x}| \leq 2\rho\right) \leq \rho m.$$

Thus, with probability at least $1 - 2\exp\left(-\frac{\rho^4 m}{2\|\boldsymbol{C}_{\boldsymbol{x}}\|^2}\right)$, we have

$$\left\|\mathbf{1}_{|\boldsymbol{A}\boldsymbol{x}| \leq \rho}\right\|_1 \leq \|g_\rho(\boldsymbol{A}\boldsymbol{x})\|_1 \leq 2\rho m$$

holds. Thus, for any set $\mathcal{S}$ such that $|\mathcal{S}| \geq 4\rho m$, we have

$$\|(\boldsymbol{A}\boldsymbol{x}) \odot \mathbf{1}_{\mathcal{S}}\|^2 \geq \left\|(\boldsymbol{A}\boldsymbol{x}) \odot \mathbf{1}_{|\boldsymbol{A}\boldsymbol{x}| \leq \rho}\right\|^2 \geq 2\rho^3 m.$$

Thus, by replacing 4ρ with ρ, we complete the proof. ■

Lemma 24.11 *Given any scalar $\delta > 0$, let $\rho \in (0, c_\delta log^{-4} n)$ with c_δ be some constant depending on δ, whenever $m \geq C\delta^{-2} n \log^4 n$, with probability at least $1 - 2m^{-c\log^2 n}$, for any set $\mathcal{S} \in [m]$ with $|\mathcal{S}| < \rho m$ and for all*

$\boldsymbol{w} \in \mathbb{C}^n$, we have

$$\|(\boldsymbol{A}\boldsymbol{w}) \odot \mathbf{1}_{\mathcal{S}}\| \leq \delta\sqrt{m}\,\|\boldsymbol{w}\|$$

holds. Here $c, C > 0$ are some numerical constants.

Proof Without loss of generality, let us assume that $\|\boldsymbol{w}\| = 1$. First, notice that

$$\|\boldsymbol{A}\boldsymbol{w} \odot \mathbf{1}_{\mathcal{S}}\| \;=\; \sup_{\boldsymbol{v}\in\mathbb{CS}^{m-1},\ \mathrm{supp}(\boldsymbol{v})\subseteq\mathcal{S}} \langle \boldsymbol{v}, \boldsymbol{A}\boldsymbol{w}\rangle \;\leq\; \sup_{\boldsymbol{v}\in\mathbb{CS}^{m-1},\ \mathrm{supp}(\boldsymbol{v})\subseteq\mathcal{S}} \|\boldsymbol{A}^*\boldsymbol{v}\|\,.$$

By Lemma B.11, for any positive scalar $\delta > 0$ and any $\rho \in (0, c\delta^2 \log^{-4} n)$, whenever $m \geq C\delta^{-2} n \log^4 n$, with probability at least $1 - m^{-c' \log^2 n}$, we have

$$\sup_{\boldsymbol{v}\in\mathbb{CS}^{n-1},\ \mathrm{supp}(\boldsymbol{v})\subseteq\mathcal{S}} \|\boldsymbol{A}^*\boldsymbol{v}\| \leq \delta\sqrt{m}$$

holds. Here $c, c', C > 0$ are some positive numerical constants. ∎

Lemma 24.12 *For a variable $u \in \mathbb{C}$ and a fixed positive scalar $v \in \mathbb{R}$, let us define*

$$g_v(u) = \begin{cases} 1 & \text{if } |u| \leq v, \\ \frac{1}{v}\,(2v - |u|) & v < |u| \leq 2v \\ 0 & \text{otherwise}, \end{cases} \tag{24.4.3}$$

then $g_v(u)$ is $1/v$-Lipschitz. Moreover, the following bound

$$g_v(u) \geq \mathbb{1}_{|u|\leq v}$$

holds uniformly for u over the whole space.

Proof The proof of Lipschitz continuity of $g_v(u)$ is straight forward, and the inequality directly follows from the definition of $g_v(u)$. ∎

24.5 Proof of Lemma 24.6

Given some scalar $\varepsilon > 0$ and $\sigma^2 > 1/2$, let us define a quantity

$$\Delta_\infty(\varepsilon) \doteq \left(1 + 2\sigma^2\right) \left\|(1+\varepsilon)\mathbb{E}_{s\sim\mathcal{CN}(0,1)}\left[\psi(t+s)\right] - \zeta_{\sigma^2}(t)\psi(t)\right\|_{L^\infty}, \tag{24.5.1}$$

where $\psi(t) = \exp\left(-2\mathrm{i}\phi(t)\right)$ and ζ_{σ^2} is defined in (24.2.1). Assuming $\|\boldsymbol{x}\| = 1$, we show the following result.

Lemma 24.13 *For any scalar $\delta \in (0,1)$, whenever $m \geq C \|\boldsymbol{C}_{\boldsymbol{x}}\|^2 \delta^{-2} n \log^4 n$, with probability at least $1 - cm^{-c' \log^3 n}$ for all $\boldsymbol{w} \in \mathbb{C}^n$ with $\boldsymbol{w} \perp \boldsymbol{x}$, we have*

$$\left\| \sqrt{\frac{2\sigma^2+1}{m}} \operatorname{diag}\left(\zeta_{\sigma^2}^{1/2}(\boldsymbol{y})\right) \Im\left(\boldsymbol{A}\boldsymbol{w} \odot \exp\left(-\mathrm{i}\phi(\boldsymbol{A}\boldsymbol{x})\right)\right) \right\|^2 \leq \frac{1+(2+\varepsilon)\delta+(1+\delta)\Delta_\infty(\varepsilon)}{2} \|\boldsymbol{w}\|^2$$

holds. Here c, c' are some numerical constants. In particular, when $\sigma^2 = 0.51$ and $\varepsilon = 0.2$, we have $\Delta(\varepsilon) \leq 0.404$. With the same probability for all $\boldsymbol{w} \in \mathbb{C}^n$ with $\boldsymbol{w} \perp \boldsymbol{x}$, we have

$$\left\| \sqrt{\frac{2\sigma^2+1}{m}} \operatorname{diag}\left(\zeta_{\sigma^2}^{1/2}(\boldsymbol{y})\right) \Im\left(\boldsymbol{A}\boldsymbol{w} \odot \exp\left(-\mathrm{i}\phi(\boldsymbol{A}\boldsymbol{x})\right)\right) \right\|^2 \leq \frac{1+2.2\delta+0.404(1+\delta)}{2} \|\boldsymbol{w}\|^2 .$$

Proof Without loss of generality, let us assume $\boldsymbol{w} \in \mathbb{CS}^{n-1}$. For any $\boldsymbol{w} \in \mathbb{CS}^{n-1}$ with $\boldsymbol{w} \perp \boldsymbol{x}$, we observe

$$\begin{aligned}
&\left\| \sqrt{\frac{2\sigma^2+1}{m}} \operatorname{diag}\left(\zeta_{\sigma^2}^{1/2}(\boldsymbol{y})\right) \Im\left(\boldsymbol{A}\boldsymbol{w} \odot \exp\left(-\mathrm{i}\phi(\boldsymbol{A}\boldsymbol{x})\right)\right) \right\|^2 \\
&= \left\| \frac{1}{2} \sqrt{\frac{2\sigma^2+1}{m}} \operatorname{diag}\left(\zeta_{\sigma^2}^{1/2}(\boldsymbol{y})\right) \left[(\boldsymbol{A}\boldsymbol{w}) \odot \exp\left(-\mathrm{i}\phi(\boldsymbol{A}\boldsymbol{x})\right) - (\overline{\boldsymbol{A}\boldsymbol{w}}) \odot \exp\left(\mathrm{i}\phi(\boldsymbol{A}\boldsymbol{x})\right)\right] \right\|^2 \\
&\leq \frac{1}{2} \left| \boldsymbol{w}^* \boldsymbol{P}_{\boldsymbol{x}^\perp} \boldsymbol{M} \boldsymbol{P}_{\boldsymbol{x}^\perp} \boldsymbol{w} \right| + \frac{1}{2} \left| \frac{2\sigma^2+1}{m} \boldsymbol{w}^\top \boldsymbol{P}_{\boldsymbol{x}^\perp}^\top \boldsymbol{A}^\top \operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{A}\boldsymbol{x}) \psi(\boldsymbol{A}\boldsymbol{x})\right) \boldsymbol{A} \boldsymbol{P}_{\boldsymbol{x}^\perp} \boldsymbol{w} \right| \\
&\leq \frac{1}{2} \|\mathbb{E}[\boldsymbol{H}]\| + \frac{1}{2} \|\boldsymbol{H} - \mathbb{E}[\boldsymbol{H}]\| + \frac{1}{2} \left| \frac{2\sigma^2+1}{m} \boldsymbol{w}^\top \boldsymbol{P}_{\boldsymbol{x}^\perp}^\top \boldsymbol{A}^\top \operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{A}\boldsymbol{x}) \psi(\boldsymbol{A}\boldsymbol{x})\right) \boldsymbol{A} \boldsymbol{P}_{\boldsymbol{x}^\perp} \boldsymbol{w} \right|,
\end{aligned}$$

holds for all $\boldsymbol{w} \in \mathbb{CS}^{n-1}$ with $\boldsymbol{w} \perp \boldsymbol{x}$, where $\boldsymbol{M}$ and $\boldsymbol{H}$ are defined in (B.3.2) and (B.3.6), and $\psi(t) = \left(\bar{t}/|t|\right)^2$. By Lemma B.30, we know that

$$\|\mathbb{E}[\boldsymbol{H}]\| = \|\boldsymbol{P}_{\boldsymbol{x}^\perp}\| \leq 1. \tag{24.5.2}$$

By Theorem B.24, we know that for any $\delta > 0$, whenever $m \geq C_1 \delta^{-2} \|\boldsymbol{C}_{\boldsymbol{x}}\|^2 n \log^4 n$, we have

$$\|\boldsymbol{H} - \mathbb{E}[\boldsymbol{H}]\| \leq \delta,$$

with probability at least $1 - c_1 m^{-c_2 \log^3 n}$. Here c_1, c_2 and C_1 are some numerical constants. In addition, Lemma 24.14 implies that for any $\delta > 0$, when $m \geq C_2 \delta^{-2} n \log^4 n$ for some constant $C_2 > 0$, we have

$$\left| \frac{2\sigma^2+1}{m} \boldsymbol{w}^\top \boldsymbol{P}_{\boldsymbol{x}^\perp}^\top \boldsymbol{A}^\top \operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{A}\boldsymbol{x}) \psi(\boldsymbol{A}\boldsymbol{x})\right) \boldsymbol{A} \boldsymbol{P}_{\boldsymbol{x}^\perp} \boldsymbol{w} \right| \leq (1+\delta)\Delta_\infty(\varepsilon) + (1+\varepsilon)\delta,$$

holds with probability at least $1 - 2m^{-c_3 \log^3 n}$ for some constant $c_3 > 0$. Combining the results above, we

obtain

$$\left\|\sqrt{\frac{2\sigma^2+1}{m}}\operatorname{diag}\left(\zeta_{\sigma^2}^{1/2}(\boldsymbol{y})\right)\Im\left(\boldsymbol{A}\boldsymbol{w}\odot\exp\left(-\mathrm{i}\phi(\boldsymbol{A}\boldsymbol{x})\right)\right)\right\|^2 \le \frac{1+(2+\varepsilon)\delta+(1+\delta)\Delta_\infty(\varepsilon)}{2}.$$

Finally, by using Lemma 24.15, when $\sigma^2=0.51$ and $\varepsilon=0.2$, we have

$$\left\|\sqrt{\frac{2\sigma^2+1}{m}}\operatorname{diag}\left(\zeta_{\sigma^2}^{1/2}(\boldsymbol{y})\right)\Im\left(\boldsymbol{A}\boldsymbol{w}\odot\exp\left(-\mathrm{i}\phi(\boldsymbol{A}\boldsymbol{x})\right)\right)\right\|^2 \le \frac{1+2.2\delta+0.404(1+\delta)}{2},$$

as desired. ■

Lemma 24.14 *For a fixed scalar $\varepsilon>0$, let $\Delta_\infty(\varepsilon)$ be defined as (24.5.1). For any $\delta>0$, whenever $m\ge C\delta^{-2}n\log^4 n$, with probability at least $1-m^{-c\log^3 n}$ for all $\boldsymbol{w}\in\mathbb{CS}^{n-1}$ with $\boldsymbol{w}\perp\boldsymbol{x}$, we have*

$$\left|\frac{2\sigma^2+1}{m}\boldsymbol{w}^\top\boldsymbol{A}^\top\operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{A}\boldsymbol{x})\psi(\boldsymbol{A}\boldsymbol{x})\right)\boldsymbol{A}\boldsymbol{w}\right|\le(1+\delta)\,\Delta_\infty(\varepsilon)+(1+\varepsilon)\delta$$

holds. Here $c,\ C>0$ are some numerical constants.

Proof First, let $\boldsymbol{g}=\boldsymbol{a}$ and let $\boldsymbol{\delta}\sim\mathcal{CN}(\boldsymbol{0},\boldsymbol{I})$ independent of $\boldsymbol{g}$, given a small scalar $\varepsilon>0$, we have

$$\begin{aligned}
&\left|\frac{2\sigma^2+1}{m}\boldsymbol{w}^\top\boldsymbol{R}_{[1:n]}\boldsymbol{C}_{\boldsymbol{g}}^\top\operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{g}\circledast\boldsymbol{x})\psi(\boldsymbol{g}\circledast\boldsymbol{x})\right)\boldsymbol{C}_{\boldsymbol{g}}\boldsymbol{R}_{[1:n]}^\top\boldsymbol{w}\right|\\
\le&\left|\frac{2\sigma^2+1}{m}\boldsymbol{w}^\top\boldsymbol{R}_{[1:n]}\boldsymbol{C}_{\boldsymbol{g}}^\top\operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{g}\circledast\boldsymbol{x})\psi(\boldsymbol{g}\circledast\boldsymbol{x})-(1+\varepsilon)\mathbb{E}_{\boldsymbol{\delta}}\left[\psi((\boldsymbol{g}-\boldsymbol{\delta})\circledast\boldsymbol{x})\right]\right)\boldsymbol{C}_{\boldsymbol{g}}\boldsymbol{R}_{[1:n]}^\top\boldsymbol{w}\right|\\
&+(1+\varepsilon)\left|\frac{2\sigma^2+1}{m}\boldsymbol{w}^\top\boldsymbol{R}_{[1:n]}\boldsymbol{C}_{\boldsymbol{g}}^\top\operatorname{diag}\left(\mathbb{E}_{\boldsymbol{\delta}}\left[\psi((\boldsymbol{g}-\boldsymbol{\delta})\circledast\boldsymbol{x})\right]\right)\boldsymbol{C}_{\boldsymbol{g}}\boldsymbol{R}_{[1:n]}^\top\boldsymbol{w}\right|\\
\le&\frac{\Delta_\infty(\varepsilon)}{m}\left\|\boldsymbol{R}_{[1:n]}\boldsymbol{C}_{\boldsymbol{g}}^*\right\|^2+(1+\varepsilon)\underbrace{\left|\frac{2\sigma^2+1}{m}\boldsymbol{w}^\top\boldsymbol{R}_{[1:n]}\boldsymbol{C}_{\boldsymbol{g}}^\top\operatorname{diag}\left(\mathbb{E}_{\boldsymbol{\delta}}\left[\psi((\boldsymbol{g}-\boldsymbol{\delta})\circledast\boldsymbol{x})\right]\right)\boldsymbol{C}_{\boldsymbol{g}}\boldsymbol{R}_{[1:n]}^\top\boldsymbol{w}\right|}_{\mathcal{D}(\boldsymbol{g},\boldsymbol{w})}.
\end{aligned}$$

By Corollary B.13, for any $\delta>0$, whenever $m\ge C_0\delta^{-2}n\log^4 n$ for some constant $C_0>0$, we have

$$\left\|\boldsymbol{R}_{[1:n]}\boldsymbol{C}_{\boldsymbol{g}}^*\right\|^2\le(1+\delta)m$$

with probability at least $1-m^{-c_0\log^3 m}$ for some constant $c_0>0$. Next, let us define a decoupled version of $\mathcal{D}(\boldsymbol{g},\boldsymbol{w})$,

$$\mathcal{Q}_{dec}^{\mathcal{D}}(\boldsymbol{g}^1,\boldsymbol{g}^2,\boldsymbol{w})=\frac{2\sigma^2+1}{m}\boldsymbol{w}^\top\boldsymbol{R}_{[1:n]}\boldsymbol{C}_{\boldsymbol{g}^1}^\top\operatorname{diag}\left(\psi(\boldsymbol{g}^2\circledast\boldsymbol{x})\right)\boldsymbol{C}_{\boldsymbol{g}^1}\boldsymbol{R}_{[1:n]}^\top\boldsymbol{w}. \tag{24.5.3}$$

where $\boldsymbol{g}^1=\boldsymbol{g}+\boldsymbol{\delta}$ and $\boldsymbol{g}^2=\boldsymbol{g}-\boldsymbol{\delta}$. Then by using the fact that $\boldsymbol{w}\perp\boldsymbol{x}$, we have

$$\left|\mathbb{E}_{\boldsymbol{\delta}}\left[\mathcal{Q}_{dec}^{\mathcal{D}}(\boldsymbol{g}^1,\boldsymbol{g}^2,\boldsymbol{w})\right]\right|=\left|\frac{2\sigma^2+1}{m}\boldsymbol{w}^\top\boldsymbol{R}_{[1:n]}\boldsymbol{C}_{\boldsymbol{g}}^\top\operatorname{diag}\left(\mathbb{E}_{\boldsymbol{\delta}}\left[\psi\left((\boldsymbol{g}-\boldsymbol{\delta})\circledast\boldsymbol{x}\right)\right]\right)\boldsymbol{C}_{\boldsymbol{g}}\boldsymbol{R}_{[1:n]}^\top\boldsymbol{w}\right|.$$

Then for any positive integer $p \geq 1$, by Jensen's inequality and Theorem B.14, we have

$$\begin{aligned}\left\| \sup_{\|\boldsymbol{w}\|=1,\, \boldsymbol{w}\perp\boldsymbol{x}} |\mathcal{D}(\boldsymbol{g},\boldsymbol{w})| \right\|_{L^p} &\leq \left\| \sup_{\|\boldsymbol{w}\|=1} \left|\mathcal{Q}_{dec}^{\mathcal{D}}(\boldsymbol{g}^1,\boldsymbol{g}^2,\boldsymbol{w})\right| \right\|_{L^p} \\ &\leq C_{\sigma^2}\left(\sqrt{\frac{n}{m}}\log^{3/2} n \log^{1/2} m + \sqrt{\frac{n}{m}}\sqrt{p} + \frac{n}{m}p\right),\end{aligned}$$

where $C_{\sigma^2} > 0$ is some positive constant depending on σ^2, and we used the fact that $\left\|\psi(\boldsymbol{g}^2 \circledast \boldsymbol{x})\right\|_\infty \leq 1$ holds uniformly for all $\boldsymbol{g}^2$. Thus, by Lemma B.6, then for any $\delta > 0$, whenever $m \geq C_1\delta^{-2} n \log^3 n \log m$, we have

$$\sup_{\|\boldsymbol{w}\|=1,\, \boldsymbol{w}\perp\boldsymbol{x}} |\mathcal{D}(\boldsymbol{g},\boldsymbol{w})| \leq \delta,$$

holds with probability at least $1 - m^{-c_1 \log^3 m}$. Here $c_1, C_1 > 0$ are some numerical constants. Combining the results above, we complete the proof. ■

24.5.0.1 Bounding $\Delta_\infty(\varepsilon)$

Let us define

$$h(t) = \mathbb{E}_{s\sim\mathcal{N}(0,1)}\left[\psi(t+s)\right], \tag{24.5.4}$$

in this subsection, given $\sigma^2 > 1/2$, we bound the following quantity

$$\Delta_\infty(\varepsilon) = \left(1+2\sigma^2\right)\left\|(1+\varepsilon)h(t) - \zeta_{\sigma^2}(t)\psi(t)\right\|_{L^\infty}.$$

with $\zeta_{\sigma^2}(t) = 1 - \exp\left(-|t|^2/(2\sigma^2)\right)$ and $\psi(t) = \left(\bar{t}/|t|\right)^2$. The result is as follows.

Lemma 24.15 *Given $\sigma^2 = 0.51$ and $\varepsilon = 0.2$, we have*

$$\Delta_\infty(\varepsilon) \leq 0.404.$$

Proof First, by Lemma 24.16, notice that the function $h(t)$ can be decomposed as

$$h(t) = g(t)\psi(t)$$

where $g(t) : \mathbb{C} \mapsto [0,1)$ is rotational invariant with respect to t. Since $\zeta_{\sigma^2}(t)$ is also rotational invariant with

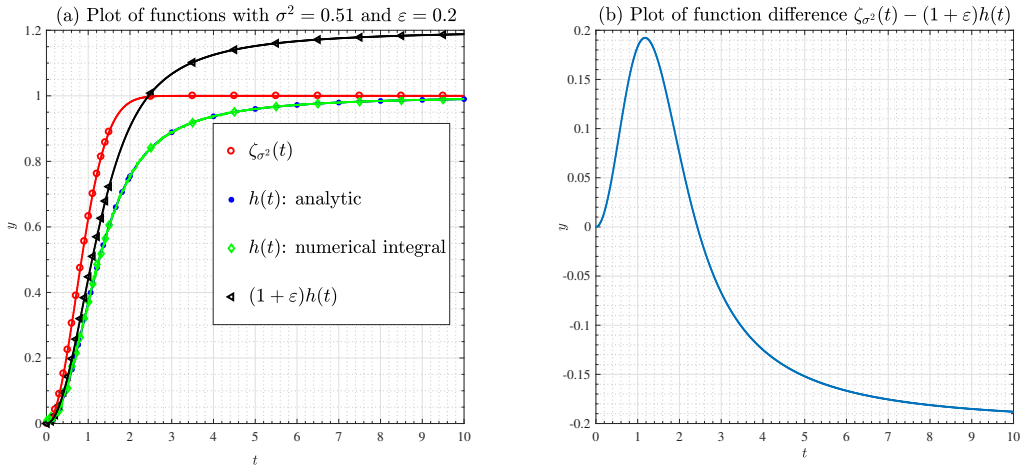


Figure 24.1: Computer simulation of the functions $\zeta_{\sigma^2}(t)$ and $h(t)$. Fig. (a) displays the functions $\zeta_{\sigma^2}(t)$ and $h(t)$ with $\sigma^2 = 0.51$. Fig. (b) shows differences two function $\zeta_{\sigma^2}(t) - (1+\varepsilon)h(t)$ with $\varepsilon = 0.2$.

respect to t, it is enough to consider the case when $t \in [0, +\infty)$, and bounding the following quantity

$$\sup_{t\in[0,+\infty)} \left|(1+\varepsilon)h(t) - \zeta_{\sigma^2}(t)\right|.$$

Lemma 24.17 implies that

$$h(t) \;=\; \mathbb{E}_{s\sim\mathcal{CN}(0,1)}\left[\psi(t+s)\right] \;=\; \begin{cases} 1 - t^{-2} + t^{-2}e^{-t^2} & t > 0, \\ 0 & t = 0. \end{cases}$$

Then if $t = 0$, then it is obvious that $|(1+\varepsilon)h(t) - \zeta_{\sigma^2}(t)| = 0$. For $t > 0$, when $\varepsilon = 0.2$ and $\sigma^2 = 0.51$, we have

$$\zeta_{\sigma^2}(t) - (1+\varepsilon)h(t) = -0.2 - e^{-\frac{t^2}{1.02}} + 1.2t^{-2} - 1.2t^{-2}e^{-t^2}.$$

Based on the observation of Fig. 24.1, we can prove that $\|\zeta_{\sigma^2}(t) - (1+\varepsilon)h(t)\|_{L^\infty} \leq 0.2$ by a tight approximation of the function $\zeta_{\sigma^2}(t) - (1+\varepsilon)h(t)$. Therefore, we have

$$\Delta_\infty(\varepsilon) = (1+2\sigma^2)\left\|\zeta_{\sigma^2}(t) - (1+\varepsilon)h(t)\right\|_{L^\infty} \leq 0.2 \times (1 + 2 \times 0.51) = 0.404,$$

when $\sigma^2 = 0.51$ and $\varepsilon = 0.2$. ■

Lemma 24.16 *Let* $\psi(t) = \left(\bar{t}/\left|t\right|\right)^2$ *, then we have*

$$h(t) = \mathbb{E}_{s\sim\mathcal{CN}(0,1)}\left[\psi(t+s)\right] = g(t)\psi(t).$$

where $g(t) : \mathbb{C} \mapsto [0,1)$*, such that*

$$g(t) = \mathbb{E}_{v_1,v_2\sim\mathcal{N}(0,1/2)}\left[\frac{(|t|+v_1)^2 - v_2^2}{(|t|+v_1)^2 + v_2^2}\right],$$

where $v_1 \sim \mathcal{N}(0,1/2)$*, and* $v_2 \sim \mathcal{N}(0,1/2)$*.*

Proof By definition, we know that

$$\mathbb{E}_{s\sim\mathcal{CN}(0,1)}\left[\psi(t+s)\right] = \mathbb{E}_s\left[\left(\frac{\overline{t+s}}{|t+s|}\right)^2\right] = \underbrace{\mathbb{E}_s\left[\left(\frac{t}{|t|}\frac{\overline{t+s}}{|t+s|}\right)^2\right]}_{g(t)}\psi(t).$$

Next, we estimate $g(t)$ and show that it is indeed real. We decompose the random variable s as

$$s = \Re\left(\frac{\bar{t}s}{|t|}\right)\frac{t}{|t|} + \mathrm{i}\Im\left(\frac{\bar{t}s}{|t|}\right)\frac{t}{|t|} = v_1\frac{t}{|t|} + \mathrm{i}v_2\frac{t}{|t|},$$

where $v_1 = \Re\left(\frac{\bar{t}s}{|t|}\right)$ and $v_2 = \Im\left(\frac{\bar{t}s}{|t|}\right)$ are the real and imaginary parts of a complex Gaussian variable $\bar{t}s/\left|t\right| \sim \mathcal{CN}(0,1)$. By rotation invariant property, we have $v_1 \sim \mathcal{N}(0,1/2)$ and $v_2 \sim \mathcal{N}(0,1/2)$, and v_1 and v_2 are independent. Thus, we have

$$\begin{aligned} h(t) = \mathbb{E}_s\left[\left(\frac{|t|+v_1-\mathrm{i}v_2}{|t+s|}\right)^2\right] &= \mathbb{E}_s\left[\frac{\left(|t|+v_1\right)^2 - v_2^2}{\left|t+s\right|^2}\right] - 2\mathrm{i}\mathbb{E}_s\left[\frac{\left(|t|+v_1\right)v_2}{\left|t+s\right|^2}\right] \\ &= \mathbb{E}_{v_1,v_2}\left[\frac{\left(|t|+v_1\right)^2 - v_2^2}{\left(|t|+v_1\right)^2 + v_2^2}\right] - 2\mathrm{i}\mathbb{E}_{v_1,v_2}\left[\frac{\left(|t|+v_1\right)v_2}{\left(|t|+v_1\right)^2 + v_2^2}\right]. \end{aligned}$$

We can see that $\frac{(|t|+v_1)v_2}{(|t|+v_1)^2+v_2^2}$ is an odd function of v_2. Therefore, the expectation of $\frac{(|t|+v_1)v_2}{(|t|+v_1)^2+v_2^2}$ with respect to v_2 is zero. Thus, we have

$$g(t) = \mathbb{E}_s\left[\left(\frac{t}{|t|}\frac{\overline{t+s}}{|t+s|}\right)^2\right] = \mathbb{E}_{v_1,v_2\sim\mathcal{N}(0,1/2)}\left[\frac{(|t|+v_1)^2 - v_2^2}{(|t|+v_1)^2 + v_2^2}\right],$$

which is real. ∎

Lemma 24.17 *For $t \in [0, +\infty)$, we have*

$$f(t) \;=\; \mathbb{E}_{s\sim\mathcal{CN}(0,1)}\left[\psi(t+s)\right] \;=\; \begin{cases} 1 - t^{-2} + t^{-2}e^{-t^2} & t > 0, \\ 0 & t = 0. \end{cases} \tag{24.5.5}$$

Proof Let $s_r = \Re(s)$ and $s_i = \Im(s)$, and let $s = r\exp(\mathrm{i}\theta)$ with $r = |s|$ and $\exp(\mathrm{i}\theta) = s/|s|$. We observe

$$\begin{aligned}
\mathbb{E}_{s\sim\mathcal{CN}(0,1)}\left[\psi(t+s)\right] &= \frac{1}{\pi}\int_{s_r=-\infty}^{+\infty}\int_{s_i=-\infty}^{+\infty}\frac{(\overline{s_r+\mathrm{i}s_i})^2}{|s_r+\mathrm{i}s_i|^2}e^{-|s_r+\mathrm{i}s_i-t|^2}ds_r ds_i \\
&= \frac{1}{\pi}\int_{r=0}^{+\infty}\int_{\theta=0}^{2\pi} e^{-\mathrm{i}2\theta}e^{-r^2-t^2}e^{2rt\cos\theta}r\,d\theta dr \\
&= \frac{1}{\pi}e^{-t^2}\int_{r=0}^{+\infty}\int_{\theta=0}^{2\pi}\cos(2\theta)re^{-r^2}e^{2rt\cos\theta}d\theta dr \\
&= \frac{2}{\pi}e^{-t^2}\int_{r=0}^{+\infty}\int_{\theta=0}^{\pi}\cos(2\theta)re^{-r^2}\cosh\left(2rt\cos\theta\right)d\theta dr \\
&= \frac{2}{\pi}e^{-t^2}\int_{r=0}^{+\infty}\int_{\theta=0}^{\pi/2}\cos(2\theta)re^{-r^2}\left[\cosh\left(2rt\cos\theta\right)-\cosh(2rt\sin\theta)\right]d\theta dr
\end{aligned}$$

where the third equality uses the fact that the integral of odd function is zero. By using Taylor expansion of $\cosh(x)$, and by using the *dominated convergence theorem* to exchange the summation and integration, we observe

$$\begin{aligned}
&\mathbb{E}_{s\sim\mathcal{CN}(0,1)}\left[\psi(t+s)\right] \\
&= \frac{2}{\pi}e^{-t^2}\int_{r=0}^{+\infty}\int_{\theta=0}^{\pi}\cos(2\theta)re^{-r^2}\sum_{k=0}^{+\infty}\left[\frac{(2rt\cos\theta)^{2k}}{(2k)!}-\frac{(2rt\sin\theta)^{2k}}{(2k)!}\right]d\theta dr \\
&= \frac{2}{\pi}e^{-t^2}\int_{r=0}^{+\infty}\int_{\theta=0}^{\pi}\cos(2\theta)\sum_{k=0}^{+\infty}\left[\frac{(2t\cos\theta)^{2k}r^{2k+1}e^{-r^2}}{(2k)!}-\frac{(2t\sin\theta)^{2k}r^{2k+1}e^{-r^2}}{(2k)!}\right]d\theta dr \\
&= \frac{2}{\pi}e^{-t^2}\sum_{k=0}^{+\infty}\frac{(2t)^{2k}}{(2k)!}\int_{r=0}^{+\infty}r^{2k+1}e^{-r^2}dr\left[\int_{\theta=0}^{\pi}\cos(2\theta)\cos^{2k}\theta d\theta-\int_{\theta=0}^{\pi}\cos(2\theta)\sin^{2k}\theta d\theta\right].
\end{aligned}$$

We have the integrals

$$\begin{aligned}
\int_{r=0}^{+\infty}r^{2k+1}e^{-r^2}dr &= \frac{\Gamma(k+1)}{2}, \\
\int_{\theta=0}^{\pi}\cos(2\theta)\cos^{2k}\theta d\theta &= \frac{\sqrt{\pi}}{2}\frac{k\Gamma(k+1/2)}{\Gamma(k+2)}, \\
\int_{\theta=0}^{\pi}\cos(2\theta)\sin^{2k}\theta d\theta &= -\frac{\sqrt{\pi}}{2}\frac{k\Gamma(k+1/2)}{\Gamma(k+2)}
\end{aligned}$$

holds for any integer $k \geq 0$, where $\Gamma(k)$ is the Gamma function such that

$$\Gamma(k+1) = k!, \quad \Gamma(k+1/2) = \frac{(2k)!}{4^k k!}\sqrt{\pi}.$$

Thus, for $t > 0$, we have

$$\begin{aligned}
\mathbb{E}_{s\sim\mathcal{CN}(0,1)}\left[\psi(t+s)\right] &= \frac{2}{\pi}e^{-t^2}\sum_{k=0}^{+\infty}\frac{(2t)^{2k}}{(2k)!} \times \frac{\Gamma(k+1)}{2} \times \sqrt{\pi}\frac{k\Gamma(k+1/2)}{\Gamma(k+2)} \\
&= e^{-t^2}\sum_{k=0}^{+\infty}\frac{kt^{2k}}{(k+1)!} = e^{-t^2}\left(\sum_{k=0}^{+\infty}\frac{t^{2k}}{k!} - \sum_{k=0}^{+\infty}\frac{t^{2k}}{(k+1)!}\right) \\
&= e^{-t^2}\left[e^{t^2} - t^{-2}\left(\sum_{k=0}^{+\infty}\frac{t^{2k}}{k!} - 1\right)\right] = 1 - t^{-2} + t^{-2}e^{-t^2}.
\end{aligned}$$

When $t = 0$, by using L'Hopital's rule, we have

$$\eta(0) = \lim_{t\to 0}\mathbb{E}_{s\sim\mathcal{CN}(0,1)}\left[\psi(t+s)\right] = \lim_{t\to 0}\left[1 + \frac{1-e^{t^2}}{t^2 e^{t^2}}\right] = 1 + \lim_{t\to 0}\frac{-1}{1+t} = 0.$$

■

Part VI

Discussion and Future Directions

Chapter 25

Future Directions in Broad Perspective

The thesis has been focused on nonconvex optimization methods. In particular, we focused on two problems: (i) phase retrieval, (ii) sparse subspace learning, where both are of broad interest in signal processing and machine learning. Chapter II and Chapter V demonstrate that for certain structured random models, nonconvex problems we can construct a model specialized initialization that is close to the optimal solution, so that simple and efficient methods provably converge to the global solution. Chapter III and Chapter IV studies the complete dictionary learning and phase retrieval under more general assumptions, for which the problems have global geometric structures, that allows efficient and initialization free global optimization. The theories developed in this thesis laid a solid foundation for studying nonconvex problems of broader interest. In the following, we discuss about potential directions moving forward.

25.1 Broader Applications of Nonconvex Optimization

The practical benefits of heuristic nonconvex approaches are well-known in industry. However, nonconvex recovery in practice is still widely viewed as a "dark art". Shedding light on the global guarantees of nonconvex optimizations will not only have substantial theoretical impacts, but also huge impacts in practice that we can efficiently cope with much broader classes of signal structures in a near optimal way.

Scientific/computational imaging Computational imaging problems abound in the modern world. Medical imaging (CT, MRI, PET, ultrasound), remote sensing, seismography, non-destructive inspection, digital photography, astronomy, all involve at their computational core the solution of *inverse problems*. These prob-

lems are often ill-posed with missing or only partial observations. Many inverse problems such as *Fourier phase retrieval* [GS72, Fie82], their variational formulations are naturally nonconvex. However, most of the nonconvex methods that have been proposed lack global convergence guarantees and require "tricks" in order to work well (e.g., careful initialization and continuation procedures), making it hard to trace their (non)success to the behavior of the optimization algorithm or the (in)adequacy of the objective function. I believe our new theoretical insights into those problems will advance the practice by enabling design of better sensing modalities with reduced measurements, and more efficient and guaranteed reconstruction methods. I would like to work with practitioners from sensing, imaging, and a wide range of application domains to investigate on nonconvex methods for those problems with global theoretical guarantees and without careful user interaction.

Deep neural network The success of deep neutral networks in various disciplines is another demonstration of the power of nonconvex optimization. However, its spectacular success is purely empirical — the nonconvexity and nonlinearity of the networks pose significant challenges for theoretical understanding. The lack of theoretical guarantee limits its application to scientific discovery, and many other mission-critical applications. Towards theoretical understanding of deep networks, I would like to: (i) build up our understanding from shallow networks with generative models – what does the function landscape of simple two/three layer network look like in high dimensions? how and why does depth (not) create spurious local minima? (ii) investigate deep network on solving nonlinear inverse problems, where the target functions and solutions are often mathematically precise – for instance, when and why can(not) deep network solve Fourier phase retrieval problem, which cannot be solved by traditional optimization methodologies? The advances would help us provide theoretical guidance of deep networks in broader applications, and shed light on developing better optimization methods.

25.2 General Methodologies of Nonconvex Modeling and Optimization?

A general framework of nonconvex modeling One of our crucial discoveries is that certain nonconvex objective functions arising in structured signal recovery have special structures which enable efficient algorithms to find the global optimum. In the context of the sparse vector in a subspace problem and phase retrieval under random sensing model, this discovery allowed us to break known barriers for convex methods. This is illustrative of a general phenomena: *when the data are large and random, certain seemingly challenging*

nonconvex problems become easy! Inspired by these observations, we aim to attack nonconvex problems falling in the following form

$$\min_{\boldsymbol{z}} \ \mathcal{L}_n(\boldsymbol{z}), \quad \text{s.t.} \quad \boldsymbol{z} \in \mathcal{M}, \tag{25.2.1}$$

where $\mathcal{M}$ is a smooth manifold, $\mathcal{L}_n(\cdot) : \mathcal{M} \mapsto \mathbb{R}$ is a random nonconvex function depends on the observed data $\{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n\}$, and the function value $\mathcal{L}_n(\boldsymbol{z})$ provides the measure of fitness to the given observation. I would like to develop a *global, geometric, and generic framework* for theoretically justifying the correctness of many other nonconvex learning and inverse/recovery problems in the form of (25.2.1). Correspondingly, we will develop a corresponding suite of *efficient, scalable algorithms* which are customized to the special geometric structure of these problems.

More general properties of nonconvex problems Currently, verifying the ridable saddle properties on specific problems, based on first and second derivatives, is highly technical. This limits our ability to identify the benign structure for new nonconvex problems – there is a pressing need for simple analytic tools. Similar to the study of convex functions, one promising direction is to identify conditions and operations that preserve the ridable saddle property: our case study on (overcomplete) dictionary learning and tensor decomposition suggests that adding 4-th order random Gaussian polynomials does not create bad local minima over the sphere – I believe this observation could lead to the discovery of a much more general phenomena.

Furthermore, the computational challenges of globally solving many other nonconvex problems (i.e., deep neutral network and Fourier phase retrieval) cannot be dealt with using strict saddle property. Those problems can have much more complicated landscapes due to their rich inherent symmetry. Studying and understanding symmetries in those problems would potentially provide theoretical insight of solving those problems globally.

Chapter 26

Potential Problems of Particular Interest

In this chapter, we will discuss about several problems that could be of immediate interest and possible extensions of the thesis. We will discuss about several problems whose optimization objective could be (locally) ridable saddle function, which is conjectured with strong numerical evidence.

26.1 Convolutional Dictionary Learning

Given the data $\boldsymbol{y} \in \mathbb{R}^m$, the convolutional dictionary learning (CDL) problem is to seek a compact representation of the data in the following form

$$\boldsymbol{y} = \sum_{k=1}^{N} \boldsymbol{a}_{0k} \circledast \boldsymbol{x}_{0k}, \quad \{\boldsymbol{a}_{0k}\}_{k=1}^{N} \quad \text{convolution kernel}, \quad \{\boldsymbol{x}_{0k}\}_{k=1}^{N} \quad \text{sparse spike train},$$

where $\circledast$ denotes the circulant convolution of $\boldsymbol{a}_{0k} \in \mathbb{R}_0^n$ and $\boldsymbol{x}_{0k} \in \mathbb{R}^m$, and both $\{\boldsymbol{a}_{0k}\}_{k=1}^N$ and $\{\boldsymbol{x}_{0k}\}_{k=1}^N$ are unknown. This problem can be thought as a more general problem of blind deconvolution [ZLK+17], and it appears in many applications of signal processing, astronomy, and computational imaging, etc. For example, the spike sorting problem, which is a crucial step to extract information from extracellular recordings in neural science, can naturally formulated as the CDL problem [SFB18]. Motivated by these applications, we assume the spike trains $\{\boldsymbol{x}_{0k}\}_{k=1}^N$ are sparse, and $\{\boldsymbol{a}_{0k}\}_{k=1}^N$ to be short kernels (i.e.,$n_0 \ll m$) and satisfies the following incoherent conditions

- **Shift incoherence for each kernel $\boldsymbol{a}_{0k}$:** The first assumption is that distinct self-shifts of $\boldsymbol{a}_{0k}$ have small inner product for each $1 \le k \le N$. For each kernel $\boldsymbol{a}_{0k}$ $(1 \le k \le N)$, we define the shift coherence of

$\boldsymbol{a}_{0k}$ to be the largest inner product between distinct self-shifts:

$$\mu_s(\boldsymbol{a}_{0k}) \doteq \max_{\ell} \left|\langle \boldsymbol{a}_{0k}, s_\ell\left[\boldsymbol{a}_{0k}\right]\rangle\right|.$$

The quantity $\mu_s(\boldsymbol{a}_{0k}) \in [0,1]$.

- **Incoherence between kernels $\{\boldsymbol{a}_{0k}\}_{k=1}^m$:** Moreover, we also assume that all the shifts of different kernels $\boldsymbol{a}_{0i}$ and $\boldsymbol{a}_{0j}$ has small correlation. Let $\boldsymbol{A}_0 = [\boldsymbol{a}_{01}, \boldsymbol{a}_{02}, \cdots, \boldsymbol{a}_{0N}]$, we define the incoherence

$$\mu_d(\boldsymbol{A}_0) = \max_{1\le i,j\le N, i\neq j} \left\|\boldsymbol{C}^*_{\boldsymbol{a}_{0i}}\boldsymbol{a}_{0j}\right\|_\infty.$$

Thus, our problem of interest can be stated as follows.

Problem 26.1 *Given the convolutional measurement $\boldsymbol{y} = \sum_{k=1}^N \boldsymbol{a}_{0k} \circledast \boldsymbol{x}_{0k} \in \mathbb{R}^m$, with the kernels $\{\boldsymbol{a}_{0k}\}_{k=1}^N \in \mathbb{R}^{n_0}$ and representations $\{\boldsymbol{x}_{0k}\}_{k=1}^N \in \mathbb{R}^m$ sparse, recover $\{\boldsymbol{a}_{0k}\}_{k=1}^N \in \mathbb{R}^{n_0}$ and $\{\boldsymbol{x}_{0k}\}_{k=1}^N \in \mathbb{R}^m$.*

The problem is notoriously difficult to solve, due to its intrinsic symmetries, which can be classified into three categories,

- **scale symmetry:** It is obvious that the solution of CDL can only be optimal up to scale ambiguity: suppose $\{\boldsymbol{a}^\star_{0k}\}_{k=1}^N$ and $\{\boldsymbol{x}^\star_{0k}\}_{k=1}^N$ are optimal solutions, then for any $\{\alpha_k\}_{k=1}^N \neq 0$, $\left\{\frac{1}{\alpha_k}\boldsymbol{a}^\star_{0k}\right\}_{k=1}^N$ and $\{\alpha_k \boldsymbol{x}^\star_{0k}\}_{k=1}^N$ are also equivalent optimal solutions. We assume $\|\boldsymbol{a}_{0k}\| = 1$ to reduce the scale ambiguities to sign ambiguities.

- **shift symmetry:** Let $s_\ell[\cdot]$ denote the cyclic shift of a signal of length ℓ. Obviously, we have $s_\ell[\boldsymbol{a}_{0k}] \circledast s_{-\ell}[\boldsymbol{x}_{0k}] = \boldsymbol{a}_{0k} \circledast \boldsymbol{x}_{0k}$, so we can only hope to find the solutions up to shift ambiguities.

- **permutation symmetry of $\{\boldsymbol{a}_{0k}\}_{k=1}^N$.** Changing the order of $\boldsymbol{a}_{0k}$ does not change the solution.

Therefore, we can only hope to solve this problem up to scaling, shift and permutation symmetries. To count for all shifts of $\boldsymbol{a}_{0k}$, we consider optimization variables $\widetilde{\boldsymbol{A}}$ of longer length $n = 3n_0 - 2$. Let

$$\boldsymbol{a}_k = \begin{bmatrix} \boldsymbol{0}_{n_0-1} \\ \boldsymbol{a}_{0k} \\ \boldsymbol{0}_{n_0-1} \end{bmatrix} \in \mathbb{S}^{n-1}, \qquad \boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_N \end{bmatrix}, \tag{26.1.1}$$

then we are hoping to recover $\boldsymbol{A}$ up to permutation. The problem can be naturally casted as

$$\min_{\widetilde{\boldsymbol{A}},\boldsymbol{X}} \Phi_L^N\left(\widetilde{\boldsymbol{A}}, \boldsymbol{X}\right) = \frac{1}{2}\left\|\boldsymbol{y} - \sum_{k=1}^N \widetilde{\boldsymbol{a}}_k \circledast \boldsymbol{x}_k\right\|^2 + \lambda \sum_{k=1}^N \|\boldsymbol{x}_k\|_1, \qquad \text{s.t.} \qquad \widetilde{\boldsymbol{a}}_k \in \mathbb{S}^{n-1},$$

where we minimize the least squares loss plus a sparsity promoting penalty for $\boldsymbol{X}$, and $\lambda > 0$ is a scalar. We constrain the kernels $\{\widetilde{\boldsymbol{a}}_k\}_{k=1}^m$ over the spheres (the oblique manifold) to reduce the scale ambiguities. The problem is bilinear in $\widetilde{\boldsymbol{A}}$ and $\boldsymbol{X}$, and the constraint $\widetilde{\boldsymbol{a}}_k \in \mathbb{S}^{n-1}$ is nonconvex, so the overall problem is nonconvex. Nonetheless, simple alternating minimization methods have shown empirical success in many applications [GCW17]. However, as the lasso formulation for $\boldsymbol{x}_k$ does not have closed-form solution, it makes the marginalized function

$$\varphi_L^N\left(\widetilde{\boldsymbol{A}}\right) = \min_{\boldsymbol{X}} \frac{1}{2}\left\|\boldsymbol{y} - \sum_{k=1}^N \widetilde{\boldsymbol{a}}_k \circledast \boldsymbol{x}_k\right\|^2 + \lambda \sum_{k=1}^N \|\boldsymbol{x}_k\|_1 ,$$

very difficult to analyze. By our incoherence assumption of $\boldsymbol{a}_k$ with small incoherence parameter μ, the quadratic term approximately equals to

$$\begin{aligned}
\left\|\boldsymbol{y} - \sum_{k=1}^N \widetilde{\boldsymbol{a}}_k \circledast \boldsymbol{x}_k\right\|^2 &= \|\boldsymbol{y}\|^2 - 2\left\langle \boldsymbol{y}, \sum_{k=1}^N \widetilde{\boldsymbol{a}}_k \circledast \boldsymbol{x}_k \right\rangle + \left\|\sum_{k=1}^N \widetilde{\boldsymbol{a}}_k \circledast \boldsymbol{x}_k\right\|^2 \\
&\approx \|\boldsymbol{y}\|^2 - 2\left\langle \boldsymbol{y}, \sum_{k=1}^N \widetilde{\boldsymbol{a}}_k \circledast \boldsymbol{x}_k \right\rangle + \sum_{k=1}^N \boldsymbol{x}_k^* \underbrace{\boldsymbol{C}^*_{\boldsymbol{\iota}_{n\to m}\widetilde{\boldsymbol{a}}_k} \boldsymbol{C}_{\boldsymbol{\iota}_{n\to m}\widetilde{\boldsymbol{a}}_k}}_{\approx \boldsymbol{I}} \boldsymbol{x}_k \\
&\approx \|\boldsymbol{y}\|^2 - 2\left\langle \boldsymbol{y}, \sum_{k=1}^N \widetilde{\boldsymbol{a}}_k \circledast \boldsymbol{x}_k \right\rangle + \sum_{k=1}^N \|\boldsymbol{x}_k\|^2 .
\end{aligned}$$

Therefore, we could consider a variant of the lasso formulation,

$$\begin{aligned}
\Phi_{DQ}^N\left(\widetilde{\boldsymbol{A}}, \boldsymbol{X}\right) &= \frac{1}{2}\|\boldsymbol{y}\|^2 - \left\langle \boldsymbol{y}, \sum_{k=1}^N \widetilde{\boldsymbol{a}}_k \circledast \boldsymbol{x}_k \right\rangle + \frac{1}{2}\sum_{k=1}^N \|\boldsymbol{x}_k\|^2 + \lambda \sum_{k=1}^N \|\boldsymbol{x}_k\|_1 , \\
&= \frac{1}{2}\|\boldsymbol{y}\|^2 + \sum_{k=1}^N \left(\frac{1}{2}\|\boldsymbol{x}_k\|^2 - \langle \boldsymbol{y}, \widetilde{\boldsymbol{a}}_k \circledast \boldsymbol{x}_k\rangle + \lambda \|\boldsymbol{x}_k\|_1\right),
\end{aligned}$$

which we call it the drop quadratic (DQ) loss. we have closed-form solutions for $\boldsymbol{X}$ with $\widetilde{\boldsymbol{A}}$ fixed,

$$\boldsymbol{x}_k^\star = \arg\min_{\boldsymbol{x}_k} \widetilde{\Phi}_{DQ}(\widetilde{\boldsymbol{A}}, \boldsymbol{X}) = S_\lambda\left\{(\mathcal{R}_n \boldsymbol{a}_k) \circledast \boldsymbol{y}\right\}.$$

Plugging $\boldsymbol{x}_k^\star$ back, we could obtain the marginalized objective function of $\boldsymbol{a}_k$ as

$$\varphi_{DQ}^N\left(\widetilde{\boldsymbol{A}}\right) = \frac{1}{2}\|\boldsymbol{y}\|^2 - \frac{1}{2}\sum_{k=1}^N \|S_\lambda\left[(\mathcal{R}_n \widetilde{\boldsymbol{a}}_k) \circledast \boldsymbol{y}\right]\|^2 . \tag{26.1.2}$$

However, the drop quadratic loss decouples the dependence of $\{\boldsymbol{a}_k\}_{k=1}^N$ across k. If we minimize the objective over the oblique manifold, we could obtain multiple duplicate solutions for the kernels. Instead, we could consider finding all the kernels one by one using a deflation approach. As observed from (26.1.2), the

objective function $\varphi_{DQ}(\widetilde{\boldsymbol{A}})$ is decoupled with respect to $\{\boldsymbol{a}_k\}_{k=1}^N$, we can try to find the kernels $\{\boldsymbol{a}_k\}_{k=1}^N$ one by one via minimizing

$$\varphi_{\mathrm{DQ}}(\boldsymbol{a}) \doteq \frac{1}{2}\left\|\boldsymbol{y}\right\|^2 - \frac{1}{2}\left\|S_\lambda\left[\boldsymbol{C}_{\boldsymbol{a}}^*\boldsymbol{y}\right]\right\|^2, \qquad \boldsymbol{a} \in \mathbb{S}^{n-1} \tag{26.1.3}$$

The optimal solution $\boldsymbol{a}_\star = \arg\min_{\boldsymbol{a}} \varphi_{DQ}(\boldsymbol{a})$ produces an approximation of one of those $\{\boldsymbol{a}_k\}_{k=1}^K$. Using the approximation, we solve a lasso problem

$$\min_{\boldsymbol{x}} \; F(\boldsymbol{x}) = \underbrace{\frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{a}_\star \circledast \boldsymbol{x}\right\|^2}_{f(\boldsymbol{x})} + \underbrace{\lambda\left\|\boldsymbol{x}\right\|_1}_{g(\boldsymbol{x})}$$

to find the corresponding $\boldsymbol{x}_\star$. After subtracting $\boldsymbol{y}$ by $\boldsymbol{a}_\star \circledast \boldsymbol{x}_\star$, i.e.,

$$\boldsymbol{y} \;\leftarrow\; \boldsymbol{y} - \boldsymbol{a}_\star \circledast \boldsymbol{x}_\star,$$

we repeat the whole process until all $\{\boldsymbol{a}_k\}_{k=1}^N$ and $\{\boldsymbol{x}_k\}_{k=1}^N$ are recovered. Our premature analysis implies that the objective function (26.1.3) is a ridable saddle function over a local portion on the sphere, which implies that we could find an approximate solution of one of the kernels with efficient methods.

Last but not least, it should be noticed that our numerical simulations implies that the quadratic-free approximation provides a problem formulation amenable to analysis, but at a significant trade-off to statistical efficiency. Specifically, for $N = 1$, solving a typical drop quadratic problem to high statistical precision would require

$$n \geq \Omega(10^5), \quad m \geq \Omega(10^7), \quad \theta \leq n^{-0.5},$$

while for bilinear lasso the optimal solution can be reliably recovered with problem size as small as

$$n \sim O(1), \quad m \geq \Omega(10^2), \quad \theta \leq n^{-0.5}.$$

Given advantage of the statistical efficiency, it would be interesting of how to directly analyze the original bilinear formulation .

26.2 Overcomplete Dictionary Learning/Tensor Decomposition

Overcomplete tensor decomposition. Another nonconvex problem of great interest in theoretical computer science is the overcomplete tensor decomposition problem. For example, consider decomposing a

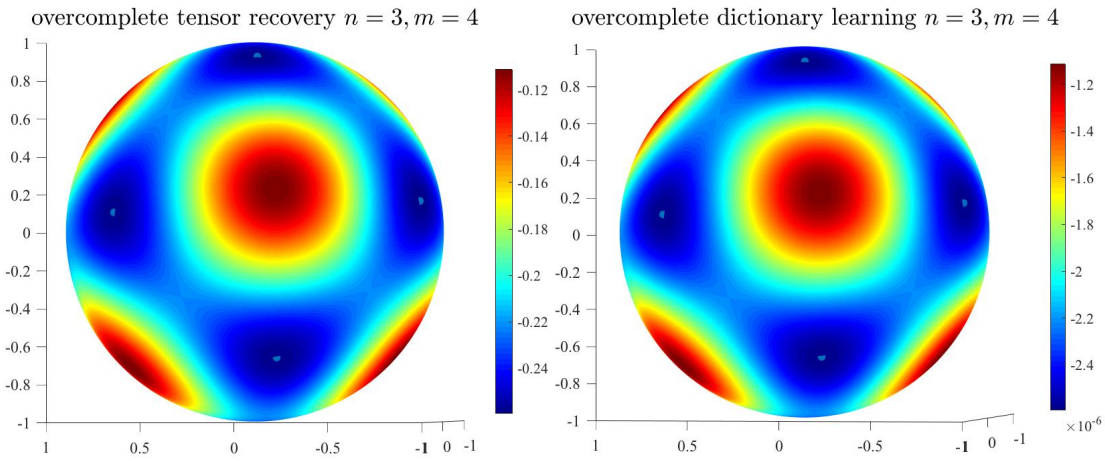


Figure 26.1: Function landscape of $\varphi(\boldsymbol{q})$ (left) and $\overline{\varphi}(\boldsymbol{q})$ (right) over the sphere $\mathbb{S}^2$, with preconditioned $\boldsymbol{A} \in \mathbb{R}^{3\times 4}$.

4-th order tensor $\mathcal{T}$ of rank m in the following form,

$$\mathcal{T} = \sum_{k=1}^{m} \boldsymbol{a}_k \otimes \boldsymbol{a}_k \otimes \boldsymbol{a}_k \otimes \boldsymbol{a}_k,$$

where $\boldsymbol{a}_k \in \mathbb{R}^n$ are the true components. We are interested in the overcomplete regime where the number of components $m \gg n$. Suppose we are given all the entries of the tensor $\mathcal{T}$, our goal is to recover all the components $\{\boldsymbol{a}_k\}_{k=1}^m$. Previously, Ge et al. [GHJY15] show that for the orthogonal case where $n \leq m$ and all the $\boldsymbol{a}_k$ are orthogonal, the objective function $\varphi(\boldsymbol{q})$ have only $2m$ local minima that are approximately the true components. However, the technique heavily uses the orthogonality of the components and is not generalizable to the overcomplete case. The overcomplete setting is much more challenging, but it is crucial for unsupervised learning applications where the hidden representations have higher dimension than the data [DLCC07, KB09, AGMM15, AGJ15]. Previous algorithmic results either require access to high order tensors [BCMV14, GVX14], or use complicated techniques such as FOOBI [DLCC07] or sum-of-squares relaxation [BKS15, GM15, HSSS16, MSS16, SS17], whose computational complexity is quasi-polynomial. Instead, we could directly analyze the following non-convex objective

$$\min \varphi(\boldsymbol{q}) \doteq -\mathcal{T}(\boldsymbol{q},\boldsymbol{q},\boldsymbol{q},\boldsymbol{q}) = -\left\|\boldsymbol{A}^\top \boldsymbol{q}\right\|_4^4, \quad \text{s.t.} \quad \|\boldsymbol{q}\| = 1, \tag{26.2.1}$$

where $\boldsymbol{A} = \left[\boldsymbol{a}_1^\top, \cdots, \boldsymbol{a}_m^\top\right]$. Empirically, under proper assumptions of $\boldsymbol{a}_k$, (Riemannian) gradient decent of $\varphi(\boldsymbol{q})$ with random initialization finds one of the solution even if m is significantly larger than n. In the literature, the local geometry for the over-complete case around the true components is known: in a small neigh-

borhood of each component, the function is strongly convex and there is a unique local minima [AGJ14a]. Ge and Ma [GM17] further expand the "nice" region by showing that there is no spurious local minima whenever the objective is a little bit smaller than its expected value. However, the size of the enlarged region they characterize decreases exponentially as data dimension increases. It remains a major open question whether there are any other spurious local minima over the rest of the sphere. Based on extensive simulations and function landscape in low dimension (Fig. 26.1), our conjecture is that when $\boldsymbol{A}$ is i.i.d. Gaussian, the function is ridable saddle and there is no spurious local minimizer over the sphere.

Overcomplete dictionary learning. Another important problem is the overcomplete dictionary learning, which has many applications in signal processing and machine learning [Ela10, MBP14]. Given the underlying generative model of the observed data $\boldsymbol{Y}$,

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}, \qquad \boldsymbol{A} \in \mathbb{R}^{n\times m},\ \boldsymbol{X} \in \mathbb{R}^{m\times p},$$

where $\boldsymbol{A}$ is called the dictionary and $\boldsymbol{X}$ is the sparse code, the problem of dictionary learning is to find the underlying dictionary $\boldsymbol{A}$ from $\boldsymbol{Y}$. When the dictionary $\boldsymbol{A}$ is complete (i.e., square and nonsingular), the row space of $\boldsymbol{Y}$ equals to the row space of $\boldsymbol{X}$ (i.e., $\mathrm{row}(\boldsymbol{Y}) = \mathrm{row}(\boldsymbol{X})$). As discussed in this thesis, the dictionary learning problem is equivalent to *finding the sparsest vector in the subspace* $\mathcal{S} = \mathrm{row}(\boldsymbol{Y})$ [SWW12b, QSW14, DH14]. Let $h(\cdot)$ be a sparse promoting function, Chapter xx in this thesis reveals that the nonconvex problem

$$\min_{\boldsymbol{q}}\ h(\boldsymbol{q}^\top \boldsymbol{Y}), \quad \text{s.t.} \quad \|\boldsymbol{q}\| = 1,$$

has no spurious local minima, and every the local minima corresponds to an approximation of one row of $\boldsymbol{X}$. The new discovery has lead to the development of efficient optimization methods [SQW15c]. Recently, [SS17] proposed a spectral method for dictionary learning based on sum of squares relaxation. However, all of these methods exploit the fact that $\mathrm{row}(\boldsymbol{Y}) = \mathrm{row}(\boldsymbol{X})$ when $\boldsymbol{A}$ is complete, and it cannot be generalized to the overcomplete setting $m > n$.

In this work, we are interested in the case when $\boldsymbol{A}$ is overcomplete $m \gg n$. Instead of recovering rows of $\boldsymbol{X}$, we seek to find the columns of $\boldsymbol{A}$ by solving the following nonconvex objective,

$$\min_{\boldsymbol{q}} \overline{\varphi}(\boldsymbol{q}) = -\frac{1}{4m}\left\|\boldsymbol{q}^\top \boldsymbol{Y}\right\|_4^4, \quad \text{s.t.} \quad \|\boldsymbol{q}\| = 1. \tag{26.2.2}$$

We show that under proper random assumptions of $\boldsymbol{A}$ and $\boldsymbol{X}$, the optimal solution of $\overline{\varphi}(\boldsymbol{q})$ corresponds to one column of $\boldsymbol{A}$, and the objective function has no spurious local minima. More specifically, when

$\boldsymbol{A}\boldsymbol{A}^\top \approx \boldsymbol{I}$ and assume that $\boldsymbol{X}$ is Bernoulli-Gaussian, we can show that

$$\mathbb{E}_{\boldsymbol{X}}\left[\overline{\varphi}(\boldsymbol{q})\right] \;\approx\; c_1\varphi(\boldsymbol{q}) + c_2,$$

where c_1 and c_2 are some numerical constants. This implies that, with respect to the randomness of $\boldsymbol{X}$, the expectation of optimization landscape $\overline{\varphi(\boldsymbol{q})}$ of the overcomplete dictionary learning can be reduced to that of the overcomplete tensor decomposition. Therefore, if the conjecture that the overcomplete tensor problem is ridable, one can expect a similar benign geometric structure for overcomplete dictionary problem by an expectation-concentration type analysis.

Bibliography

Bibliography

[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.

[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.

[AG93] Miguel A Arcones and Evarist Giné. On decoupling, series expansions, and tail behavior of chaos processes. *Journal of Theoretical Probability*, 6(1):101–122, 1993.

[AG16] Anima Anandkumar and Rong Ge. Efficient approaches for escaping higher order saddle points in non-convex optimization. *arXiv preprint arXiv:1602.05908*, 2016.

[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.

[AGJ15] Animashree Anandkumar, Rong Ge, and Majid Janzamin. Learning overcomplete latent variable models through tensor methods. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 36–112, Paris, France, 03–06 Jul 2015. PMLR.

[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.

[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.

[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. *IEEE Transactions on Information Theory*, 60(3):1711–1732, 2014.

[BAC16] Nicolas Boumal, P-A Absil, and Coralia Cartis. Global rates of convergence for nonconvex optimization on manifolds. *arXiv preprint arXiv:1605.08101*, 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.

[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.

[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.

[BCMV14] Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan. Smoothed analysis of tensor decompositions. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 594–603. ACM, 2014.

[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.

[BE16] Tamir Bendory and Yonina C Eldar. Non-convex phase retrieval from STFT measurements. *arXiv preprint arXiv:1607.08218*, 2016.

[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. 1999.

[BH89] Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

[BHK+85] M.W. Berry, M.T. Heath, I. Kaneko, M. Lawo, R.J. Plemmons, and R.C. Ward. An algorithm to compute a sparse basis of the null space. *Numerische Mathematik*, 47(4):483–504, 1985.

[BJ03] Ronen Basri and David W Jacobs. Lambertian reflectance and linear subspaces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(2):218–233, 2003.

[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.

[BKS15] Boaz Barak, Jonathan A Kelner, and David Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 143–151. ACM, 2015.

[BL14] Hilton Bristow and Simon Lucey. Optimization methods for convolutional sparse coding. *arXiv preprint arXiv:1406.2407*, 2014.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

[BM05] Gregory Beylkin and Lucas Monzón. On approximation of functions by exponential sums. *Applied and Computational Harmonic Analysis*, 19(1):17–48, 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.

[BNS16] Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Global optimality of local search for low rank matrix recovery. *arXiv preprint arXiv:1605.07221*, 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.

[BR16] Sohail Bahmani and Justin Romberg. Phase retrieval meets statistical learning theory: A flexible convex relaxation. *arXiv preprint arXiv:1610.04210*, 2016.

[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.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[BVB16] Nicolas Boumal, Vladislav Voroninski, and Afonso S Bandeira. The non-convex burer-monteiro approach works on smooth semidefinite programs. *arXiv preprint arXiv:1606.04970*, 2016.

[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.

[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.

[CFL] Pengwen Chen, Albert Fannjiang, and Gi-Ren Liu. Phase retrieval with one or two diffraction patterns by alternating projections with the null initialization. *Journal of Fourier Analysis and Applications*, pages 1–40.

[CGT00] Andrew R. Conn, Nicholas I.M. Gould, and Philippe L. Toint. *Trust region methods*, volume 1. SIAM, 2000.

[CGT12] Coralia Cartis, Nicholas IM Gould, and Ph L Toint. Complexity bounds for second-order optimality in unconstrained optimization. *Journal of Complexity*, 28(1):93–108, 2012.

[CJ16] Valerio Cambareri and Laurent Jacques. Through the haze: A non-convex approach to blind calibration for linear random sensing models. *arXiv preprint arXiv:1610.09028*, 2016.

[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.

[CLMW11a] Emmanuel Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM*, 58(3), May 2011.

[CLMW11b] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[CLS14] Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *arXiv preprint arXiv:1407.1065*, 2014.

[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.

[CM14a] Sunav Choudhary and Urbashi Mitra. Identifiability scaling laws in bilinear inverse problems. *arXiv preprint arXiv:1402.2637*, 2014.

[CM14b] Sunav Choudhary and Urbashi Mitra. Sparse blind deconvolution: What cannot be done. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 3002–3006. IEEE, 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.

[CP86] Thomas F Coleman and Alex Pothen. The null space problem i. complexity. *SIAM Journal on Algebraic Discrete Methods*, 7(4):527–537, 1986.

[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.

[CRT06a] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

[CRT06b] Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

[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.

[CT05] Emmanuel J Candès and Terence Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.

[CT06] Emmanuel J Candes and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE transactions on information theory*, 52(12):5406–5425, 2006.

[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.

[dEGJL07] Alexandre d'Aspremont, Laurent El Ghaoui, Michael I Jordan, and Gert RG Lanckriet. A direct formulation for sparse pca using semidefinite programming. *SIAM review*, 49(3):434–448, 2007.

[dEJL07] A. d'Aspremont, L. El Ghaoui, M. Jordan, and G. Lanckriet. A direct formulation of sparse PCA using semidefinite programming. *SIAM Review*, 49(3), 2007.

[DER86] Iain S Duff, Albert M Erisman, and John K Reid. *Direct Methods for Sparse Matrices*. Oxford University Press, Inc., New York, NY, USA, 1986.

[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.

[DLCC07] Lieven De Lathauwer, Josphine Castaing, and Jean-Franois Cardoso. Fourth-order cumulant-based blind identification of underdetermined mixtures. *IEEE Transactions on Signal Processing*, 55(6):2965–2973, 2007.

[DLH12] Yuchao Dai, Hongdong Li, and Mingyi He. A simple prior-free method for non-rigid structure-from-motion factorization. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2018–2025. IEEE, 2012.

[DlPG99] Victor De la Pena and Evarist Giné. *Decoupling: from dependence to independence*. Springer, 1999.

[Don06] David L Donoho. For most large underdetermined systems of linear equations the minimal ℓ^1-norm solution is also the sparsest solution. *Communications on pure and applied mathematics*, 59(6):797–829, 2006.

[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.

[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.

[EK12] Yonina C Eldar and Gitta Kutyniok. *Compressed sensing: theory and applications*. Cambridge University Press, 2012.

[Ela10] Michael Elad. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer, 2010.

[ETS11] Chaitanya Ekanadham, Daniel Tranchina, and Eero P. Simoncelli. A blind sparse deconvolution method for neural spike identification. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1440–1448. Curran Associates, Inc., 2011.

[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.

[FLM77] Tadeusz Figiel, Joram Lindenstrauss, and Vitali D Milman. The dimension of almost spherical sections of convex bodies. *Acta Mathematica*, 139(1):53–94, 1977.

[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.

[GCW17] Cristina Garcia-Cardona and Brendt Wohlberg. Convolutional dictionary learning. *arXiv preprint arXiv:1709.02893*, 2017.

[GG84] Andrej Y Garnaev and Efim D Gluskin. The widths of a euclidean ball. In *Dokl. Akad. Nauk SSSR*, volume 277, pages 1048–1052, 1984.

[GH87] John R Gilbert and Michael T. Heath. Computing a sparse basis for the null space. *SIAM Journal on Algebraic Discrete Methods*, 8(3):446–459, 1987.

[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.

[GK76] Robert M Gagliardi and Sherman Karp. Optical communications. *New York, Wiley-Interscience, 1976. 445 p.*, 1, 1976.

[GKK13] David Gross, Felix Krahmer, and Richard Kueng. A partial derandomization of phaselift using spherical designs. *arXiv preprint arXiv:1310.2267*, 2013.

[GLM16] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. *arXiv preprint arXiv:1605.07272*, 2016.

[GM03] E Gluskin and V Milman. Note on the geometric-arithmetic mean inequality. In *Geometric aspects of Functional analysis*, pages 131–135. Springer, 2003.

[GM15] Rong Ge and Tengyu Ma. Decomposing overcomplete 3rd order tensors using sum-of-squares algorithms. *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, page 829, 2015.

[GM17] Rong Ge and Tengyu Ma. On the optimization landscape of tensor decompositions. *arXiv preprint arXiv:1706.05598*, 2017.

[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.

[GS16] Tom Goldstein and Christoph Studer. Phasemax: Convex phase retrieval via basis pursuit. *arXiv preprint arXiv:1610.07531*, 2016.

[GVX14] Navin Goyal, Santosh Vempala, and Ying Xiao. Fourier pca and robust tensor decomposition. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 584–593. ACM, 2014.

[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.

[GX16] Bing Gao and Zhiqiang Xu. Gauss-newton method for phase retrieval. *arXiv preprint arXiv:1606.08135*, 2016.

[HA] Furong Huang and Animashree Anandkumar. Convolutional dictionary learning through tensor factorization.

[Har13] Moritz Hardt. On the provable convergence of alternating minimization for matrix completion. *arXiv preprint arXiv:1312.0925*, 2013.

[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.

[HD13] Paul Hand and Laurent Demanet. Recovering the sparsest element in a subspace. *arXiv preprint arXiv:1310.1654*, 2013.

[HHW15] Felix Heide, Wolfgang Heidrich, and Gordon Wetzstein. Fast and flexible convolutional sparse coding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5135–5143, 2015.

[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.

[HMR16] Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *arXiv preprint arXiv:1609.05191*, 2016.

[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.

[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.

[HSSS16] Samuel B Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer. Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 178–191. ACM, 2016.

[HV16] Paul Hand and Vladislav Voroninski. An elementary proof of convex phase retrieval in the natural parameter space via the linear program phasemax. *arXiv preprint arXiv:1611.03935*, 2016.

[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.

[HXV13] Jeffrey Ho, Yuchen Xie, and Baba Vemuri. On a nonlinear generalization of sparse coding and dictionary learning. In *Proceedings of The 30th International Conference on Machine Learning*, pages 1480–1488, 2013.

[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.

[JGN+17] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. *arXiv preprint arXiv:1703.00887*, 2017.

[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.

[JL09] Iain M Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486), 2009.

[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.

[Kaw16] Kenji Kawaguchi. Deep learning without poor local minima. *arXiv preprint arXiv:1605.07110*, 2016.

[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.

[KMMP04] Telikepalli Kavitha, Kurt Mehlhorn, Dimitrios Michail, and Katarzyna Paluch. A faster algorithm for minimum cycle basis of graphs. In *31st International Colloquium on Automata, Languages and Programming*, pages 846–857. Springer, 2004.

[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.

[KMR14] Felix Krahmer, Shahar Mendelson, and Holger Rauhut. Suprema of chaos processes and the restricted isometry property. *Communications on Pure and Applied Mathematics*, 67(11):1877–1904, 2014.

[KNV+15] Robert Krauthgamer, Boaz Nadler, Dan Vilenchik, et al. Do semidefinite relaxations solve sparse PCA up to the information limit? *The Annals of Statistics*, 43(3):1300–1322, 2015.

[KÖ16] Ritesh Kolte and Ayfer Özgür. Phase retrieval via incremental truncated wirtinger flow. *arXiv preprint arXiv:1606.03196*, 2016.

[KR14] Felix Krahmer and Holger Rauhut. Structured random measurements in signal processing. *GAMM-Mitteilungen*, 37(2):217–238, 2014.

[Kwa87] Stanislaw Kwapien. Decoupling inequalities for polynomial chaos. *The Annals of Probability*, pages 1062–1071, 1987.

[led07] On measure concentration of vector-valued maps. *Bulletin of the Polish Academy of Sciences. Mathematics*, 55(3):261–278, 2007.

[LGBB05] Sylvain Lesage, Rémi Gribonval, Frédéric Bimbot, and Laurent Benaroya. Learning unions of orthonormal bases with thresholded singular value decomposition. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages v–293. IEEE, 2005.

[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.

[LLSW16] Xiaodong Li, Shuyang Ling, Thomas Strohmer, and Ke Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *arXiv preprint arXiv:1606.04933*, 2016.

[LO79] Jae S Lim and Alan V Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, 67(12):1586–1604, 1979.

[Loh15] Po-Ling Loh. Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *arXiv preprint arXiv:1501.00312*, 2015.

[LS15] Shuyang Ling and Thomas Strohmer. Self-calibration and biconvex compressive sensing. *Inverse Problems*, 31(11):115002, 2015.

[LS16] Shuyang Ling and Thomas Strohmer. Self-calibration via linear least squares. *arXiv preprint arXiv:1611.04196*, 2016.

[LS17] Shuyang Ling and Thomas Strohmer. Regularized gradient descent: A nonconvex recipe for fast joint blind deconvolution and demixing. *arXiv preprint arXiv:1703.08642*, 2017.

[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.

[LTR16] Kiryung Lee, Ning Tian, and Justin Romberg. Fast and guaranteed blind multichannel deconvolution under a bilinear system model. *arXiv preprint arXiv:1610.06469*, 2016.

[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.

[LV+15a] Jing Lei, Vincent Q Vu, et al. Sparsistency and agnostic inference in sparse pca. *The Annals of Statistics*, 43(1):299–322, 2015.

[LV15b] 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.

[LWDF09] Anat Levin, Yair Weiss, Fredo Durand, and William T Freeman. Understanding and evaluating blind deconvolution algorithms. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1964–1971. IEEE, 2009.

[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.

[McC83] S Thomas McCormick. A combinatorial approach to some sparse matrix problems. Technical report, DTIC Document, 1983.

[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.

[MHWG13] Cun Mu, Bo Huang, John Wright, and Donald Goldfarb. Square deal: Lower bounds and improved relaxations for tensor recovery. *arXiv preprint arXiv:1307.5870*, 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.

[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.

[MR15] C. Tsakiris Manolis and Vidal Rene. Dual principal component pursuit. *arXiv preprint arXiv:1510.04390*, 2015.

[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.

[MSS16] Tengyu Ma, Jonathan Shi, and David Steurer. Polynomial-time tensor decompositions with sum-of-squares. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 438–446. IEEE, 2016.

[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.

[MW15] Tengyu Ma and Avi Wigderson. Sum-of-squares lower bounds for sparse pca. *arXiv preprint arXiv:1507.06370*, 2015.

[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.

[NSU15] Yuji Nakatsukasa, Tasuku Soma, and André Uschmajew. Finding a low-rank basis in a matrix subspace. *CoRR*, abs/1503.08601, 2015.

[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.

[Pis99] Gilles Pisier. *The volume of convex bodies and Banach space geometry*, volume 94. Cambridge University Press, 1999.

[PKCS16] Dohyung Park, Anastasios Kyrillidis, Constantine Caramanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the burer-monteiro approach. *arXiv preprint arXiv:1609.03240*, 2016.

[PP16] Ioannis Panageas and Georgios Piliouras. Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions. *CoRR, vol. abs/1605.00405*, 2016.

[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.

[QZEW17] Qing Qu, Yuqian Zhang, Yonina C Eldar, and John Wright. Convolutional phase retrieval via gradient descent. *arXiv preprint arXiv:1712.00716*, 2017.

[Rau10] Holger Rauhut. Compressive sensing and structured random matrices. *Theoretical foundations and numerical methods for sparse recovery*, 9:1–92, 2010.

[Rei65] H. Reichenbach. In *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.

[SA14] 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.

[SC16] Daniel Soudry and Yair Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361*, 2016.

[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.

[SCP94] Milica Stojanovic, Josko A Catipovic, and John G Proakis. Phase-coherent digital communications for underwater acoustic channels. *IEEE Journal of Oceanic Engineering*, 19(1):100–111, 1994.

[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.

[SFB18] Andrew H Song, Francisco Flores, and Demba Ba. Spike sorting by convolutional dictionary learning. *arXiv preprint arXiv:1806.01979*, 2018.

[SGD+15] Arash Shahmansoori, Gabriel E Garcia, Giuseppe Destino, Gonzalo Seco-Granados, and Henk Wymeersch. 5g position and orientation estimation through millimeter wave mimo. In *Globecom Workshops (GC Wkshps), 2015 IEEE*, pages 1–6. IEEE, 2015.

[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.

[Sol17] Mahdi Soltanolkotabi. Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization. *CoRR*, abs/1702.06175, 2017.

[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. Complete dictionary recovery over the sphere I: Overview and the geometric picture. *arXiv preprint arXiv:1511.03607*, 2015.

[SQW15c] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere II: Recovery by Riemannian trust-region method. *arXiv preprint arXiv:1511.04777*, 2015.

[SQW15d] 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.

[SS00] Alex J. Smola and Bernhard Schölkopf. Sparse greedy matrix approximation for machine learning. pages 911–918. Morgan Kaufmann, 2000.

[SS17] Tselil Schramm and David Steurer. Fast and robust tensor decomposition with applications to dictionary learning. *arXiv preprint arXiv:1706.08672*, 2017.

[SWW12a] 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.

[SWW12b] Daniel A. Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *Conference on Learning Theory*, 2012.

[Tal14a] Michel Talagrand. *Upper and Lower Bounds for Stochastic Processes*. Springer, 2014.

[Tal14b] Michel Talagrand. *Upper and Lower Bounds for Stochastic Processes: Modern Methods and Classical Problems*, volume 60. Springer Science & Business Media, 2014.

[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.

[Tem03] Vladimir N Temlyakov. Nonlinear methods of approximation. *Foundations of Computational Mathematics*, 3(1):33–107, 2003.

[TKD+96] Thomas Tsang, Marco A Krumbügel, Kenneth W DeLong, David N Fittinghoff, and Rick Trebino. Frequency-resolved optical-gating measurements of ultrashort pulses using surface third-harmonic generation. *Optics letters*, 21(17):1381–1383, 1996.

[Tse01] Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.

[VCLR13] Vincent Q Vu, Juhee Cho, Jing Lei, and Karl Rohe. Fantope projection and selection: A near-optimal convex relaxation of sparse pca. In *Advances in Neural Information Processing Systems*, pages 2670–2678, 2013.

[Ver10] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[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.

[Wal16] Irène Waldspurger. Phase retrieval with random gaussian sensing vectors by alternating projections. *arXiv preprint arXiv:1609.03088*, 2016.

[WBJ15] P. Walk, H. Becker, and P. Jung. Ofdm channel estimation via phase retrieval. In *Asilomar 2015*, 2015.

[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.

[WGE16] Gang Wang, Georgios B Giannakis, and Yonina C Eldar. Solving systems of random quadratic equations via truncated amplitude flow. *arXiv preprint*, 2016.

[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.

[ZCL16] Huishuai Zhang, Yuejie Chi, and Yingbin Liang. Provable non-convex phase retrieval with outliers: Median truncated wirtinger flow. *arXiv preprint arXiv:1603.03805*, 2016.

[ZF13] Yun-Bin Zhao and Masao Fukushima. Rank-one solutions for homogeneous linear matrix equations over the positive semidefinite cone. *Applied Mathematics and Computation*, 219(10):5569–5583, 2013.

[ZHT06] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

[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.

[ZL16] Huishuai Zhang and Yingbin Liang. Reshaped wirtinger flow for solving quadratic systems of equations. *arXiv preprint arXiv:1605.07719*, 2016.

[ZLK+17] Yuqian Zhang, Yenson Lau, Han-wen Kuo, Sky Cheung, Abhay Pasupathy, and John Wright. On the global geometry of sphere-constrained sparse blind deconvolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[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

Auxillary Results for Finding a Sparse Vector in a Subspace

A.1 Technical Tools and Preliminaries

In this appendix, we record several lemmas that are useful for our analysis.

Lemma A.1 *Let $\psi(x)$ and $\Psi(x)$ to denote the probability density function (pdf) and the cumulative distribution function (cdf) for the standard normal distribution:*

$$\begin{aligned}
&(\textit{Standard Normal pdf}) \quad \psi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \\
&(\textit{Standard Normal cdf}) \quad \Psi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left\{-\frac{t^2}{2}\right\} dt,
\end{aligned}$$

Suppose a random variable $X \sim \mathbb{N}(0, \sigma^2)$, with the pdf $f_\sigma(x) = \frac{1}{\sigma}\psi\left(\frac{x}{\sigma}\right)$, then for any $t_2 > t_1$ we have

$$\begin{aligned}
\int_{t_1}^{t_2} f_\sigma(x)dx &= \Psi\left(\frac{t_2}{\sigma}\right) - \Psi\left(\frac{t_1}{\sigma}\right), \\
\int_{t_1}^{t_2} x f_\sigma(x)dx &= -\sigma\left[\psi\left(\frac{t_2}{\sigma}\right) - \psi\left(\frac{t_1}{\sigma}\right)\right], \\
\int_{t_1}^{t_2} x^2 f_\sigma(x)dx &= \sigma^2\left[\Psi\left(\frac{t_2}{\sigma}\right) - \Psi\left(\frac{t_1}{\sigma}\right)\right] - \sigma\left[t_2\psi\left(\frac{t_2}{\sigma}\right) - t_1\psi\left(\frac{t_1}{\sigma}\right)\right].
\end{aligned}$$

Lemma A.2 (Taylor Expansion of Standard Gaussian *cdf* and *pdf*) *Assume $\psi(x)$ and $\Psi(x)$ be defined as above. There exists some universal constant $C_\psi > 0$ such that for any $x_0,\ x \in \mathbb{R}$,*

$$|\psi(x) - [\psi(x_0) - x_0\psi(x_0)(x - x_0)]| \le C_\psi (x - x_0)^2,$$
$$|\Psi(x) - [\Psi(x_0) + \psi(x_0)(x - x_0)]| \le C_\psi (x - x_0)^2.$$

Lemma A.3 (Matrix Induced Norms) *For any matrix $\boldsymbol{A} \in \mathbb{R}^{p\times n}$, the induced matrix norm from $\ell^p \to \ell^q$ is defined as*

$$\|\boldsymbol{A}\|_{\ell^p\to\ell^q} \doteq \sup_{\|\boldsymbol{x}\|_p=1} \|\boldsymbol{A}\boldsymbol{x}\|_q .$$

In particular, let $\boldsymbol{A} = [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n] = \left[\boldsymbol{a}^1, \cdots, \boldsymbol{a}^p\right]^\top$, we have

$$\|\boldsymbol{A}\|_{\ell^2\to\ell^1} = \sup_{\|\boldsymbol{x}\|=1} \sum_{k=1}^{p} \left|\boldsymbol{a}_k^\top \boldsymbol{x}\right|, \quad \|\boldsymbol{A}\|_{\ell^2\to\ell^\infty} = \max_{1\le k\le p} \left\|\boldsymbol{a}^k\right\|,$$
$$\|\boldsymbol{A}\boldsymbol{B}\|_{\ell^p\to\ell^r} \le \|\boldsymbol{A}\|_{\ell^q\to\ell^r} \|\boldsymbol{B}\|_{\ell^p\to\ell^q},$$

and $\boldsymbol{B}$ is any matrix of size compatible with $\boldsymbol{A}$.

Lemma A.4 (Moments of the Gaussian Random Variable) *If $X \sim \mathcal{N}\left(0, \sigma_X^2\right)$, then it holds for all integer $m \ge 1$ that*

$$\mathbb{E}|X|^m = \sigma_X^m (m-1)!! \left[\sqrt{\frac{2}{\pi}}\mathbb{1}_{m=2k+1} + \mathbb{1}_{m=2k}\right] \le \sigma_X^m (m-1)!!,\ k = \lfloor m/2 \rfloor.$$

Lemma A.5 (Moments of the χ Random Variable) *If $X \sim \chi(n)$, i.e., $X = \|\boldsymbol{x}\|$ for $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, then it holds for all integer $m \ge 1$ that*

$$\mathbb{E}X^m = 2^{m/2} \frac{\Gamma(m/2 + n/2)}{\Gamma(n/2)} \le m!!\ n^{m/2}.$$

Lemma A.6 (Moments of the χ^2 Random Variable) *If $X \sim \chi^2(n)$, i.e., $X = \|\boldsymbol{x}\|^2$ for $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, then it holds for all integer $m \ge 1$ that*

$$\mathbb{E}X^m = 2^m \frac{\Gamma(m + n/2)}{\Gamma(n/2)} = \prod_{k=1}^{m}(n + 2k - 2) \le \frac{m!}{2}(2n)^m.$$

Lemma A.7 (Moment-Control Bernstein's Inequality for Random Variables [FR13]) *Let $X_1, \ldots, X_p$ be i.i.d. real-valued random variables. Suppose that there exist some positive numbers R and σ_X^2 such that*

$$\mathbb{E}|X_k|^m \leq \frac{m!}{2}\sigma_X^2 R^{m-2}, \quad \text{for all integers } m \geq 2.$$

Let $S \doteq \frac{1}{p}\sum_{k=1}^p X_k$, then for all $t > 0$, it holds that

$$\mathbb{P}|S - \mathbb{E}S| \geq t \leq 2\exp\left(-\frac{pt^2}{2\sigma_X^2 + 2Rt}\right).$$

Lemma A.8 (Moment-Control Bernstein's Inequality for Random Vectors [SQW15a]) *Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p \in \mathbb{R}^d$ be i.i.d. random vectors. Suppose there exist some positive number R and σ_X^2 such that*

$$\mathbb{E}\left[\|\boldsymbol{x}_k\|^m\right] \leq \frac{m!}{2}\sigma_X^2 R^{m-2}, \quad \text{for all integers } m \geq 2.$$

Let $\boldsymbol{s} = \frac{1}{p}\sum_{k=1}^p \boldsymbol{x}_k$, then for any $t > 0$, it holds that

$$\mathbb{P}\left[\|\boldsymbol{s} - \mathbb{E}\left[\boldsymbol{s}\right]\| \geq t\right] \leq 2(d+1)\exp\left(-\frac{pt^2}{2\sigma_X^2 + 2Rt}\right).$$

Lemma A.9 (Gaussian Concentration Inequality) *Let $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$. Let $f : \mathbb{R}^p \mapsto \mathbb{R}$ be an L-Lipschitz function. Then we have for all $t > 0$ that*

$$\mathbb{P}\left[f(\boldsymbol{X}) - \mathbb{E}f(\boldsymbol{X}) \geq t\right] \leq \exp\left(-\frac{t^2}{2L^2}\right).$$

Lemma A.10 (Bounding Maximum Norm of Gaussian Vector Sequence) *Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_1}$ be a sequence of (not necessarily independent) standard Gaussian vectors in $\mathbb{R}^{n_2}$. It holds that*

$$\mathbb{P}\max_{i\in[n_1]}\|\boldsymbol{x}_i\| > \sqrt{n_2} + 2\sqrt{2\log(2n_1)} \leq (2n_1)^{-3}.$$

Proof Since the function $\|\cdot\|$ is 1-Lipschitz, by Gaussian concentration inequality, for any $i \in [n_1]$, we have

$$\mathbb{P}\|\boldsymbol{x}_i\| - \sqrt{\mathbb{E}\left\|\boldsymbol{x}_i\right\|^2} > t \leq \mathbb{P}\|\boldsymbol{x}_i\| - \mathbb{E}\left\|\boldsymbol{x}_i\right\| > t \leq \exp\left(-\frac{t^2}{2}\right)$$

for all $t > 0$. Since $\mathbb{E}\left\|\boldsymbol{x}_i\right\|^2 = n_2$, by a simple union bound, we obtain

$$\mathbb{P}\max_{i\in[n_1]}\|\boldsymbol{x}_i\|_> \sqrt{n_2} + t \leq \exp\left(-\frac{t^2}{2} + \log n_1\right)$$

for all $t > 0$. Taking $t = 2\sqrt{2\log(2n_1)}$ gives the claimed result. ∎

Corollary A.11 *Let* $\mathbf{\Phi} \in \mathbb{R}^{n_1 \times n_2} \sim_{i.i.d.} \mathcal{N}(0,1)$. *It holds that*

$$\|\mathbf{\Phi}\boldsymbol{x}\|_\infty \leq \left(\sqrt{n_2} + 2\sqrt{2\log(2n_1)}\right)\|\boldsymbol{x}\| \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^{n_2},$$

with probability at least $1-(2n_1)^{-3}$.

Proof Let $\mathbf{\Phi} = \left[\boldsymbol{\phi}^1, \cdots, \boldsymbol{\phi}^{n_1}\right]^\top$. Without loss of generality, let us only consider $\boldsymbol{x} \in \mathbb{S}^{n_2-1}$, we have

$$\|\mathbf{\Phi}\boldsymbol{x}\|_\infty = \max_{i\in[n_1]} \left|\boldsymbol{x}^\top \boldsymbol{\phi}^i\right| \leq \max_{i\in[n_1]} \left\|\boldsymbol{\phi}^i\right\|. \tag{A.1.1}$$

Invoking Lemma A.10 returns the claimed result. ∎

Lemma A.12 (Covering Number of a Unit Sphere [Ver10]) *Let* $\mathbb{S}^{n-1} = \{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x}\| = 1\}$ *be the unit sphere. For any* $\varepsilon \in (0,1)$, *there exists some* ε *cover of* $\mathbb{S}^{n-1}$ *w.r.t. the* ℓ^2 *norm, denoted as* $\mathcal{N}_\varepsilon$, *such that*

$$|\mathcal{N}_\varepsilon| \leq \left(1+\frac{2}{\varepsilon}\right)^n \leq \left(\frac{3}{\varepsilon}\right)^n.$$

Lemma A.13 (Spectrum of Gaussian Matrices, [Ver10]) *Let* $\mathbf{\Phi} \in \mathbb{R}^{n_1 \times n_2}$ $(n_1 > n_2)$ *contain i.i.d. standard normal entries. Then for every* $t \geq 0$, *with probability at least* $1 - 2\exp\left(-t^2/2\right)$, *one has*

$$\sqrt{n_1} - \sqrt{n_2} - t \leq \sigma_{\min}(\mathbf{\Phi}) \leq \sigma_{\max}(\mathbf{\Phi}) \leq \sqrt{n_1} + \sqrt{n_2} + t.$$

Lemma A.14 *For any* $\varepsilon \in (0,1)$, *there exists a constant* $C(\varepsilon) > 1$, *such that provided* $n_1 > C(\varepsilon)\, n_2$, *the random matrix* $\mathbf{\Phi} \in \mathbb{R}^{n_1 \times n_2} \sim_{i.i.d.} \mathcal{N}(0,1)$ *obeys*

$$(1-\varepsilon)\sqrt{\frac{2}{\pi}} n_1 \|\boldsymbol{x}\| \leq \|\mathbf{\Phi}\boldsymbol{x}\|_1 \leq (1+\varepsilon)\sqrt{\frac{2}{\pi}} n_1 \|\boldsymbol{x}\| \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^{n_2},$$

with probability at least $1 - 2\exp\left(-c(\varepsilon)\, n_1\right)$ *for some* $c(\varepsilon) > 0$.

Geometrically, this lemma roughly corresponds to the well known almost spherical section theorem [FLM77, GG84], see also [GM03]. A slight variant of this version has been proved in [Don06], borrowing ideas from [Pis99].

Proof By homogeneity, it is enough to show that the bounds hold for every $\boldsymbol{x}$ of unit ℓ^2 norm. For a fixed $\boldsymbol{x}_0$ with $\|\boldsymbol{x}_0\| = 1$, $\mathbf{\Phi}\boldsymbol{x}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. So $\mathbb{E}\|\mathbf{\Phi}\boldsymbol{x}\|_1 = \sqrt{\frac{2}{\pi}} n_1$. Note that $\|\cdot\|_1$ is $\sqrt{n_1}$-Lipschitz, by concentration of measure for Gaussian vectors in Lemma A.9, we have

$$\mathbb{P}\left|\|\mathbf{\Phi}\boldsymbol{x}\|_1 - \mathbb{E}\|\mathbf{\Phi}\boldsymbol{x}\|_1\right| > t \leq 2\exp\left(-\frac{t^2}{2n_1}\right)$$

for any $t > 0$. For a fixed $\delta \in (0,1)$, $\mathcal{S}^{n_2-1}$ can be covered by a δ-net N_δ with cardinality $\#N_\delta \le (1+2/\delta)^{n_2}$. Now consider the event

$$\mathcal{E} \doteq \left\{ (1-\delta)\sqrt{\frac{2}{\pi}}n_1 \le \|\boldsymbol{\Phi}\boldsymbol{x}\|_1 \le (1+\delta)\sqrt{\frac{2}{\pi}}n_1 \ \forall\, \boldsymbol{x} \in N_\delta \right\}.$$

A simple application of union bound yields

$$\mathbb{P}\mathcal{E}^c \le 2\exp\left(-\frac{\delta^2 n_1}{\pi} + n_2 \log\left(1+\frac{2}{\delta}\right)\right).$$

Choosing δ small enough such that

$$(1-3\delta)(1-\delta)^{-1} \ge 1-\varepsilon \text{ and } (1+\delta)(1-\delta)^{-1} \le 1+\varepsilon,$$

then conditioned on $\mathcal{E}$, we can conclude that

$$(1-\varepsilon)\sqrt{\frac{2}{\pi}}n_1 \le \|\boldsymbol{\Phi}\boldsymbol{x}\|_1 \le (1+\varepsilon)\sqrt{\frac{2}{\pi}}n_1 \ \forall\, \boldsymbol{x} \in \mathbb{S}^{n_2-1}.$$

Indeed, suppose $\mathcal{E}$ holds. Then it can easily be seen that any $\boldsymbol{z} \in \mathbb{S}^{n_2-1}$ can be written as

$$\boldsymbol{z} = \sum_{k=0}^{\infty} \lambda_k \boldsymbol{x}_k, \qquad \text{with } |\lambda_k| \le \delta^k, \boldsymbol{x}_k \in N_\delta \text{ for all } k.$$

Hence we have

$$\|\boldsymbol{\Phi}\boldsymbol{z}\|_1 = \left\|\boldsymbol{\Phi}\sum_{k=0}^{\infty}\lambda_k\boldsymbol{x}_k\right\|_1 \le \sum_{k=0}^{\infty}\delta^k\|\boldsymbol{\Phi}\boldsymbol{x}_k\|_1 \le (1+\delta)(1-\delta)^{-1}\sqrt{\frac{2}{\pi}}n_1.$$

Similarly,

$$\|\boldsymbol{\Phi}\boldsymbol{z}\|_1 = \left\|\boldsymbol{\Phi}\sum_{k=0}^{\infty}\lambda_k\boldsymbol{x}_k\right\|_1 \ge \left[1-\delta-\delta(1+\delta)(1-\delta)^{-1}\right]\sqrt{\frac{2}{\pi}}n_1 = (1-3\delta)(1-\delta)^{-1}\sqrt{\frac{2}{\pi}}n_1.$$

Hence, the choice of δ above leads to the claimed result. Finally, given $n_1 > Cn_2$, to make the probability $\mathbb{P}\mathcal{E}^c$ decaying in n_1, it is enough to set $C = \frac{2\pi}{\delta^2}\log\left(1+\frac{2}{\delta}\right)$. This completes the proof. ■

A.2 The Random Basis vs. Its Orthonormalized Version

In this appendix, we consider the planted sparse model

$$\overline{\boldsymbol{Y}} \;=\; [\boldsymbol{x}_0 \mid \boldsymbol{g}_1 \mid \cdots \mid \boldsymbol{g}_{n-1}] \;=\; [\boldsymbol{x}_0 \mid \boldsymbol{G}] \in \mathbb{R}^{p\times n}$$

as defined in (3.0.5), where

$$x_0(k) \sim_{i.i.d.} \frac{1}{\sqrt{\theta p}}\mathrm{Ber}\,(\theta)\,, \quad \boldsymbol{g}_\ell \sim_{i.i.d.} \mathcal{N}\left(\boldsymbol{0}, \frac{1}{p}\boldsymbol{I}\right), \quad 1 \le k \le p,\ 1 \le \ell \le n-1. \tag{A.2.1}$$

Recall that one "natural/canonical" orthonormal basis for the subspace spanned by columns of $\overline{\boldsymbol{Y}}$ is

$$\boldsymbol{Y} = \left[\frac{\boldsymbol{x}_0}{\|\boldsymbol{x}_0\|} \mid \mathcal{P}_{\boldsymbol{x}_0^\perp}\boldsymbol{G}\left(\boldsymbol{G}^\top\mathcal{P}_{\boldsymbol{x}_0^\perp}\boldsymbol{G}\right)^{-1/2}\right],$$

which is well-defined with high probability as $\mathcal{P}_{\boldsymbol{x}_0^\perp}\boldsymbol{G}$ is well-conditioned (proved in Lemma A.16). We write

$$\boldsymbol{G}' \doteq \mathcal{P}_{\boldsymbol{x}_0^\perp}\boldsymbol{G}\left(\boldsymbol{G}^\top\mathcal{P}_{\boldsymbol{x}_0^\perp}\boldsymbol{G}\right)^{-1/2} \tag{A.2.2}$$

for convenience. When p is large, $\overline{\boldsymbol{Y}}$ has nearly orthonormal columns, and so we expect that $\boldsymbol{Y}$ closely approximates $\overline{\boldsymbol{Y}}$. In this section, we make this intuition rigorous. We prove several results that are needed for the proof of Theorem 2.1, and for translating results for $\overline{\boldsymbol{Y}}$ to results for $\boldsymbol{Y}$ in Section 7.3.4.

For any realization of $\boldsymbol{x}_0$, let $\mathcal{I} = \mathrm{supp}(\boldsymbol{x}_0) = \{i \mid \boldsymbol{x}_0(i) \neq 0\}$. By Bernstein's inequality in Lemma A.7 with $\sigma_X^2 = 2\theta$ and $R = 1$, the event

$$\mathcal{E}_0 \doteq \left\{\frac{1}{2}\theta p \le |\mathcal{I}| \le 2\theta p\right\} \tag{A.2.3}$$

holds with probability at least $1 - 2\exp\left(-\theta p/16\right)$. Moreover, we show the following:

Lemma A.15 *When $p \ge Cn$ and $\theta > 1/\sqrt{n}$, the bound*

$$\left|1 - \frac{1}{\|\boldsymbol{x}_0\|}\right| \le \frac{4\sqrt{2}}{5}\sqrt{\frac{n\log p}{\theta^2 p}} \tag{A.2.4}$$

holds with probability at least $1 - cp^{-2}$. Here C, c are positive constants.

Proof Because $\mathbb{E}\|\boldsymbol{x}_0\|^2 = 1$, by Bernstein's inequality in Lemma A.7 with $\sigma_X^2 = 2/(\theta p^2)$ and $R = 1/(\theta p)$, we have

$$\mathbb{P}\left|\|\boldsymbol{x}_0\|^2 - \mathbb{E}\|\boldsymbol{x}_0\|^2\right| > t = \mathbb{P}\left|\|\boldsymbol{x}_0\|^2 - 1\right| > t \le 2\exp\left(-\frac{\theta p t^2}{4+2t}\right)$$

for all $t > 0$, which implies

$$\mathbb{P}|\|\boldsymbol{x}_0\| - 1| > \frac{t}{\|\boldsymbol{x}_0\|+1} = \mathbb{P}|\|\boldsymbol{x}_0\| - 1|\,(\|\boldsymbol{x}_0\|+1) > t \le 2\exp\left(-\frac{\theta p t^2}{4+2t}\right).$$

On the intersection with $\mathcal{E}_0$, $\|\boldsymbol{x}_0\| + 1 \geq \frac{1}{\sqrt{2}} + 1 \geq 5/4$ and setting $t = \sqrt{\frac{n \log p}{\theta^2 p}}$, we obtain

$$\mathbb{P}|\|\boldsymbol{x}_0\| - 1| \geq \frac{4}{5}\sqrt{\frac{n \log p}{\theta^2 p}} \;\Big|\; \mathcal{E}_0 \leq 2\exp\left(-\sqrt{np \log p}\right).$$

Unconditionally, this implies that with probability at least $1 - 2\exp(-p\theta/16) - 2\exp\left(-\sqrt{np\log p}\right)$, we have

$$\left|1 - \frac{1}{\|\boldsymbol{x}_0\|}\right| = \frac{|1 - \|\boldsymbol{x}_0\||}{\|\boldsymbol{x}_0\|} \leq \frac{4\sqrt{2}}{5}\sqrt{\frac{n \log p}{\theta^2 p}},$$

as desired. ∎

Let $\boldsymbol{M} \doteq \left(\boldsymbol{G}^\top \mathcal{P}_{\boldsymbol{x}_0^\perp} \boldsymbol{G}\right)^{-1/2}$. Then $\boldsymbol{G}' = \boldsymbol{G}\boldsymbol{M} - \frac{\boldsymbol{x}_0 \boldsymbol{x}_0^\top}{\|\boldsymbol{x}_0\|^2}\boldsymbol{G}\boldsymbol{M}$. We show the following results hold:

Lemma A.16 *Provided $p \geq Cn$, it holds that*

$$\|\boldsymbol{M}\|_{\leq} 2, \quad \|\boldsymbol{M} - \boldsymbol{I}\|_{\leq} 4\sqrt{\frac{n}{p}} + 4\sqrt{\frac{\log(2p)}{p}}$$

with probability at least $1 - (2p)^{-2}$. Here C is a positive constant.

Proof First observe that

$$\|\boldsymbol{M}\|_{=} \left(\sigma_{\min}\left(\boldsymbol{G}^\top \mathcal{P}_{\boldsymbol{x}_0^\perp}\boldsymbol{G}\right)\right)^{-1/2} = \sigma_{\min}^{-1}\left(\mathcal{P}_{\boldsymbol{x}_0^\perp}\boldsymbol{G}\right).$$

Now suppose $\boldsymbol{B}$ is an orthonormal basis spanning $\boldsymbol{x}_0^\perp$. Then it is not hard to see the spectrum of $\mathcal{P}_{\boldsymbol{x}_0^\perp}\boldsymbol{G}$ is the same as that of $\boldsymbol{B}^\top \boldsymbol{G} \in \mathbb{R}^{(p-1)\times(n-1)}$; in particular,

$$\sigma_{\min}\left(\mathcal{P}_{\boldsymbol{x}_0^\perp}\boldsymbol{G}\right) = \sigma_{\min}\left(\boldsymbol{B}^\top \boldsymbol{G}\right).$$

Since each entry of $\boldsymbol{G} \sim_{i.i.d.} \mathcal{N}\left(0, \frac{1}{p}\right)$, and $\boldsymbol{B}^\top$ has orthonormal rows, $\boldsymbol{B}^\top\boldsymbol{G} \sim_{i.i.d.} \mathcal{N}\left(0, \frac{1}{p}\right)$, we can invoke the spectrum results for Gaussian matrices in Lemma A.13 and obtain that

$$\sqrt{\frac{p-1}{p}} - \sqrt{\frac{n-1}{p}} - 2\sqrt{\frac{\log(2p)}{p}} \leq \sigma_{\min}\left(\boldsymbol{B}^\top\boldsymbol{G}\right) \leq \sigma_{\max}\left(\boldsymbol{B}^\top\boldsymbol{G}\right) \leq \sqrt{\frac{p-1}{p}} + \sqrt{\frac{n-1}{p}} + 2\sqrt{\frac{\log(2p)}{p}}$$

with probability at least $1 - (2p)^{-2}$. Thus, when $p \geq C_1 n$ for some sufficiently large constant C_1, by using the results above we have

$$\|\boldsymbol{M}\| = \sigma_{\min}^{-1}\left(\boldsymbol{B}^\top\boldsymbol{G}\right) = \left(\sqrt{\frac{p-1}{p}} - \sqrt{\frac{n-1}{p}} - 2\sqrt{\frac{\log(2p)}{p}}\right)^{-1} \leq 2,$$

$$\begin{aligned}\|\boldsymbol{I} - \boldsymbol{M}\| &= \max\left(\left|\sigma_{\max}(\boldsymbol{M}) - 1\right|, \left|\sigma_{\min}(\boldsymbol{M}) - 1\right|\right) \\ &= \max\left(\left|\sigma_{\min}^{-1}\left(\boldsymbol{B}^\top\boldsymbol{G}\right) - 1\right|, \left|\sigma_{\max}^{-1}\left(\boldsymbol{B}^\top\boldsymbol{G}\right) - 1\right|\right)\end{aligned}$$

$$
\begin{aligned}
&\leq \max\left\{\left(\sqrt{\frac{p-1}{p}}-\sqrt{\frac{n-1}{p}}-2\sqrt{\frac{\log(2p)}{p}}\right)^{-1}-1, 1-\left(\sqrt{\frac{p-1}{p}}+\sqrt{\frac{n-1}{p}}+2\sqrt{\frac{\log(2p)}{p}}\right)^{-1}\right\}\\
&= \max\left\{\left(1-\sqrt{\frac{p-1}{p}}+\sqrt{\frac{n-1}{p}}+2\sqrt{\frac{\log(2p)}{p}}\right)\left(\sqrt{\frac{p-1}{p}}-\sqrt{\frac{n-1}{p}}-2\sqrt{\frac{\log(2p)}{p}}\right)^{-1},\right.\\
&\qquad\left.\left(\sqrt{\frac{p-1}{p}}-1+\sqrt{\frac{n-1}{p}}+2\sqrt{\frac{\log(2p)}{p}}\right)\left(\sqrt{\frac{p-1}{p}}+\sqrt{\frac{n-1}{p}}+2\sqrt{\frac{\log(2p)}{p}}\right)^{-1}\right\}\\
&\leq 2\left(1-\sqrt{\frac{p-1}{p}}+\sqrt{\frac{n-1}{p}}+2\sqrt{\frac{\log(2p)}{p}}\right)\\
&\leq 4\sqrt{\frac{n}{p}}+4\sqrt{\frac{\log(2p)}{p}},
\end{aligned}
$$

with probability at least $1-(2p)^{-2}$. ■

Lemma A.17 *Let $\boldsymbol{Y}_{\mathcal{I}}$ be a submatrix of $\boldsymbol{Y}$ whose rows are indexed by the set $\mathcal{I}$. There exists a constant $C>0$, such that when $p \geq Cn$ and $1/2 > \theta > 1/\sqrt{n}$, the following*

$$
\begin{aligned}
\left\|\overline{\boldsymbol{Y}}\right\|_{\ell^2\to\ell^1} &\leq 3\sqrt{p},\\
\left\|\boldsymbol{Y}_{\mathcal{I}}\right\|_{\ell^2\to\ell^1} &\leq 7\sqrt{2\theta p},\\
\left\|\boldsymbol{G}-\boldsymbol{G}'\right\|_{\ell^2\to\ell^1} &\leq 4\sqrt{n}+7\sqrt{\log(2p)},\\
\left\|\overline{\boldsymbol{Y}}_{\mathcal{I}}-\boldsymbol{Y}_{\mathcal{I}}\right\|_{\ell^2\to\ell^1} &\leq 20\sqrt{\frac{n\log p}{\theta}},\\
\left\|\overline{\boldsymbol{Y}}-\boldsymbol{Y}\right\|_{\ell^2\to\ell^1} &\leq 20\sqrt{\frac{n\log p}{\theta}}
\end{aligned}
$$

hold simultaneously with probability at least $1-cp^{-2}$ for a positive constant c.

Proof First of all, we have

$$
\left\|\frac{\boldsymbol{x}_0\boldsymbol{x}_0^\top}{\|\boldsymbol{x}_0\|^2}\boldsymbol{G}\boldsymbol{M}\right\|_{\ell^2\to\ell^1} \leq \frac{1}{\|\boldsymbol{x}_0\|^2}\|\boldsymbol{x}_0\|_{\ell^2\to\ell^1}\left\|\boldsymbol{x}_0^\top\boldsymbol{G}\boldsymbol{M}\right\|_{\ell^2\to\ell^2} = \frac{2}{\|\boldsymbol{x}_0\|^2}\|\boldsymbol{x}_0\|_1\left\|\boldsymbol{x}_0^\top\boldsymbol{G}\right\|,
$$

where in the last inequality we have applied the fact $\|\boldsymbol{M}\| \leq 2$ from Lemma A.16. Now $\boldsymbol{x}_0^\top\boldsymbol{G}$ is an i.i.d. Gaussian vectors with each entry distributed as $\mathcal{N}\left(0, \frac{\|\boldsymbol{x}_0\|^2}{p}\right)$, where $\|\boldsymbol{x}_0\|^2 = \frac{|\mathcal{I}|}{\theta p}$. So by Gaussian concentration inequality in Lemma A.9, we have

$$
\left\|\boldsymbol{x}_0^\top\boldsymbol{G}\right\| \leq 2\|\boldsymbol{x}_0\|\sqrt{\frac{\log(2p)}{p}}
$$

with probability at least $1 - c_1 p^{-2}$. On the intersection with $\mathcal{E}_0$, this implies

$$\left\| \frac{\boldsymbol{x}_0 \boldsymbol{x}_0^\top}{\|\boldsymbol{x}_0\|^2} \boldsymbol{G}\boldsymbol{M} \right\|_{\ell^2 \to \ell^1} \le 2\sqrt{2\theta \log(2p)},$$

with probability at least $1 - c_2 p^{-2}$ provided $\theta > 1/\sqrt{n}$. Moreover, when intersected with $\mathcal{E}_0$, Lemma A.14 implies that when $p \ge C_1 n$,

$$\|\boldsymbol{G}\|_{\ell^2 \to \ell^1} \le \sqrt{p}, \quad \|\boldsymbol{G}_{\mathcal{I}}\|_{\ell^2 \to \ell^1} \le \sqrt{2\theta p}$$

with probability at least $1 - c_3 p^{-2}$ provided $\theta > 1/\sqrt{n}$. Hence, by Lemma A.16, when $p > C_2 n$,

$$\begin{aligned}
\|\boldsymbol{G} - \boldsymbol{G}'\|_{\ell^2 \to \ell^1} &\le \|\boldsymbol{G}\|_{\ell^2 \to \ell^1} \|\boldsymbol{I} - \boldsymbol{M}\| + \left\| \frac{\boldsymbol{x}_0 \boldsymbol{x}_0^\top}{\|\boldsymbol{x}_0\|^2} \boldsymbol{G}\boldsymbol{M} \right\|_{\ell^2 \to \ell^1} \\
&\le \sqrt{p} \left(4\sqrt{\frac{n}{p}} + 4\sqrt{\frac{\log(2p)}{p}} \right) + 2\sqrt{2\theta \log(2p)} \le 4\sqrt{n} + 7\sqrt{\log(2p)}, \\
\left\|\overline{\boldsymbol{Y}}\right\|_{\ell^2 \to \ell^1} &\le \|\boldsymbol{x}_0\|_{\ell^2 \to \ell^1} + \|\boldsymbol{G}\|_{\ell^2 \to \ell^1} \le \|\boldsymbol{x}_0\|_1 + \sqrt{p} \le 2\sqrt{\theta p} + \sqrt{p} \le 3\sqrt{p}, \\
\|\boldsymbol{G}'_{\mathcal{I}}\|_{\ell^2 \to \ell^1} &\le \|\boldsymbol{G}_{\mathcal{I}}\|_{\ell^2 \to \ell^1} \|\boldsymbol{M}\| + \left\| \frac{\boldsymbol{x}_0 \boldsymbol{x}_0^\top}{\|\boldsymbol{x}_0\|^2} \boldsymbol{G}\boldsymbol{M} \right\|_{\ell^2 \to \ell^1} \le 2\sqrt{2\theta p} + 2\sqrt{2\theta \log(2p)} \le 4\sqrt{2\theta p}, \\
\|\boldsymbol{G}_{\mathcal{I}} - \boldsymbol{G}'_{\mathcal{I}}\|_{\ell^2 \to \ell^1} &\le \|\boldsymbol{G}_{\mathcal{I}}\|_{\ell^2 \to \ell^1} \|\boldsymbol{I} - \boldsymbol{M}\| + \left\| \frac{\boldsymbol{x}_0 \boldsymbol{x}_0^\top}{\|\boldsymbol{x}_0\|^2} \boldsymbol{G}\boldsymbol{M} \right\|_{\ell^2 \to \ell^1} \\
&\le \sqrt{2\theta p} \left(4\sqrt{\frac{n}{p}} + 4\sqrt{\frac{\log(2p)}{p}} \right) + 2\sqrt{2\theta \log(2p)} \le 4\sqrt{2\theta n} + 6\sqrt{2\theta \log(2p)}, \\
\|\boldsymbol{Y}_{\mathcal{I}}\|_{\ell^2 \to \ell^1} &\le \left\| \frac{\boldsymbol{x}_0}{\|\boldsymbol{x}_0\|} \right\|_{\ell^2 \to \ell^1} + \|\boldsymbol{G}'_{\mathcal{I}}\|_{\ell^2 \to \ell^1} \le \frac{\|\boldsymbol{x}_0\|_1}{\|\boldsymbol{x}_0\|} + 6\sqrt{2\theta p} \le 7\sqrt{2\theta p}
\end{aligned}$$

with probability at least $1 - c_4 p^{-2}$ provided $\theta > 1/\sqrt{n}$. Finally, by Lemma A.15 and the results above, we obtain

$$\begin{aligned}
\left\|\overline{\boldsymbol{Y}} - \boldsymbol{Y}\right\|_{\ell^2 \to \ell^1} &\le \left| 1 - \frac{1}{\|\boldsymbol{x}_0\|} \right| \|\boldsymbol{x}_0\|_1 + \|\boldsymbol{G} - \boldsymbol{G}'\|_{\ell^2 \to \ell^1} \le 20\sqrt{\frac{n \log p}{\theta}}, \\
\left\|\overline{\boldsymbol{Y}}_{\mathcal{I}} - \boldsymbol{Y}_{\mathcal{I}}\right\|_{\ell^2 \to \ell^1} &\le \left| 1 - \frac{1}{\|\boldsymbol{x}_0\|} \right| \|\boldsymbol{x}_0\|_1 + \|\boldsymbol{G}_{\mathcal{I}} - \boldsymbol{G}'_{\mathcal{I}}\|_{\ell^2 \to \ell^1} \le 20\sqrt{\frac{n \log p}{\theta}},
\end{aligned}$$

holding with probability at least $1 - c_5 p^{-2}$. ∎

Lemma A.18 *Provided $p \ge Cn$ and $\theta > 1/\sqrt{n}$, the following*

$$\|\boldsymbol{G}'\|_{\ell^2 \to \ell^\infty} \le 2\sqrt{\frac{n}{p}} + 8\sqrt{\frac{2\log(2p)}{p}},$$

$$\left\| \|\boldsymbol{G}-\boldsymbol{G}'\|_{\ell^2\to\ell^\infty} \le \frac{4n}{p} + \frac{8\sqrt{2}\log(2p)}{p} + \frac{21\sqrt{n\log(2p)}}{p} \right.$$

hold simultaneously with probability at least $1-cp^{-2}$ *for some constant* $c>0$.

Proof First of all, we have when $p \ge C_1 n$, it holds with probability at least $1-c_2p^{-2}$ that

$$\left\|\frac{\boldsymbol{x}_0\boldsymbol{x}_0^\top}{\|\boldsymbol{x}_0\|^2}\boldsymbol{G}\boldsymbol{M}\right\|_{\ell^2\to\ell^\infty} \le \frac{1}{\|\boldsymbol{x}_0\|^2}\|\boldsymbol{x}_0\|_{\ell^2\to\ell^\infty}\left\|\boldsymbol{x}_0^\top\boldsymbol{G}\boldsymbol{M}\right\|_{\ell^2\to\ell^2} \le \frac{2}{\|\boldsymbol{x}_0\|^2}\|\boldsymbol{x}_0\|_\infty\left\|\boldsymbol{x}_0^\top\boldsymbol{G}\right\|,$$

where at the last inequality we have applied the fact $\|\boldsymbol{M}\| \le 2$ from Lemma A.16. Moreover, from proof of Lemma A.17, we know that $\left\|\boldsymbol{x}_0^\top\boldsymbol{G}\right\| \le 2\sqrt{\log(2p)/p}\,\|\boldsymbol{x}_0\|$ with probability at least $1-c_3p^{-2}$ provided $p \ge C_4 n$. Therefore, conditioned on $\mathcal{E}_0$, we obtain that

$$\left\|\frac{\boldsymbol{x}_0\boldsymbol{x}_0^\top}{\|\boldsymbol{x}_0\|^2}\boldsymbol{G}\boldsymbol{M}\right\|_{\ell^2\to\ell^\infty} \le \frac{4\|\boldsymbol{x}_0\|_\infty}{\|\boldsymbol{x}_0\|}\sqrt{\frac{\log(2p)}{p}} \le \frac{4\sqrt{2\log(2p)}}{\sqrt{\theta}p}$$

holds with probability at least $1-c_5p^{-2}$ provided $\theta > 1/\sqrt{n}$. Now by Corollary A.11, we have that

$$\|\boldsymbol{G}\|_{\ell^2\to\ell^\infty} \le \sqrt{\frac{n}{p}} + 2\sqrt{\frac{2\log(2p)}{p}}$$

with probability at least $1-c_6p^{-2}$. Combining the above estimates and Lemma A.16, we have that with probability at least $1-c_7p^{-2}$

$$\begin{aligned}
\|\boldsymbol{G}'\|_{\ell^2\to\ell^\infty} &\le \|\boldsymbol{G}\boldsymbol{M}\|_{\ell^2\to\ell^\infty} + \left\|\frac{\boldsymbol{x}_0\boldsymbol{x}_0^\top}{\|\boldsymbol{x}_0\|^2}\boldsymbol{G}\boldsymbol{M}\right\|_{\ell^2\to\ell^\infty} \\
&\le \|\boldsymbol{G}\|_{\ell^2\to\ell^\infty}\|\boldsymbol{M}\| + \left\|\frac{\boldsymbol{x}_0\boldsymbol{x}_0^\top}{\|\boldsymbol{x}_0\|^2}\boldsymbol{G}\boldsymbol{M}\right\|_{\ell^2\to\ell^\infty} \\
&\le 2\sqrt{\frac{n}{p}} + 4\sqrt{\frac{2\log(2p)}{p}} + \frac{4\sqrt{2\log(2p)}}{\sqrt{\theta}p} \le 2\sqrt{\frac{n}{p}} + 8\sqrt{\frac{2\log(2p)}{p}},
\end{aligned}$$

where the last simplification is provided that $\theta > 1/\sqrt{n}$ and $p \ge C_8 n$ for a sufficiently large C_8. Similarly,

$$\begin{aligned}
\|\boldsymbol{G}-\boldsymbol{G}'\|_{\ell^2\to\ell^\infty} &\le \|\boldsymbol{G}\|_{\ell^2\to\ell^\infty}\|\boldsymbol{I}-\boldsymbol{M}\| + \left\|\frac{\boldsymbol{x}_0\boldsymbol{x}_0^\top}{\|\boldsymbol{x}_0\|^2}\boldsymbol{G}\boldsymbol{M}\right\|_{\ell^2\to\ell^\infty} \\
&\le \frac{4n}{p} + \frac{8\sqrt{2}\log(2p)}{p} + \frac{(8\sqrt{2}+4)\sqrt{n\log(2p)}}{p} + \frac{4\sqrt{2\log(2p)}}{\sqrt{\theta}p} \\
&\le \frac{4n}{p} + \frac{8\sqrt{2}\log(2p)}{p} + \frac{21\sqrt{n\log(2p)}}{p},
\end{aligned}$$

completing the proof. ∎

Appendix B

Auxillary Results for Convolutional Phase Retrieval

In the appendix, we provide details of proof for some supporting results. Appendix B.1 provided us the very basic tools used throughout the analysis. In Appendix B.2, we provide results of bounding the suprema of chaos processes for random circulant matrices. In Appendix B.3, we provide concentration results for suprema of some dependent random processes via decoupling.

B.1 Elementary Tools and Results

Lemma B.1 *Given a fixed number $\rho > 0$, for any $z,\ z' \in \mathbb{C}$, we have*

$$\left|\exp\left(\mathrm{i}\phi(z'+z)\right) - \exp\left(\mathrm{i}\phi(z')\right)\right| \leq 2\mathbb{1}_{|z|\geq\rho|z'|} + \frac{1}{1-\rho}\left|\Im\left(z/z'\right)\right|. \tag{B.1.1}$$

Proof Please refer to the proof of Lemma 3.2 of [Wal16]. ∎

Lemma B.2 *Let $\rho \in (0,1)$, for any $\boldsymbol{z} \in \mathbb{C}$ with $|z| \leq \rho$, we have*

$$\left|1 - \exp\left(\mathrm{i}\phi(1+z)\right) + \mathrm{i}\Im(z)\right| \leq \frac{2-\rho}{(1-\rho)^2}\left|z\right|^2. \tag{B.1.2}$$

Proof For any $t \in \mathbb{R}^+$, let $g(t) = \sqrt{\left(1+\Re(z)\right)^2 + t^2}$, then

$$g'(t) = \frac{t}{\sqrt{\left(1+\Re(z)\right)^2 + t^2}} \leq \frac{t}{\left|1+\Re(z)\right|}.$$

Hence, for any $z \in \mathbb{C}$ with $|z| \leq \rho$, we have

$$\begin{aligned} ||1+z| - (1+\Re(z))| &= \left|\sqrt{(1+\Re(z))^2 + \Im^2(z)} - (1+\Re(z))\right| \\ &= |g(\Im(z)) - g(0)| \leq \frac{\Im^2(z)}{|1+\Re(z)|} \leq \frac{1}{1-\rho}\Im^2(z). \end{aligned}$$

Let $f(z) = 1 - \exp(\mathrm{i}\phi(1+z))$, then by using the estimates above, we observe

$$\begin{aligned} |f(z) + \mathrm{i}\Im(z)| &= \left|\frac{|1+z| - (1+z)}{|1+z|} + \mathrm{i}\Im(z)\right| \\ &= \frac{1}{|1+z|}\left||1+z| - (1+z) + \mathrm{i}\Im(z)\,|1+z|\right| \\ &\leq \frac{1}{|1+z|}\left(|\Im(z)|\,|1-|1+z|| + ||1+z| - (1+\Re(z))|\right) \\ &\leq \frac{1}{|1+z|}\left(|z|\,|\Im(z)| + \frac{1}{1-\rho}\Im^2(z)\right) \leq \frac{2-\rho}{(1-\rho)^2}\,|z|^2\,. \end{aligned}$$

■

Lemma B.3 (Gaussian Concentration Inequality) *Let $\boldsymbol{w} \in \mathbb{R}^n$ be a standard Gaussian random variable $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, and let $g : \mathbb{R}^n \mapsto \mathbb{R}$ denote an L-Lipschitz function. Then for all $t > 0$,*

$$\mathbb{P}\left(|g(\boldsymbol{w}) - \mathbb{E}\left[g(\boldsymbol{w})\right]| \geq t\right) \leq 2\exp\left(-t^2/(2L^2)\right).$$

Moreover, if $\boldsymbol{w} \in \mathbb{C}^n$ with $\boldsymbol{w} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$, and $g : \mathbb{C}^n \mapsto \mathbb{R}$ is L-Lipschitz, then the inequality above still holds.

Proof The result for real-valued Gaussian random variables is standard, please refer to [BLM13, Chapter 5] for detailed proof. For the complex case, let

$$\boldsymbol{v} = \underbrace{\frac{1}{\sqrt{2}}\begin{bmatrix}\boldsymbol{I} & \mathrm{i}\boldsymbol{I}\end{bmatrix}}_{h}\begin{bmatrix}\boldsymbol{v}_r \\ \boldsymbol{v}_i\end{bmatrix}, \qquad \boldsymbol{v}_r, \boldsymbol{v}_i \sim_{i.i.d.} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}).$$

By composition theorem, we know that $g' \circ h : \mathbb{R}^{2n} \mapsto \mathbb{R}$ is L-Lipschitz. Therefore, by applying the Gaussian concentration inequality for $g' \circ h$ and $\begin{bmatrix}\boldsymbol{v}_r \\ \boldsymbol{v}_i\end{bmatrix}$, we get the desired result. ■

Theorem B.4 (Gaussian tail comparison for vector-valued functions, Theorem 3, [led07]) *Let $\boldsymbol{w} \in \mathbb{R}^n$ be standard Gaussian variable $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, and let $f : \mathbb{R}^n \mapsto \mathbb{R}^\ell$ be an L-Lipschitz function. Then for any*

$t > 0$, we have

$$\mathbb{P}\left(\|f(\boldsymbol{w}) - \mathbb{E}\left[f(\boldsymbol{w})\right]\| \geq t\right) \leq e\mathbb{P}\left(\|\boldsymbol{v}\| \geq \frac{t}{L}\right),$$

where $\boldsymbol{v} \in \mathbb{R}^\ell$ such that $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Moreover, if $\boldsymbol{w} \in \mathbb{C}^n$ with $\boldsymbol{w} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$ and $f : \mathbb{C}^n \mapsto \mathbb{R}^\ell$ is L-Lipschitz, then the inequality above still holds.

The proof is similar to that of Lemma B.3.

Lemma B.5 (sub-Gaussian Random Variables) *Let X be a centered σ^2 sub-Gaussian random variable, such that*

$$\mathbb{P}\left(|X| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right),$$

then for any integer $p \geq 1$, we have

$$\mathbb{E}\left[|X|^p\right] \leq \left(2\sigma^2\right)^{p/2} p\Gamma(p/2).$$

In particular, we have

$$\|X\|_{L^p} = \left(\mathbb{E}\left[|X|^p\right]\right)^{1/p} \leq \sigma e^{1/e}\sqrt{p}, \quad p \geq 2,$$

and $\mathbb{E}\left[|X|\right] \leq \sigma\sqrt{2\pi}$.

Lemma B.6 (Sub-exponential tail bound via moment control) *Suppose X is a centered random variable satisfying*

$$\left(\mathbb{E}\left[|X|^p\right]\right)^{1/p} \leq \alpha_0 + \alpha_1\sqrt{p} + \alpha_2 p, \quad \textit{for all } p \geq p_0$$

for some $\alpha_0, \alpha_1, \alpha_2, p_0 > 0$. Then, for any $u \geq p_0$, we have

$$\mathbb{P}\left(|X| \geq e(\alpha_0 + \alpha_1\sqrt{u} + \alpha_2 u)\right) \leq 2\exp\left(-u\right).$$

This further implies that for any $t > \alpha_1\sqrt{p_0} + \alpha_2 p_0$, we have

$$\mathbb{P}\left(|X| \geq c_1\alpha_0 + t\right) \leq 2\exp\left(-c_2\min\left\{\frac{t^2}{\alpha_1^2}, \frac{t}{\alpha_2}\right\}\right),$$

for some positive constants $c_1,\ c_2 > 0$.

Proof The first inequality directly comes from Proposition 2.6 of [KMR14] via Markov inequality, also see

Proposition 7.11 and Proposition 7.15 of [FR13]. For the second, let $t = \alpha_1\sqrt{u} + \alpha_2 u$, if $\alpha_1\sqrt{u} \leq \alpha_2 u$, then

$$t = \alpha_1\sqrt{u} + \alpha_2 u \leq 2\alpha_2 u \;\Rightarrow\; u \geq \frac{t}{2\alpha_2}.$$

Otherwise, similarly, we have $u \geq t^2/(4\alpha_1^2)$. Combining the two cases above, we get the desired result. ■

Lemma B.7 (Tail bound for heavy-tailed distribution via moment control) *Suppose X is a centered random variable satisfying*

$$\left(\mathbb{E}\left[|X|^p\right]\right)^{1/p} \leq p\left(\alpha_0 + \alpha_1\sqrt{p} + \alpha_2 p\right), \quad \textit{for all } p \geq p_0,$$

for some $\alpha_0, \alpha_1, \alpha_2, p_0 \geq 0$. Then, for any $u \geq p_0$, we have

$$\mathbb{P}\left(|X| \geq eu\left(\alpha_0 + \alpha_1\sqrt{u} + \alpha_2 u\right)\right) \leq 2\exp\left(-u\right).$$

This further implies that for any $t > p_0\left(\alpha_0 + \alpha_1\sqrt{p_0} + \alpha_2 p_0\right)$, we have

$$\mathbb{P}\left(|X| \geq c_1 t\right) \leq 2\exp\left(-c_2 \min\left\{\sqrt{\frac{t}{2(\alpha_1+\alpha_2)}}, \frac{t}{2\alpha_0}\right\}\right),$$

for some positive constant $c_1, c_2 > 0$.

Proof The proof of the first tail bound is similar to that of Lemma B.6 by using Markov inequality. Notice that

$$\mathbb{P}\left(|X| \geq eu\left(\alpha_0 + (\alpha_1+\alpha_2)u\right)\right) \leq \mathbb{P}\left(|X| \geq eu\left(\alpha_0 + \alpha_1\sqrt{u} + \alpha_2 u\right)\right) \leq 2\exp\left(-u\right).$$

Let $t = \alpha_0 u + (\alpha_1+\alpha_2)\,u^2$, if $\alpha_0 u \leq (\alpha_1+\alpha_2)\,u^2$, then

$$t = \alpha_0 u + (\alpha_1+\alpha_2)\,u^2 \leq 2\,(\alpha_1+\alpha_2)\,u^2 \;\Rightarrow\; u \geq \sqrt{\frac{t}{2(\alpha_1+\alpha_2)}}.$$

Otherwise, we have $u \geq t/(2\alpha_0)$. Combining the two cases above, we get the desired result. ■

Definition B.8 ($d_2(\cdot), d_F(\cdot)$ and γ_β functional) *For a given set of matrices $\mathcal{B}$, we define*

$$d_F(\mathcal{B}) \doteq \sup_{\boldsymbol{B}\in\mathcal{B}} \|\boldsymbol{B}\|_F, \quad d_2(\mathcal{B}) \doteq \sup_{\boldsymbol{B}\in\mathcal{B}} \|\boldsymbol{B}\|,$$

For a metric space (T, d), an admissible sequence of T is a collection of subsets of T, $\{T_r : r > 0\}$, such that for

every $s > 1$, $|T_r| \leq 2^{2^r}$ and $|T_0| = 1$. For $\beta \geq 1$, define the γ_β functional by

$$\gamma_\beta(T, d) \doteq \inf \sup_{t \in T} \sum_{r=0}^{\infty} 2^{r/\beta} d(t, T_r),$$

where the infimum is taken with respect to all admissible sequences of T. In particular, for γ_2 functional of the set $\mathcal{B}$ equipped with distance $\|\|$, [Tal14a] shows that

$$\gamma_2(\mathcal{B}, \|\|) \leq c \int_0^{d_2(\mathcal{B})} \log^{1/2} \mathcal{N}(\mathcal{B}, \|\|, \epsilon) d\epsilon, \tag{B.1.3}$$

where $\mathcal{N}(\mathcal{B}, \|\|, \epsilon)$ is the covering number of the set $\mathcal{B}$ with diameter $\epsilon \in (0, 1)$.

Theorem B.9 (Theorem 3.5, [KMR14]) *Let $\sigma_\xi^2 \geq 1$ and $\boldsymbol{\xi} = (\xi_j)_{j=1}^n$, where $\{\xi_j\}_{j=1}^n$ are independent zero-mean, variance one, σ_ξ^2-subgaussian random variables, and let $\mathcal{B}$ be a class of matrices. Let us define a quantity*

$$\mathcal{C}_{\mathcal{B}}(\boldsymbol{\xi}) \doteq \sup_{\boldsymbol{B} \in \mathcal{B}} \left| \|\boldsymbol{B}\boldsymbol{\xi}\|^2 - \mathbb{E}\left[\|\boldsymbol{B}\boldsymbol{\xi}\|^2\right] \right|. \tag{B.1.4}$$

For every $p \geq 1$, we have

$$\left\| \sup_{\boldsymbol{B} \in \mathcal{B}} \|\boldsymbol{B}\boldsymbol{\xi}\| \right\|_{L^p} \leq C_{\sigma_\xi^2} \left[\gamma_2(\mathcal{B}, \|\|) + d_F(\mathcal{B}) + \sqrt{p} d_2(\mathcal{B})\right]$$

$$\left\| \sup_{\boldsymbol{B} \in \mathcal{B}} \left| \|\boldsymbol{B}\boldsymbol{\xi}\|^2 - \mathbb{E}\left[\|\boldsymbol{B}\boldsymbol{\xi}\|^2\right] \right| \right\|_{L^p} \leq C_{\sigma_\xi^2} \left\{ \gamma_2(\mathcal{B}, \|\|) \left[\gamma_2(\mathcal{B}, \|\|) + d_F(\mathcal{B})\right] \right.$$

$$\left. + \sqrt{p} d_2(\mathcal{B}) \left[\gamma_2(\mathcal{B}, \|\|) + d_F(\mathcal{B})\right] + p d_2^2(\mathcal{B}) \right\},$$

where $C_{\sigma_\xi^2}$ is some positive numerical constant only depending on σ_ξ^2, and $d_2(\cdot), d_F(\cdot)$ and $\gamma_2(\mathcal{B}, \|\cdot\|)$ are given in Definition B.8.

The following theorem establishes the *restricted isometry property* (RIP) of the Gaussian random convolution matrix.

Theorem B.10 (Theorem 4.1, [KMR14]) *Let $\boldsymbol{\xi} \in \mathbb{C}^m$ be a random vector with $\xi_i \sim_{i.i.d.} \mathcal{CN}(0, 1)$, and let Ω be a fixed subset of $[m]$ with $|\Omega| = n$. Define a set $\mathcal{E}_s = \{\boldsymbol{v} \in \mathbb{C}^m \mid \|\boldsymbol{v}\|_0 \leq s\}$, and define a matrix*

$$\boldsymbol{\Phi} = \boldsymbol{R}_\Omega \boldsymbol{C}_{\boldsymbol{\xi}}^* \in \mathbb{C}^{n \times m}$$

where $\boldsymbol{R}_\Omega : \mathbb{C}^m \mapsto \mathbb{C}^n$ is an operator that restrict a vector to its entries in Ω. Then for any $s \leq m$, and $\eta, \delta_s \in (0, 1)$ such that

$$n \geq C \delta_s^{-2} s \log^2 s \log^2 m,$$

the partial random circulant matrix $\boldsymbol{\Phi} \in \mathbb{R}^{n\times m}$ satisfies the restricted isometry property

$$(1-\delta_s)\sqrt{n}\,\|\boldsymbol{v}\| \le \|\boldsymbol{\Phi}\boldsymbol{v}\| \le (1+\delta_s)\sqrt{n}\,\|\boldsymbol{v}\| \tag{B.1.5}$$

for all $\boldsymbol{v} \in \mathcal{E}_s$, with probability at least $1 - m^{-\log^2 s \log m}$.

Lemma B.11 *Let the random vector $\boldsymbol{\xi} \in \mathbb{C}^m$ and the random matrix $\boldsymbol{\Phi} \in \mathbb{C}^{n\times m}$ be defined the same as Theorem B.10, and let $\mathcal{E}_s = \{\boldsymbol{v} \in \mathbb{C}^m \mid \|\boldsymbol{v}\|_0 \le s\}$ for some positive integer $s \le n$. For any positive scalar $\delta > 0$ and any positive integer $s \le n$, whenever $m \ge C\delta^{-2} n \log^4 n$, we have*

$$\|\boldsymbol{\Phi}\boldsymbol{v}\| \le \delta\sqrt{m}\,\|\boldsymbol{v}\|,$$

for all $\boldsymbol{v} \in \mathcal{E}_s$, with probability at least $1 - m^{-c\log^2 s}$. Here c, C are some positive numerical constants.

Proof The proof follows from the results in [KMR14]. Without loss of generality, we assume $\|\boldsymbol{v}\| = 1$. Let us define sets

$$\mathcal{D}_{s,m} \doteq \{\boldsymbol{v} \in \mathbb{C}^m : \ \|\boldsymbol{v}\| = 1,\ \|\boldsymbol{v}\|_0 \le s\},$$
$$\mathcal{V} \doteq \left\{ \frac{1}{\sqrt{n}} \boldsymbol{R}_{[1:n]} \boldsymbol{F}_m^{-1} \operatorname{diag}(\boldsymbol{F}_m \boldsymbol{v})\, \boldsymbol{F}_m \mid \boldsymbol{v} \in \mathcal{D}_{s,m} \right\},$$

Section 4 of [KMR14] shows that

$$\sup_{\boldsymbol{v}\in\mathcal{D}_{s,m}} \left| \frac{1}{n}\|\boldsymbol{\Phi}\boldsymbol{v}\|^2 - 1 \right| = \sup_{\boldsymbol{V}_{\boldsymbol{v}}\in\mathcal{V}} \left| \|\boldsymbol{V}_{\boldsymbol{v}}\boldsymbol{\xi}\|^2 - \mathbb{E}_{\boldsymbol{\xi}}\left[\|\boldsymbol{V}_{\boldsymbol{v}}\boldsymbol{\xi}\|^2\right] \right| = \mathcal{C}_{\mathcal{V}}(\boldsymbol{\xi}),$$

where $\mathcal{C}_{\mathcal{V}}(\boldsymbol{\xi})$ is defined in (B.1.4). Theorem 4.1 and Lemma 4.2 of [KMR14] implies that

$$d_F(\mathcal{V}) = 1, \quad d_2(\mathcal{V}) \le \sqrt{\frac{s}{n}}, \quad \gamma_2(\mathcal{V}, \|\|) \le c\sqrt{\frac{s}{n}} \log s \log m,$$

for some constant $c > 0$. By using the estimates above, Theorem 3.1 of [KMR14] further implies for any $t > 0$

$$\mathbb{P}\left(\mathcal{C}_{\mathcal{V}}(\boldsymbol{\xi}) \ge c_1\sqrt{\frac{s}{n}} \log^2 s \log^2 m + t \right) \le 2\exp\left(-c_2 \min\left\{ \frac{nt^2}{s\log^2 s \log^2 m}, \frac{nt}{s} \right\} \right).$$

For any positive constant $\delta > 0$, choosing $t = \delta^2 m/n$, whenever $m \ge C\delta^{-2} n \log^2 s \log^2 m$ for some constant $C > 0$ large enough, we have

$$\sup_{\boldsymbol{v}\in\mathcal{D}_{s,m}} \left| \frac{1}{n}\|\boldsymbol{\Phi}\boldsymbol{v}\|^2 - 1 \right| \le c_1\sqrt{\frac{s}{n}} \log^2 s \log^2 m + \delta\frac{m}{n} \le 2\delta^2 \frac{m}{n},$$

with probability at least $1 - m^{-c_3 \log^2 s}$ for some positive constant $c_3 > 0$. Therefore, we have

$$\|\boldsymbol{\Phi} \boldsymbol{v}\| \leq \sqrt{n + 2\delta^2 m} \leq C' \delta \sqrt{m},$$

holds for any $\boldsymbol{v} \in \mathcal{D}_{s,m}$ with high probability, where $C' > 0$ is a numerical constant. ■

B.2 Moments and Spectral Norm of Partial Random Circulant Matrix

Let $\boldsymbol{g} \in \mathbb{C}^m$ be a random complex Gaussian vector with $\boldsymbol{g} \sim \mathcal{CN}(\boldsymbol{0}, \sigma_g^2 \boldsymbol{I})$. Given a partial random circulant matrix $\boldsymbol{C_g} \boldsymbol{R}_{[1:n]}^\top \in \mathbb{C}^{m \times n}$ $(m \geq n)$, we control the moments and the tail bound of the terms in the following form

$$\begin{aligned} \boldsymbol{T}_1(\boldsymbol{g}) &= \frac{1}{m} \boldsymbol{R}_{[1:n]} \boldsymbol{C}_{\boldsymbol{g}}^* \operatorname{diag}(\boldsymbol{b}) \boldsymbol{C_g} \boldsymbol{R}_{[1:n]}^\top, \\ \boldsymbol{T}_2(\boldsymbol{g}) &= \frac{1}{m} \boldsymbol{R}_{[1:n]} \boldsymbol{C}_{\boldsymbol{g}}^\top \operatorname{diag}\left(\widetilde{\boldsymbol{b}}\right) \boldsymbol{C_g} \boldsymbol{R}_{[1:n]}^\top, \end{aligned}$$

where $\boldsymbol{b} \in \mathbb{R}^m$, and $\widetilde{\boldsymbol{b}} \in \mathbb{C}^m$. The concentration of these quantities plays an important role in our arguments, and the proof mimics the arguments in [Rau10, KMR14]. Prior to that, let us define sets

$$\mathcal{D} = \left\{ \boldsymbol{v} \in \mathbb{CS}^{m-1} : \ \operatorname{supp}(\boldsymbol{v}) \in [n] \right\}, \tag{B.2.1}$$

$$\mathcal{V}(\boldsymbol{d}) = \left\{ \boldsymbol{V_v} : \ \boldsymbol{V_v} = \frac{1}{\sqrt{m}} \operatorname{diag}(\boldsymbol{d})^{1/2} \boldsymbol{F}_m^{-1} \operatorname{diag}(\boldsymbol{F}_m \boldsymbol{v}) \boldsymbol{F}_m, \ \boldsymbol{v} \in \mathcal{D} \right\}. \tag{B.2.2}$$

B.2.1 Controlling the Moments and Tail of $\boldsymbol{T}_1(\boldsymbol{g})$

Theorem B.12 *Let $\boldsymbol{g} \in \mathbb{C}^m$ be a random complex Gaussian vector with $\boldsymbol{g} \sim \mathcal{CN}(\boldsymbol{0}, \sigma_g^2 \boldsymbol{I})$ and any fixed vector $\boldsymbol{b} = [b_1, \cdots, b_m]^\top \in \mathbb{R}^m$. Given a partial random circulant matrix $\boldsymbol{C_g} \boldsymbol{R}_{[1:n]}^\top \in \mathbb{C}^{m \times n}$ $(m \geq n)$, let us define*

$$\mathcal{L}(\boldsymbol{g}) \doteq \left\| \frac{1}{m} \boldsymbol{R}_{[1:n]} \boldsymbol{C}_{\boldsymbol{g}}^* \operatorname{diag}(\boldsymbol{b}) \boldsymbol{C_g} \boldsymbol{R}_{[1:n]}^\top - \frac{1}{m} \left(\sum_{k=1}^m b_k \right) \boldsymbol{I} \right\|.$$

Then for any integer $p \geq 1$, we have

$$\|\mathcal{L}(\boldsymbol{g})\|_{L^p} \leq C_{\sigma_g^2} \|\boldsymbol{b}\|_\infty \left(\sqrt{\frac{n}{m}} \log^{3/2} n \log^{1/2} m + \sqrt{p} \sqrt{\frac{n}{m}} + p \frac{n}{m} \right).$$

In addition, For any $\delta > 0$, whenever $m \geq C'_{\sigma_g^2} \delta^{-2} \|\boldsymbol{b}\|_\infty^2 n \log^4 n$, we have

$$\mathcal{L}(\boldsymbol{g}) \leq \delta \tag{B.2.3}$$

holds with probability at least $1-2m^{-c_{\sigma_g^2}\log^3 n}$. *Here,* $c_{\sigma_g^2},\ C_{\sigma_g^2}$, *and* $C'_{\sigma_g^2}$ *are some numerical constants only depending on* σ_g^2.

Proof Without loss of generality, let us assume that $\sigma_g^2=1$. Let us first consider the case $\boldsymbol{b}\geq\boldsymbol{0}$, and let $\boldsymbol{\Lambda}=\mathrm{diag}\,(\boldsymbol{b})$, then

$$\begin{aligned}\mathcal{L}(\boldsymbol{g}) &= \sup_{\boldsymbol{w}\in\mathbb{CS}^{n-1}}\left|\frac{1}{m}\boldsymbol{w}^*\boldsymbol{R}_{[1:n]}\boldsymbol{C}_{\boldsymbol{g}}^*\boldsymbol{\Lambda}\boldsymbol{C}_{\boldsymbol{g}}\boldsymbol{R}_{[1:n]}^\top\boldsymbol{w}-\frac{1}{m}\sum_{k=1}^m b_k\right| \\ &= \sup_{\boldsymbol{v}\in\mathbb{CS}^{m-1},\mathrm{supp}(\boldsymbol{v})\in[n]}\left|\frac{1}{m}\boldsymbol{v}^*\boldsymbol{C}_{\boldsymbol{g}}^*\boldsymbol{\Lambda}\boldsymbol{C}_{\boldsymbol{g}}\boldsymbol{v}-\frac{1}{m}\sum_{k=1}^m b_k\right|.\end{aligned}$$

By the convolution theorem, we know that

$$\frac{1}{\sqrt{m}}\boldsymbol{\Lambda}^{1/2}\boldsymbol{C}_{\boldsymbol{g}}\boldsymbol{v}=\frac{1}{\sqrt{m}}\boldsymbol{\Lambda}^{1/2}\left(\boldsymbol{g}\circledast\boldsymbol{v}\right)=\frac{1}{\sqrt{m}}\boldsymbol{\Lambda}^{1/2}\boldsymbol{F}_m^{-1}\,\mathrm{diag}\left(\boldsymbol{F}_m\boldsymbol{v}\right)\boldsymbol{F}_m\boldsymbol{g}=\boldsymbol{V}_{\boldsymbol{v}}\boldsymbol{g}.$$

Since $\mathbb{E}\left[\boldsymbol{R}_{[1:n]}\boldsymbol{C}_{\boldsymbol{g}}^*\boldsymbol{\Lambda}\boldsymbol{C}_{\boldsymbol{g}}\boldsymbol{R}_{[1:n]}^\top\right]=\left(\sum_{k=1}^m b_k\right)\boldsymbol{I}$, we observe

$$\mathcal{L}(\boldsymbol{g})=\sup_{\boldsymbol{V}_{\boldsymbol{v}}\in\mathcal{V}}\left|\left\|\boldsymbol{V}_{\boldsymbol{v}}\boldsymbol{g}\right\|^2-\mathbb{E}\left[\left\|\boldsymbol{V}_{\boldsymbol{v}}\boldsymbol{g}\right\|^2\right]\right|,$$

where the set $\mathcal{V}(\boldsymbol{b})$ is defined in (B.2.2). Next, we invoke Theorem B.9 to control all the moments of $\mathcal{L}(\boldsymbol{a})$, where we need to control the quantities $d_2(\cdot)$, $d_F(\cdot)$ and $\gamma_2(\cdot,\|\|)$ defined in Lemma B.8 for the set $\mathcal{V}$. By Lemma B.18 and Lemma B.19, we know that

$$d_F(\mathcal{V})\leq\|\boldsymbol{b}\|_\infty^{1/2},\quad d_2(\mathcal{V})\leq\sqrt{\frac{n}{m}}\|\boldsymbol{b}\|_\infty^{1/2}, \tag{B.2.4}$$

$$\gamma_2(\mathcal{V},\|\cdot\|)\leq C_0\sqrt{\frac{n}{m}}\|\boldsymbol{b}\|_\infty^{1/2}\log^{3/2}n\log^{1/2}m, \tag{B.2.5}$$

for some constant $C_0>0$. Thus, combining the results in (B.2.4) and (B.2.5), whenever $m\geq C_1 n\log^3 n\log m$ for some constant $C_1>0$, Theorem B.9 implies that

$$\begin{aligned}\left\|\mathcal{L}(\boldsymbol{g})\right\|_{L^p} &\leq C_2\left\{\gamma_2(\mathcal{V},\|\|)\left[\gamma_2\left(\mathcal{V},\|\|\right)+d_F(\mathcal{V})\right]+\sqrt{p}d_2(\mathcal{V})\left[\gamma_2\left(\mathcal{V},\|\|\right)+d_F(\mathcal{V})\right]+pd_2^2(\mathcal{V})\right\} \\ &\leq C_3\|\boldsymbol{b}\|_\infty\left(\sqrt{\frac{n}{m}}\log^{3/2}n\log^{1/2}m+\sqrt{\frac{n}{m}}\sqrt{p}+\frac{n}{m}p\right)\end{aligned}$$

holds for some constants $C_2,\ C_3>0$. Based on the moments estimate of $\mathcal{L}(\boldsymbol{g})$, Lemma B.6 further implies that

$$\mathbb{P}\left(\mathcal{L}(\boldsymbol{g})\geq C_4\sqrt{\frac{n}{m}}\|\boldsymbol{b}\|_\infty\log^{3/2}n\log^{1/2}m+t\right)\leq 2\exp\left(-C_5\frac{m}{n}\|\boldsymbol{b}\|_\infty^{-1}\min\left\{\frac{t^2}{\|\boldsymbol{b}\|_\infty},t\right\}\right),$$

for some constants $C_4,\ C_5>0$. Thus, for any $\delta>0$, whenever $m\geq C_6\delta^{-2}\|\boldsymbol{b}\|_\infty^2 n\log^3 n\log m$ for some

constant $C_6 > 0$, we have

$$\mathcal{L}(\boldsymbol{g}) \le \delta$$

holds with probability at least $1 - 2m^{-C_7 \log^3 n}$.

Now when $\boldsymbol{b}$ is not nonnegative, let $\boldsymbol{b} = \boldsymbol{b}_+ - \boldsymbol{b}_-$, where $\boldsymbol{b}_+ = \left[b_1^+, \cdots, b_m^+\right]^\top$, $\boldsymbol{b}_- = \left[b_1^-, \cdots, b_m^-\right]^\top \in \mathbb{R}_+^m$ are the nonnegative and nonpositive part of $\boldsymbol{b}$, respectively. Let $\boldsymbol{\Lambda} = \operatorname{diag}(\boldsymbol{b})$, $\boldsymbol{\Lambda}_+ = \operatorname{diag}(\boldsymbol{b}_+)$ and $\boldsymbol{\Lambda}_- = \operatorname{diag}(\boldsymbol{b}_-)$, we have

$$\begin{aligned}
\mathcal{L}(\boldsymbol{g}) &= \sup_{\boldsymbol{w}\in\mathbb{CS}^{n-1}} \left| \frac{1}{m} \boldsymbol{w}^* \boldsymbol{R}_{[1:n]} \boldsymbol{C}_{\boldsymbol{g}}^* \boldsymbol{\Lambda} \boldsymbol{C}_{\boldsymbol{g}} \boldsymbol{R}_{[1:n]}^\top \boldsymbol{w} - \frac{1}{m} \sum_{k=1}^m b_k \right| \\
&\le \underbrace{\sup_{\boldsymbol{w}\in\mathbb{CS}^{n-1}} \left| \frac{1}{m} \boldsymbol{w}^* \boldsymbol{R}_{[1:n]} \boldsymbol{C}_{\boldsymbol{g}}^* \boldsymbol{\Lambda}_+ \boldsymbol{C}_{\boldsymbol{g}} \boldsymbol{R}_{[1:n]}^\top \boldsymbol{w} - \frac{1}{m} \sum_{k=1}^m b_k^+ \right|}_{\mathcal{L}_+(\boldsymbol{g})} + \underbrace{\sup_{\boldsymbol{w}\in\mathbb{CS}^{n-1}} \left| \frac{1}{m} \boldsymbol{w}^* \boldsymbol{R}_{[1:n]} \boldsymbol{C}_{\boldsymbol{g}}^* \boldsymbol{\Lambda}_- \boldsymbol{C}_{\boldsymbol{g}} \boldsymbol{R}_{[1:n]}^\top \boldsymbol{w} - \frac{1}{m} \sum_{k=1}^m b_k^- \right|}_{\mathcal{L}_-(\boldsymbol{g})}.
\end{aligned}$$

Now since $\boldsymbol{b}_+, \boldsymbol{b}_- \in \mathbb{R}_+^m$, we can apply the results above for $\mathcal{L}_+(\boldsymbol{g})$ and $\mathcal{L}_-(\boldsymbol{g})$, respectively. Then by Minkowski's inequality, we have

$$\|\mathcal{L}(\boldsymbol{g})\|_{L^p} \le \|\mathcal{L}_+(\boldsymbol{g})\|_{L^p} + \left\|\mathcal{L}_-(\boldsymbol{g})\right\|_{L^p} \le C_6 \|\boldsymbol{b}\|_\infty \left(\sqrt{\frac{n}{m}} \log^{3/2} n \log^{1/2} m + \sqrt{\frac{n}{m}} \sqrt{p} + \frac{n}{m} p \right)$$

for some constant $C_6 > 0$. The tail bound can be similarly derived from the moments bound. This completes the proof. ∎

The result above also implies the following result.

Corollary B.13 *Let $\boldsymbol{g} \in \mathbb{C}^m$ be a random complex Gaussian vector with $\boldsymbol{g} \sim \mathcal{CN}(\boldsymbol{0}, \sigma_g^2 \boldsymbol{I})$, and let $\boldsymbol{G} = \boldsymbol{R}_{[1:n]} \boldsymbol{C}_{\boldsymbol{g}}^* \in \mathbb{C}^{n\times m}$ $(n \le m)$. Then for any integer $p \ge 1$, we have*

$$\left(\mathbb{E}\left[\|\boldsymbol{G}\|^p\right]\right)^{1/p} \le C_{\sigma_g^2} \sqrt{m} \left(1 + \sqrt{\frac{n}{m}} \log^{3/2} n \log^{1/2} m + \sqrt{\frac{n}{m}} \sqrt{p} \right)$$

Moreover, for any $\epsilon \in (0,1)$, whenever $m \ge C\delta^{-2} n \log^4 n$ for some constant $C > 0$, we have

$$(1-\delta) m \|\boldsymbol{w}\|^2 \le \|\boldsymbol{G}^* \boldsymbol{w}\|^2 \le (1+\delta) m \|\boldsymbol{w}\|^2$$

holds for $\boldsymbol{w} \in \mathbb{C}^n$ with probability at least $1 - 2m^{-c_{\sigma_g^2} \log^3 n}$. Here $c_{\sigma_g^2}, C_{\sigma_g^2} > 0$ are some constants depending only on σ_g^2.

Proof Firstly, notice that

$$\|\boldsymbol{G}\| = \sup_{\boldsymbol{w}\in\mathbb{CS}^{n-1},\boldsymbol{r}\in\mathbb{CS}^{m-1}} |\langle \boldsymbol{w}, \boldsymbol{G}\boldsymbol{r}\rangle| \le \sup_{\boldsymbol{w}\in\mathbb{CS}^{n-1}} \|\boldsymbol{G}^*\boldsymbol{w}\| = \sup_{\boldsymbol{w}\in\mathbb{CS}^{n-1}} \left\|\boldsymbol{C}_{\boldsymbol{g}}\boldsymbol{R}_{[1:n]}^\top \boldsymbol{w}\right\|$$
$$= \sup_{\boldsymbol{v}\in\mathbb{CS}^{m-1},\mathrm{supp}(\boldsymbol{v})\in[n]} \|\boldsymbol{C}_{\boldsymbol{g}}\boldsymbol{v}\|.$$

Thus, similar to the argument of Theorem B.12, let the set $\mathcal{D}$ and $\mathcal{V}(\mathbf{1})$ define as (B.2.1) and (B.2.2), we have

$$\frac{1}{\sqrt{m}}\|\boldsymbol{G}\| \le \sup_{\boldsymbol{V}_{\boldsymbol{v}}\in\mathcal{V}} \|\boldsymbol{V}_{\boldsymbol{v}}\boldsymbol{g}\|.$$

By Lemma B.18 and Lemma B.19, we know that

$$d_F(\mathcal{V}) \le 1, \quad d_2(\mathcal{V}) \le \sqrt{\frac{n}{m}}, \quad \gamma_2(\mathcal{V}, \|\cdot\|) \le C_0\sqrt{\frac{n}{m}}\log^{3/2} n \log^{1/2} m.$$

Thus, using Theorem B.9, we obtain

$$\mathbb{E}\left[\left|\sup_{\boldsymbol{V}_{\boldsymbol{v}}\in\mathcal{V}} \|\boldsymbol{V}_{\boldsymbol{v}}\boldsymbol{g}\|\right|^p\right]^{1/p} \le C_{\sigma_g^2}\left(\sqrt{\frac{n}{m}}\log^{3/2} n\log^{1/2} m + 1 + \sqrt{\frac{n}{m}}\sqrt{p}\right),$$

where $C_{\sigma_g^2} > 0$ is constant depending only on σ_g^2. The concentration inequality can be directly derived from Theorem B.12, noticing that for any $\delta > 0$, whenever $m \ge C_1\delta^{-2}n\log^4 n$ for some positive constant $C_1 > 0$, we have

$$\sup_{\boldsymbol{w}\in\mathbb{CS}^{n-1}} \left|\frac{1}{m}(\boldsymbol{w}^*\boldsymbol{G}\boldsymbol{G}^*\boldsymbol{w} - 1)\right| \le \delta \implies (1-\delta)m \le \sup_{\boldsymbol{w}\in\mathbb{CS}^{n-1}} \|\boldsymbol{G}^*\boldsymbol{w}\|^2 \le (1+\delta)m$$

holds with probability at least $1 - 2m^{-c_{\sigma_g^2}\log^3 n}$, where $c_{\sigma_g^2} > 0$ is some constant depending only on σ_g^2.

■

B.2.2 Controlling the Moments of $T_2(g)$

Theorem B.14 *Let $\boldsymbol{g} \in \mathbb{C}^m$ are a complex random Gaussian variable with $\boldsymbol{g} \sim \mathcal{CN}(\mathbf{0}, \sigma_g^2\boldsymbol{I})$, and let*

$$\mathcal{N}(\boldsymbol{g}) \doteq \sup_{\boldsymbol{w}\in\mathbb{CS}^{n-1}} \left|\frac{1}{m}\boldsymbol{w}^\top \boldsymbol{R}_{[1:n]}\boldsymbol{C}_{\boldsymbol{g}}^\top \mathrm{diag}\left(\widetilde{\boldsymbol{b}}\right)\boldsymbol{C}_{\boldsymbol{g}}\boldsymbol{R}_{[1:n]}^\top \boldsymbol{w}\right|,$$

where $\widetilde{\boldsymbol{b}} \in \mathbb{C}^m$. Then whenever $m \ge Cn\log^4 n$ for some positive constant $C > 0$, for any positive integer $p \ge 1$, we have

$$\|\mathcal{N}(\boldsymbol{g})\|_{L^p} \le C_{\sigma_g^2}\left\|\widetilde{\boldsymbol{b}}\right\|_\infty \left(\sqrt{\frac{n}{m}}\log^{3/2} n\log^{1/2} m + \sqrt{\frac{n}{m}}\sqrt{p} + \frac{n}{m}p\right),$$

where $C_{\sigma_g^2}$ is positive constant only depending on σ_g^2.

Proof Let $\widetilde{\boldsymbol{\Lambda}} = \operatorname{diag}\left(\widetilde{\boldsymbol{b}}\right)$, similar to the arguments of Theorem B.12, we have

$$\mathcal{N}(\boldsymbol{g}) = \sup_{\boldsymbol{V_v}\in\mathcal{V}} \left|\left\langle \boldsymbol{V_v g}, \overline{\boldsymbol{V_v g}}\right\rangle\right|, \tag{B.2.6}$$

where $\mathcal{V}\left(\widetilde{\boldsymbol{b}}\right)$ is defined as (B.2.2). Let $\boldsymbol{g}'$ be an independent copy of $\boldsymbol{g}$, by Lemma B.15, for any integer $p \geq 1$ we have

$$\|\mathcal{N}(\boldsymbol{g})\|_{L^p} \leq 4 \left\| \sup_{\boldsymbol{V_v}\in\mathcal{V}} \left|\left\langle \boldsymbol{V_v g}, \overline{\boldsymbol{V_v g}'}\right\rangle\right| \right\|_{L^p}.$$

By Lemma B.16 and Lemma B.17, we know that

$$\begin{aligned}
&\left\| \sup_{\boldsymbol{V_v}\in\mathcal{V}} \left|\left\langle \boldsymbol{V_v g}, \overline{\boldsymbol{V_v g}'}\right\rangle\right| \right\|_{L^p} \\
&\leq C_{\sigma_g^2}\left[\gamma_2\left(\mathcal{V}, \|\|\right) \left\| \sup_{\boldsymbol{V_v}\in\mathcal{V}} \|\boldsymbol{V_v g}'\| \right\|_{L^p} + \sup_{\boldsymbol{V_v}\in\mathcal{V}} \left\|\left\langle \boldsymbol{V_v g}, \overline{\boldsymbol{V_v g}'}\right\rangle\right\|_{L^p}\right] \\
&\leq C'_{\sigma_g^2}\left[\gamma_2(\mathcal{V}, \|\|)\left(\gamma_2(\mathcal{V}, \|\|) + d_F(\mathcal{V})\right) + \sqrt{p} d_2(\mathcal{V})\left(d_F(\mathcal{V}) + \gamma_2(\mathcal{V}, \|\|)\right) + p d_2^2(\mathcal{V})\right].
\end{aligned}$$

By Lemma B.18 and Lemma B.19, we know that

$$d_F(\mathcal{V}) \leq \left\|\widetilde{\boldsymbol{b}}\right\|_\infty^{1/2}, \; d_2(\mathcal{V}) \leq \sqrt{\frac{n}{m}}\left\|\widetilde{\boldsymbol{b}}\right\|_\infty^{1/2}, \; \gamma_2(\mathcal{V}, \|\|) \leq C\sqrt{\frac{n}{m}}\left\|\widetilde{\boldsymbol{b}}\right\|_\infty^{1/2} \log^{3/2} n \log^{1/2} m,$$

where $C > 0$ is constant. Thus, combining the results above, we have

$$\|\mathcal{N}(\boldsymbol{g})\|_{L^p} \leq C''_{\sigma_g^2}\left(\sqrt{\frac{n}{m}}\left\|\widetilde{\boldsymbol{b}}\right\|_\infty \log^{3/2} n \log^{1/2} m + \sqrt{p}\sqrt{\frac{n}{m}}\left\|\widetilde{\boldsymbol{b}}\right\|_\infty + p\frac{n}{m}\left\|\widetilde{\boldsymbol{b}}\right\|_\infty\right),$$

where $C''_{\sigma_g^2} > 0$ is some constant depending on σ_g^2. ∎

Lemma B.15 *Let $\mathcal{N}(\boldsymbol{g})$ be defined as (B.2.6), and let $\boldsymbol{g}'$ be an independent copy of $\boldsymbol{g}$, then we have*

$$\|\mathcal{N}(\boldsymbol{g})\|_{L^p} \leq 4 \left\| \sup_{\boldsymbol{V_v}\in\mathcal{V}} \left|\left\langle \boldsymbol{V_v g}, \overline{\boldsymbol{V_v g}'}\right\rangle\right| \right\|_{L^p}.$$

Proof Let $\boldsymbol{\delta} \sim \mathcal{CN}(\boldsymbol{0}, \sigma_g^2 \boldsymbol{I})$ which is independent of $\boldsymbol{g}$, and let

$$\boldsymbol{g}^1 = \boldsymbol{g} + \boldsymbol{\delta}, \qquad \boldsymbol{g}^2 = \boldsymbol{g} - \boldsymbol{\delta},$$

so that $\boldsymbol{g}^1$ and $\boldsymbol{g}^2$ are also independent with $\boldsymbol{g}^1, \boldsymbol{g}^2 \sim \mathcal{CN}(\boldsymbol{0}, 2\sigma_g^2 \boldsymbol{I})$. Let $\mathcal{Q}_{dec}^{\mathcal{N}}(\boldsymbol{g}^1, \boldsymbol{g}^2) = \left\langle \boldsymbol{V_v g}^1, \overline{\boldsymbol{V_v g}^2}\right\rangle$, then

we have

$$\mathbb{E}_{\boldsymbol{\delta}}\left[\mathcal{Q}_{dec}^{\mathcal{N}}(\boldsymbol{g}^1,\boldsymbol{g}^2)\right]=\left\langle \boldsymbol{V_v g},\overline{\boldsymbol{V_v g}}\right\rangle.$$

Therefore, by Jensen's inequality, we have

$$\begin{aligned}\|\mathcal{N}(\boldsymbol{g})\|_{L^p}=\left(\mathbb{E}_{\boldsymbol{g}}\left[\left(\sup_{\boldsymbol{V_v}\in\mathcal{V}}\left|\mathbb{E}_{\boldsymbol{\delta}}\left[\mathcal{Q}_{dec}^{\mathcal{N}}(\boldsymbol{g}^1,\boldsymbol{g}^2)\right]\right|\right)^p\right]\right)^{1/p}&\le\left(\mathbb{E}_{\boldsymbol{g}^1,\boldsymbol{g}^2}\left[\left(\sup_{\boldsymbol{V_v}\in\mathcal{V}}\left|\mathcal{Q}_{dec}^{\mathcal{N}}(\boldsymbol{g}^1,\boldsymbol{g}^2)\right|\right)^p\right]\right)^{1/p}\\&=4\left\|\sup_{\boldsymbol{V_v}\in\mathcal{V}}\left|\left\langle \boldsymbol{V_v g},\overline{\boldsymbol{V_v g'}}\right\rangle\right|\right\|_{L^p},\end{aligned}$$

as desired. ■

Lemma B.16 *Let $\boldsymbol{g}'$ be an independent copy of $\boldsymbol{g}$, for every integer $p\ge 1$, we have*

$$\left\|\sup\left\langle \boldsymbol{V_v g},\overline{\boldsymbol{V_v g'}}\right\rangle\right\|_{L^p}\le C_{\sigma_g^2}\left[\gamma_2\left(\mathcal{V},\|\|\right)\left\|\sup_{\boldsymbol{V_v}\in\mathcal{V}}\|\boldsymbol{V_v g'}\|\right\|_{L^p}+\sup_{\boldsymbol{V_v}\in\mathcal{V}}\left\|\left\langle \boldsymbol{V_v g},\overline{\boldsymbol{V_v g'}}\right\rangle\right\|_{L^p}\right],$$

where $C_{\sigma_g^2}>0$ is a constant depending only on σ_g^2.

Proof The proof is similar to the proof of Lemma 3.2 of [KMR14], and is omitted here. ■

Lemma B.17 *Let $\boldsymbol{g}'$ be an independent copy of $\boldsymbol{g}$, for every integer $p\ge 1$, we have*

$$\begin{aligned}\left\|\sup_{\boldsymbol{V_v}\in\mathcal{V}}\|\boldsymbol{V_v g'}\|\right\|_{L^p}&\le C_{\sigma_g^2}\left[\gamma_2(\mathcal{V},\|\|)+d_F(\mathcal{V})+\sqrt{p}d_2(\mathcal{V})\right]\\\sup_{\boldsymbol{V_v}\in\mathcal{V}}\left\|\left\langle \boldsymbol{V_v g},\overline{\boldsymbol{V_v g'}}\right\rangle\right\|_{L^p}&\le C_{\sigma_g^2}\left[\sqrt{p}d_F(\mathcal{V})d_2(\mathcal{V})+pd_2^2(\mathcal{V})\right],\end{aligned}$$

where $C_{\sigma_g^2}>0$ is a constant depending only on σ_g^2.

Proof The proof is similar to the proofs of Theorem 3.5 and Lemma 3.6 of [KMR14], and is omitted here. ■

B.2.3 Auxiliary Results

The following are the auxiliary results required in the main proof.

Lemma B.18 *Let the sets $\mathcal{D}$, $\mathcal{V}(\boldsymbol{d})$ be defined as (B.2.1) and (B.2.2), we have*

$$d_F(\mathcal{V})\le\|\boldsymbol{d}\|_\infty^{1/2},\qquad d_2(\mathcal{V})\le\sqrt{\frac{n}{m}}\|\boldsymbol{d}\|_\infty^{1/2}.$$

Proof Since each row of $\boldsymbol{V}_{\boldsymbol{v}} \in \mathcal{V}$ consists of weighted shifted copies of $\boldsymbol{v}$, the ℓ_2-norm of each nonzero row of $\boldsymbol{V}_{\boldsymbol{v}}$ is $m^{-1/2}\left|d_k\right|^{1/2}\|\boldsymbol{v}\|$. Thus, we have

$$d_F(\mathcal{V}) = \sup_{\boldsymbol{V}_{\boldsymbol{v}} \in \mathcal{V}} \|\boldsymbol{V}_{\boldsymbol{v}}\|_F \leq \|\boldsymbol{d}\|_\infty^{1/2} \sup_{\boldsymbol{v} \in \mathcal{D}} \|\boldsymbol{v}\| = \|\boldsymbol{d}\|_\infty^{1/2}.$$

Also, for every $\boldsymbol{v} \in \mathcal{D}$, we observe

$$\|\boldsymbol{V}_{\boldsymbol{v}}\| \leq \frac{1}{\sqrt{m}} \|\boldsymbol{d}\|_\infty^{1/2} \|\operatorname{diag}(\boldsymbol{F}_m \boldsymbol{v})\| = \frac{1}{\sqrt{m}} \|\boldsymbol{d}\|_\infty^{1/2} \|\boldsymbol{F}_m \boldsymbol{v}\|_\infty.$$

It is obvious for any $\boldsymbol{v} \in \mathcal{D}$ that $\|\boldsymbol{F}_m \boldsymbol{v}\|_\infty \leq \|\boldsymbol{v}\|_1 \leq \sqrt{n}\|\boldsymbol{v}\|_2 = \sqrt{n}$, so that

$$d_2(\mathcal{V}) = \sup_{\boldsymbol{v} \in \mathcal{D}} \|\boldsymbol{V}_{\boldsymbol{v}}\| \leq \sqrt{\frac{n}{m}} \|\boldsymbol{d}\|_\infty^{1/2}.$$

■

Lemma B.19 *Let the sets $\mathcal{D}$, $\mathcal{V}$ be defined as (B.2.1) and (B.2.2), we have*

$$\gamma_2(\mathcal{V}, \|\cdot\|) \leq C\sqrt{\frac{n}{m}} \|\boldsymbol{d}\|_\infty^{1/2} \log^{3/2} n \log^{1/2} m,$$

where $C > 0$ is some constant.

Proof By Definition B.8, we know that

$$\gamma_2(\mathcal{V}, \|\cdot\|) \leq C \int_0^{d_2(\mathcal{V})} \log^{1/2} \mathcal{N}(\mathcal{V}, \|\cdot\|, \epsilon)\, d\epsilon,$$

for some constant $C > 0$, where the right hand side is known as the "Dudley integral". To estimate the covering number $\mathcal{N}(\mathcal{V}, \|\cdot\|, \epsilon)$, we know that for any $\boldsymbol{v},\ \boldsymbol{v}' \in \mathcal{D}$,

$$\|\boldsymbol{V}_{\boldsymbol{v}} - \boldsymbol{V}_{\boldsymbol{v}'}\| = \|\boldsymbol{V}_{\boldsymbol{v}-\boldsymbol{v}'}\| \leq \frac{1}{\sqrt{m}} \left\|\operatorname{diag}(\boldsymbol{d})^{1/2}\right\| \|\boldsymbol{F}_m(\boldsymbol{v}-\boldsymbol{v}')\|_\infty \leq \frac{1}{\sqrt{m}} \|\boldsymbol{d}\|_\infty^{1/2} \|\boldsymbol{F}_m(\boldsymbol{v}-\boldsymbol{v}')\|_\infty. \tag{B.2.7}$$

Let $\|\boldsymbol{v}\|_{\widehat{\infty}} \doteq \|\boldsymbol{F}_m \boldsymbol{v}\|_\infty$ that $\|\boldsymbol{v}\|_{\widehat{\infty}} \leq \|\boldsymbol{v}\|_1$, we have $\mathcal{N}(\mathcal{V}, \|\cdot\|, \epsilon) \leq \mathcal{N}\left(\mathcal{D}, m^{-1/2}\|\boldsymbol{d}\|_\infty^{1/2}\|\cdot\|_{\widehat{\infty}}, \epsilon\right)$. Next, we bound the covering number $\mathcal{N}\left(\mathcal{D}, m^{-1/2}\|\boldsymbol{d}\|_\infty^{1/2}\|\cdot\|_{\widehat{\infty}}, \epsilon\right)$ when ϵ is small and large, respectively.

When ϵ is small (i.e., $\epsilon \leq \mathcal{O}(1/\sqrt{m})$), let $\mathcal{B}_1^{[n]} = \{\boldsymbol{v} \in \mathbb{C}^m :\ \|\boldsymbol{v}\|_1 \leq 1, \operatorname{supp} \boldsymbol{v} \in [n]\}$, then it is obvious that $\mathcal{D} \subseteq \sqrt{n}\mathcal{B}_1^{[n]}$. By Proposition 10.1 of [Rau10], we have

$$\begin{aligned}\mathcal{N}\left(\mathcal{D}, m^{-1/2}\|\boldsymbol{d}\|_\infty^{1/2}\|\cdot\|_{\widehat{\infty}}, \epsilon\right) &\leq \mathcal{N}\left(\sqrt{n}\mathcal{B}_1^{[n]}, m^{-1/2}\|\boldsymbol{d}\|_\infty^{1/2}\|\cdot\|_1, \epsilon\right) \\ &\leq \mathcal{N}(\mathcal{B}_1^{[n]}, \|\cdot\|_1, \|\boldsymbol{d}\|_\infty^{-1/2}\sqrt{\frac{m}{n}}\epsilon) \leq \left(1 + \frac{2\sqrt{n}\|\boldsymbol{d}\|_\infty^{1/2}}{\sqrt{m}\epsilon}\right)^n.\end{aligned}$$

Thus, we have

$$\log \mathcal{N}\left(\mathcal{D}, m^{-1/2}\left\|\boldsymbol{d}\right\|_{\infty}^{1/2}\left\|\cdot\right\|_{\widehat{\infty}}, \epsilon\right) \leq n \log\left(1+\frac{2\sqrt{n}\left\|\boldsymbol{d}\right\|_{\infty}^{1/2}}{\sqrt{m}\epsilon}\right).$$

If the scalar ϵ is large, let us introduce a norm

$$\|\boldsymbol{v}\|_1^* = \sum_{k=1}^{m} |\Re(v_k)| + |\Im(v_k)|, \quad \forall\, \boldsymbol{v} \in \mathbb{C}^m, \tag{B.2.8}$$

which is the usual ℓ_1-norm after identification of $\mathbb{C}^m$ with $\mathbb{R}^{2m}$. Let $\mathcal{B}_{\|\cdot\|_1^*}^{[n]} = \left\{\boldsymbol{v} \in \mathbb{C}^m : \ \|\boldsymbol{v}\|_1^* \leq 1, \mathrm{supp}(\boldsymbol{v}) \in [n]\right\}$, then we have $\mathcal{D} \subseteq \sqrt{2n}\mathcal{B}_{\|\cdot\|_1^*}^{[n]}$. By Lemma B.20, we obtain

$$\log \mathcal{N}\left(\mathcal{D}, m^{-1/2}\left\|\boldsymbol{d}\right\|_{\infty}^{1/2}\left\|\cdot\right\|_{\widehat{\infty}}, \epsilon\right) \leq \log \mathcal{N}\left(\mathcal{B}_{\|\cdot\|_1^*}^{[n]}, \left\|\cdot\right\|_{\widehat{\infty}}, \frac{\sqrt{m}}{\sqrt{2n}}\left\|\boldsymbol{d}\right\|_{\infty}^{-1/2}\epsilon\right) \leq \frac{Cn}{m\epsilon^2}\left\|\boldsymbol{d}\right\|_{\infty} \log m \log n$$

Finally, we combine the results above to estimate the "Dudley integral",

$$\begin{aligned}
\mathcal{I} &\doteq \int_0^{d_2(\mathcal{V})} \log^{1/2} \mathcal{N}\left(\mathcal{V}, \|\cdot\|, \epsilon\right) d\epsilon \\
&\leq \sqrt{n}\int_0^{\kappa} \log^{1/2}\left(1+2\sqrt{\frac{n}{m}}\frac{\left\|\boldsymbol{d}\right\|_{\infty}^{1/2}}{\epsilon}\right) d\epsilon + C\sqrt{\frac{n}{m}\left\|\boldsymbol{d}\right\|_{\infty}\log m \log n}\int_{\kappa}^{\sqrt{\frac{n}{m}}\|\boldsymbol{v}\|_{\infty}^{1/2}} \epsilon^{-1} d\epsilon \\
&\leq \frac{2n}{\sqrt{m}}\left\|\boldsymbol{d}\right\|_{\infty}^{1/2}\int_0^{\frac{\kappa}{2}\|\boldsymbol{d}\|_{\infty}^{-1/2}\sqrt{\frac{m}{n}}} \log^{1/2}\left(1+t^{-1}\right) dt + C\sqrt{\frac{n}{m}\left\|\boldsymbol{d}\right\|_{\infty}\log m\log n}\log\left(\sqrt{\frac{n}{m}}\left\|\boldsymbol{d}\right\|_{\infty}^{1/2}/\kappa\right) \\
&\leq \kappa\sqrt{n}\sqrt{\log\left(e\left(1+\frac{2}{\kappa}\left\|\boldsymbol{d}\right\|_{\infty}^{1/2}\sqrt{\frac{n}{m}}\right)\right)} + C\sqrt{\frac{n}{m}\left\|\boldsymbol{d}\right\|_{\infty}\log m\log n}\log\left(\sqrt{\frac{n}{m}}\left\|\boldsymbol{d}\right\|_{\infty}^{1/2}/\kappa\right)
\end{aligned}$$

where the last inequality we used Lemma 10.3 of [Rau10]. Choose $\kappa = \frac{\|\boldsymbol{d}\|_{\infty}^{1/2}}{\sqrt{m}}$, we obtain the desired result.

■

Lemma B.20 *Let* $\mathcal{B}_{\|\cdot\|_1^*}^{[n]} = \left\{\boldsymbol{v} \in \mathbb{C}^m : \ \|\boldsymbol{v}\|_1^* \leq 1, \ \mathrm{supp}(\boldsymbol{v}) \in [n]\right\}$, *and* $\|\cdot\|_1^*$ *is defined in* (B.2.8), *we have*

$$\log \mathcal{N}\left(\mathcal{B}_{\|\cdot\|_1^*}^{[n]}, \left\|\cdot\right\|_{\widehat{\infty}}, \epsilon\right) \leq \frac{C}{\epsilon^2}\log m \log n \tag{B.2.9}$$

for some constant $C > 0$, *where the norm* $\|\boldsymbol{v}\|_{\widehat{\infty}} = \|\boldsymbol{F}_m\boldsymbol{v}\|_{\infty}$.

Proof Let $\mathcal{U} = \{\pm\boldsymbol{e}_1, \cdots, \pm\boldsymbol{e}_n, \pm\mathrm{i}\boldsymbol{e}_1, \cdots, \pm\mathrm{i}\boldsymbol{e}_n\}$, it is obvious that $\mathcal{B}_{\|\cdot\|_1^*}^{[n]} \subseteq \mathrm{conv}(\mathcal{U})$, where $\mathrm{conv}(\mathcal{U})$ denotes the convex hull of $\mathcal{U}$. Fix any $\boldsymbol{w} \in \mathcal{U}$, the idea is to approximate $\boldsymbol{w}$ by a finite set of very sparse vectors. We

define a random vector

$$\boldsymbol{z} = \begin{cases} \operatorname{sign}\left(\Re(w_j)\right)\boldsymbol{e}_j, & \text{with prob. } \left|\Re\left(w_j\right)\right|,\ 1\le j\le n \\ \operatorname{sign}\left(\Im(w_j)\right)\boldsymbol{e}_j, & \text{with prob. } \left|\Im\left(w_j\right)\right|,\ 1\le j\le n \\ \boldsymbol{0}, & \text{with prob. } 1-\|\boldsymbol{w}\|_1^*. \end{cases}$$

Since $\|\boldsymbol{v}\|_1^* \le 1$, this is a valid probability distribution with $\mathbb{E}\left[\boldsymbol{z}\right] = \boldsymbol{w}$. Let $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_L$ be independent copies of $\boldsymbol{z}$, where L is a number to be determined later. We attempt to approximate $\boldsymbol{w}$ with a L-sparse vector

$$\boldsymbol{z}_S = \frac{1}{L}\sum_{k=1}^{L}\boldsymbol{z}_k.$$

By using symmetrization (Lemma 6.7 of [Rau10]), we obtain

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{z}_S - \boldsymbol{v}\|_{\widehat{\infty}}\right] &= \mathbb{E}\left[\left\|\frac{1}{L}\sum_{k=1}^{L}\left(\boldsymbol{z}_k - \mathbb{E}\left[\boldsymbol{z}_k\right]\right)\right\|_{\widehat{\infty}}\right] \le \frac{2}{L}\mathbb{E}\left[\left\|\sum_{k=1}^{L}\varepsilon_k\boldsymbol{z}_k\right\|_{\widehat{\infty}}\right] \\ &= \frac{2}{L}\mathbb{E}\left[\max_{\ell\in[m]}\left|\sum_{k=1}^{L}\varepsilon_k\left\langle\boldsymbol{f}_\ell, \boldsymbol{z}_k\right\rangle\right|\right]\end{aligned}$$

where $\boldsymbol{\varepsilon} = [\varepsilon_1, \cdots, \varepsilon_L]^*$ is a Rademacher vector, independent of $\{\boldsymbol{z}_k\}_{k=1}^L$. Fix a realization of $\{\boldsymbol{z}_k\}_{k=1}^L$, by applying the Hoeffding's inequality to $\boldsymbol{\epsilon}$, we obtain

$$\mathbb{P}_{\boldsymbol{\varepsilon}}\left(\left|\sum_{k=1}^{L}\varepsilon_k\left\langle\boldsymbol{f}_\ell, \boldsymbol{z}_k\right\rangle\right| \ge \sqrt{L}t\right) \le \mathbb{P}_{\boldsymbol{\varepsilon}}\left(\left|\sum_{k=1}^{L}\varepsilon_k\left\langle\boldsymbol{f}_\ell, \boldsymbol{z}_k\right\rangle\right| \ge \left\|\sum_{k=1}^{L}\left\langle\boldsymbol{f}_\ell, \boldsymbol{z}_k\right\rangle\right\| t\right) \le 2\exp\left(-t^2/2\right)$$

for all $t > 0$ and $\ell \in [m]$. Thus, by combining the result above with Lemma 6.6 of [Rau10], it implies that

$$\mathbb{E}\left[\max_{\ell\in[m]}\left|\sum_{k=1}^{L}\varepsilon_k\left\langle\boldsymbol{f}_\ell, \boldsymbol{z}_k\right\rangle\right|\right] \le C\sqrt{L\log(8m)},$$

with $C = \sqrt{2} + \left(4\sqrt{2}\log 8\right)^{-1} < 1.5$. By Fubini's theorem, we obtain

$$\mathbb{E}\left[\|\boldsymbol{z}_S - \boldsymbol{v}\|_{\widehat{\infty}}\right] \le \frac{2}{L}\mathbb{E}_{\boldsymbol{z}}\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\max_{\ell\in[m]}\left|\sum_{k=1}^{L}\varepsilon_k\left\langle\boldsymbol{f}_\ell, \boldsymbol{z}_k\right\rangle\right|\right] \le \frac{3}{\sqrt{L}}\sqrt{\log(8m)}. \tag{B.2.10}$$

This implies that there exists a vector $\boldsymbol{z}_S = \frac{1}{L}\sum_{k=1}^{L}\boldsymbol{z}_k$ where each $\boldsymbol{z}_k \in \mathcal{U}$ such that $\|\boldsymbol{z}_S - \boldsymbol{v}\|_{\widehat{\infty}} \le \frac{3}{\sqrt{L}}\sqrt{\log(8m)}$. Since each $\boldsymbol{z}_k$ can take $4n+1$ values, so that $\boldsymbol{z}_S$ can take at most $(4n+1)^L$ values. And for each $\boldsymbol{v} \in \operatorname{conv}(\mathcal{U})$, according to (B.2.10), we can therefore find a vector $\boldsymbol{z}_S$ such that $\|\boldsymbol{v} - \boldsymbol{z}_S\|_{\widehat{\infty}} \le \epsilon$ with the choice $L \le \lfloor \frac{9}{\epsilon^2}\log(10m)\rfloor$. Thus, we have

$$\log\mathcal{N}\left(\mathcal{B}_{\|\cdot\|_1^*}^{[n]}, \|\cdot\|_{\widehat{\infty}}, \epsilon\right) \le \log\mathcal{N}\left(\operatorname{conv}(\mathcal{U}), \|\cdot\|_{\widehat{\infty}}, \epsilon\right) \le L\log(4n+1) \le \frac{9}{\epsilon^2}\log(10m)\log(4n+1)$$

as desired. ∎

B.3 Concentration via Decoupling

In this section, we assume that $\|\boldsymbol{x}\| = 1$, and we develop concentration inequalities for the following quantities

$$\boldsymbol{Y}(\boldsymbol{g}) = \frac{1}{m}\boldsymbol{R}_{[1:n]}\boldsymbol{C}_{\boldsymbol{g}}^{*}\operatorname{diag}\left(\left|\boldsymbol{g}\circledast\boldsymbol{x}\right|^{2}\right)\boldsymbol{C}_{\boldsymbol{g}}\boldsymbol{R}_{[1:n]}^{\top}, \tag{B.3.1}$$

$$\boldsymbol{M}(\boldsymbol{g}) = \frac{2\sigma^{2}+1}{m}\boldsymbol{R}_{[1:n]}\boldsymbol{C}_{\boldsymbol{g}}^{*}\operatorname{diag}\left(\zeta_{\sigma^{2}}(\boldsymbol{g}\circledast\boldsymbol{x})\right)\boldsymbol{C}_{\boldsymbol{g}}\boldsymbol{R}_{[1:n]}^{\top}, \tag{B.3.2}$$

via the decoupling technique and moments control, where $\zeta_{\sigma^2}(\cdot)$ is defined in (24.2.1) and $\sigma^2 > 1/2$. Suppose $\boldsymbol{g}\in\mathbb{C}^m$ is complex Gaussian random variable $\boldsymbol{g}\sim\mathcal{CN}(\boldsymbol{0},\boldsymbol{I})$. Once all the moments are bounded, it is easy to turn the moment bounds into a tail bound via Lemma B.6 and Lemma B.7. To bound the moments, we use the decoupling technique developed in [AG93, DlPG99, KMR14]. The basic idea is to decouple the terms above into terms like

$$\mathcal{Q}_{dec}^{\boldsymbol{Y}}(\boldsymbol{g}^{1},\boldsymbol{g}^{2}) = \frac{1}{m}\boldsymbol{R}_{[1:n]}\boldsymbol{C}_{\boldsymbol{g}^{1}}^{*}\operatorname{diag}\left(\left|\boldsymbol{g}^{2}\circledast\boldsymbol{x}\right|^{2}\right)\boldsymbol{C}_{\boldsymbol{g}^{1}}\boldsymbol{R}_{[1:n]}^{\top}, \tag{B.3.3}$$

$$\mathcal{Q}_{dec}^{\boldsymbol{M}}(\boldsymbol{g}^{1},\boldsymbol{g}^{2}) = \frac{1+2\sigma^{2}}{m}\boldsymbol{R}_{[1:n]}\boldsymbol{C}_{\boldsymbol{g}^{1}}^{*}\operatorname{diag}\left(\eta_{\sigma^{2}}\left(\boldsymbol{g}^{2}\circledast\boldsymbol{x}\right)\right)\boldsymbol{C}_{\boldsymbol{g}^{1}}\boldsymbol{R}_{[1:n]}^{\top}, \tag{B.3.4}$$

where $\eta_{\sigma^2}(t) = 1 - 2\pi\sigma^2\xi_{\sigma^2-\frac{1}{2}}(t)$, and $\boldsymbol{g}^1$ and $\boldsymbol{g}^2$ are two independent random variables with

$$\boldsymbol{g}^{1} = \boldsymbol{g}+\boldsymbol{\delta}, \qquad \boldsymbol{g}^{2} = \boldsymbol{g}-\boldsymbol{\delta}, \tag{B.3.5}$$

where $\boldsymbol{\delta}\sim\mathcal{CN}(\boldsymbol{0},\boldsymbol{I})$ is an independent copy of $\boldsymbol{g}$. As we discussed in Chapter 21, it turns out that controlling the moments of the decoupled terms $\mathcal{Q}_{dec}^{\boldsymbol{Y}}(\boldsymbol{g}^1,\boldsymbol{g}^2)$ and $\mathcal{Q}_{dec}^{\boldsymbol{M}}(\boldsymbol{g}^1,\boldsymbol{g}^2)$ for convolutional random matrices is easier and sufficient for providing the tail bound of $\boldsymbol{Y}$ and $\boldsymbol{M}$. The detailed results and proofs are described in the following subsections.

B.3.1 Concentration of $\boldsymbol{Y}(\boldsymbol{g})$

In this subsection, we show that

Theorem B.21 *Let $\boldsymbol{g}\sim\mathcal{CN}(\boldsymbol{0},\boldsymbol{I})$, and let $\boldsymbol{Y}(\boldsymbol{g})$ be defined as (B.3.1). For any $\delta>0$, when $m\geq C\delta^{-2}n\log^{7}n$,*

we have

$$\|\boldsymbol{Y}(\boldsymbol{g}) - \boldsymbol{x}\boldsymbol{x}^* - \boldsymbol{I}\| \le \delta,$$

holds with probability at least $1 - 2m^{-c}$. *Here* $c,\ C > 0$ *are some numerical constants.*

Proof Suppose $\boldsymbol{g}^1, \boldsymbol{g}^2$ are defined as (B.3.5), and $\mathcal{Q}^{\boldsymbol{Y}}_{dec}(\boldsymbol{g}^1, \boldsymbol{g}^2)$ is defined as (B.3.3). Let $[\boldsymbol{g}^2 \circledast \boldsymbol{x}]_k = \left(\boldsymbol{g}^2_k\right)^* \boldsymbol{x}$ and $\boldsymbol{C}_{\boldsymbol{g}^1}\boldsymbol{R}^\top_{[1:n]} = \begin{bmatrix} \left(\boldsymbol{g}^1_1\right)^* \\ \dots \\ \left(\boldsymbol{g}^1_m\right)^* \end{bmatrix}$, then by Lemma B.22, we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\delta}}\left[\mathcal{Q}^{\boldsymbol{Y}}_{dec}(\boldsymbol{g}^1, \boldsymbol{g}^2)\right] &= \frac{1}{m}\sum_{k=1}^m \mathbb{E}_{\boldsymbol{\delta}}\left[\left|\left(\boldsymbol{g}_k - \boldsymbol{\delta}_k\right)^* \boldsymbol{x}\right|^2 \left(\boldsymbol{g}_k + \boldsymbol{\delta}_k\right)\left(\boldsymbol{g}_k + \boldsymbol{\delta}_k\right)^*\right] \\
&= \frac{1}{m}\sum_{k=1}^m \left(\left|\boldsymbol{g}_k^*\boldsymbol{x}\right|^2 \boldsymbol{g}_k\boldsymbol{g}_k^* + \boldsymbol{g}_k\boldsymbol{g}_k^* + \left|\boldsymbol{g}_k^*\boldsymbol{x}\right|^2 \boldsymbol{I} + \boldsymbol{x}\boldsymbol{x}^* + \boldsymbol{I} - \boldsymbol{x}\boldsymbol{x}^*\boldsymbol{g}_k\boldsymbol{g}_k^* - \boldsymbol{g}_k\boldsymbol{g}_k^*\boldsymbol{x}\boldsymbol{x}^*\right) \\
&= 4\boldsymbol{I} + \boldsymbol{Y}(\boldsymbol{g}) - \mathbb{E}_{\boldsymbol{g}}\left[\boldsymbol{Y}(\boldsymbol{g})\right] - \frac{1}{m}\sum_{k=1}^m \left(\boldsymbol{x}\boldsymbol{x}^*\boldsymbol{g}_k\boldsymbol{g}_k^* + \boldsymbol{g}_k\boldsymbol{g}_k^*\boldsymbol{x}\boldsymbol{x}^* - 2\boldsymbol{x}\boldsymbol{x}^*\right) \\
&\quad + \frac{1}{m}\sum_{k=1}^m \left(\boldsymbol{g}_k\boldsymbol{g}_k^* - \boldsymbol{I}\right) + \frac{1}{m}\sum_{k=1}^m \left(\left|\boldsymbol{g}_k^*\boldsymbol{x}\right|^2 - 1\right)\boldsymbol{I}.
\end{aligned}$$

Thus, by Minkowski inequality and Jensen's inequality, for any positive integer $p \ge 1$, we have

$$\begin{aligned}
&\left(\mathbb{E}_{\boldsymbol{g}}\left[\|\boldsymbol{Y}(\boldsymbol{g}) - \mathbb{E}\left[\boldsymbol{Y}(\boldsymbol{g})\right]\|^p\right]\right)^{1/p} \\
&\le 4\left(\mathbb{E}_{\boldsymbol{g}}\left[\left\|\frac{1}{m}\boldsymbol{R}_{[1:n]}\boldsymbol{C}^*_{\boldsymbol{g}}\boldsymbol{C}_{\boldsymbol{g}}\boldsymbol{R}^\top_{[1:n]} - \boldsymbol{I}\right\|^p\right]\right)^{1/p} + \left(\mathbb{E}_{\boldsymbol{g}}\left[\left\|\mathbb{E}_{\boldsymbol{\delta}}\left[\mathcal{Q}^{\boldsymbol{Y}}_{dec}(\boldsymbol{g}^1, \boldsymbol{g}^2)\right] - 4\boldsymbol{I}\right\|^p\right]\right)^{1/p} \\
&\le \underbrace{4\left(\mathbb{E}_{\boldsymbol{g}}\left[\left\|\frac{1}{m}\boldsymbol{R}_{[1:n]}\boldsymbol{C}^*_{\boldsymbol{g}}\boldsymbol{C}_{\boldsymbol{g}}\boldsymbol{R}^\top_{[1:n]} - \boldsymbol{I}\right\|^p\right]\right)^{1/p}}_{\mathcal{T}_1} + \underbrace{\left(\mathbb{E}_{\boldsymbol{g}^1,\boldsymbol{g}^2}\left[\left\|\mathcal{Q}^{\boldsymbol{Y}}_{dec}(\boldsymbol{g}^1, \boldsymbol{g}^2) - 4\boldsymbol{I}\right\|^p\right]\right)^{1/p}}_{\mathcal{T}_2}.
\end{aligned}$$

By Theorem B.12, we have

$$\mathcal{T}_1 \ \le\ C_1\left(\sqrt{\frac{n}{m}}\log^{3/2}\log^{1/2} m + \sqrt{\frac{n}{m}}\sqrt{p} + \frac{n}{m}p\right),$$

where $C_1 > 0$ is some numerical constant. For $\mathcal{T}_2$, by Theorem B.12 and Lemma B.23, we have

$$\begin{aligned}
\mathcal{T}_2 &= \left(\mathbb{E}_{\boldsymbol{g}^1,\boldsymbol{g}^2}\left[\left\|\mathcal{Q}^{\boldsymbol{Y}}_{dec}(\boldsymbol{g}^1, \boldsymbol{g}^2) - 4\boldsymbol{I}\right\|^p\right]\right)^{1/p} \\
&\le \left(\mathbb{E}_{\boldsymbol{g}^1,\boldsymbol{g}^2}\left[\left\|\mathcal{Q}^{\boldsymbol{Y}}_{dec}(\boldsymbol{g}^1, \boldsymbol{g}^2) - 2\frac{1}{m}\left\|\boldsymbol{g}^2 \circledast \boldsymbol{x}\right\|^2 \boldsymbol{I}\right\|^p\right]\right)^{1/p} \\
&\quad + 2\left(\mathbb{E}_{\boldsymbol{g}^1}\left[\left\|\frac{1}{m}\boldsymbol{R}_{[1:n]}\boldsymbol{C}^*_{\boldsymbol{g}^1}\boldsymbol{C}_{\boldsymbol{g}^1}\boldsymbol{R}^\top_{[1:n]} - 2\boldsymbol{I}\right\|^p\right]\right)^{1/p}
\end{aligned}$$

$$\leq C_2 \left(\mathbb{E}_{\boldsymbol{g}^2} \left[\left\| \boldsymbol{g}^2 \circledast \boldsymbol{x} \right\|_\infty^{2p} \right]^{1/p} + 2 \right) \left(\sqrt{\frac{n}{m}} \log^{3/2} \log^{1/2} m + \sqrt{\frac{n}{m}} \sqrt{p} + \frac{n}{m} p \right)$$
$$\leq C_3 \left(\sqrt{\frac{n}{m}} \left(\log^{3/2} n \log^{3/2} m \right) p + \sqrt{\frac{n}{m}} \left(\log m \right) p^{3/2} + \frac{n}{m} \left(\log m \right) p^2 \right).$$

where $C_2,\ C_3 > 0$ are some numerical constants. Thus, combining the estimates for $\mathcal{T}_1$ and $\mathcal{T}_2$ above, we have

$$\left(\mathbb{E}_{\boldsymbol{g}} \left[\| \boldsymbol{Y}(\boldsymbol{g}) - \mathbb{E}\left[\boldsymbol{Y}(\boldsymbol{g}) \right] \|^p \right] \right)^{1/p} \leq C_4 \left(\sqrt{\frac{n}{m}} \left(\log^{3/2} n \log^{3/2} m \right) p + \sqrt{\frac{n}{m}} \left(\log m \right) p^{3/2} + \frac{n}{m} \left(\log m \right) p^2 \right),$$

where $C_4 > 0$ is some numerical constant. Therefore, by using Lemma B.7, for any $\delta > 0$, whenever $m \geq C_5 \delta^{-2} n \log^4 m \log^3 n$

$$\| \boldsymbol{Y} - \mathbb{E}\left[\boldsymbol{Y} \right] \| \leq \delta,$$

with probability at least $1 - 2m^{-c}$, where $c > 0$ is some numerical constant . Finally, using Lemma B.22, we get the desired result. ■

Lemma B.22 *Let $\boldsymbol{g} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$, and let $\boldsymbol{Y}(\boldsymbol{g})$ be defined as (B.3.1), then we have*

$$\mathbb{E}\left[\boldsymbol{Y}(\boldsymbol{g}) \right] = \boldsymbol{x}\boldsymbol{x}^* + \boldsymbol{I}.$$

Proof Please see Lemma 6.2 of [SQW16]. ■

Lemma B.23 *Suppose $\widetilde{\boldsymbol{g}} \sim \mathcal{CN}(\boldsymbol{0}, 2\boldsymbol{I})$, for any positive integer $p \geq 1$, we have*

$$\left(\mathbb{E}_{\widetilde{\boldsymbol{g}}} \left[\| \widetilde{\boldsymbol{g}} \circledast \boldsymbol{x} \|_\infty^p \right] \right)^{1/p} \leq 6 \sqrt{\log m} \sqrt{p}.$$

Proof By Minkowski inequality, we have

$$\mathbb{E}_{\widetilde{\boldsymbol{g}}} \left[\| \widetilde{\boldsymbol{g}} \circledast \boldsymbol{x} \|_\infty^p \right]^{1/p} \leq \mathbb{E}\left[\| \widetilde{\boldsymbol{g}} \circledast \boldsymbol{x} \|_\infty \right] + \left(\mathbb{E}_{\widetilde{\boldsymbol{g}}} \left[\left(\| \widetilde{\boldsymbol{g}} \circledast \boldsymbol{x} \|_\infty - \mathbb{E}\left[\| \widetilde{\boldsymbol{g}} \circledast \boldsymbol{x} \|_\infty \right] \right)^p \right] \right)^{1/p}.$$

We know that $\| \widetilde{\boldsymbol{g}} \circledast \boldsymbol{x} \|_\infty$ is 1-Lipschitz. Thus, by Gaussian concentration inequality in Lemma B.3, we have

$$\mathbb{P}\left(\left| \| \widetilde{\boldsymbol{g}} \circledast \boldsymbol{x} \|_\infty - \mathbb{E}_{\widetilde{\boldsymbol{g}}} \left[\| \widetilde{\boldsymbol{g}} \circledast \boldsymbol{x} \|_\infty \right] \right| \geq t \right) \leq 2 \exp\left(-t^2/2 \right).$$

By Lemma B.5, we know that $\| \widetilde{\boldsymbol{g}} \circledast \boldsymbol{x} \|_\infty$ is sub-Gaussian, and satisfies

$$\left(\mathbb{E}_{\widetilde{\boldsymbol{g}}} \left[\left| \| \widetilde{\boldsymbol{g}} \circledast \boldsymbol{x} \|_\infty - \mathbb{E}_{\widetilde{\boldsymbol{g}}} \left[\| \widetilde{\boldsymbol{g}} \circledast \boldsymbol{x} \|_\infty \right] \right|^p \right] \right)^{1/p} \leq 4\sqrt{p}.$$

Besides, let $\widetilde{\boldsymbol{g}} \circledast \boldsymbol{x} = \begin{bmatrix} \widetilde{\boldsymbol{g}}_1^* \boldsymbol{x} \\ \dots \\ \widetilde{\boldsymbol{g}}_m^* \boldsymbol{x} \end{bmatrix}$, then by Jensen's inequality, for all $\lambda > 0$, we have

$$\begin{aligned} \exp\left(\lambda \mathbb{E}\left[\left\|\widetilde{\boldsymbol{g}} \circledast \boldsymbol{x}\right\|_\infty\right]\right) \leq \mathbb{E}\left[\exp\left(\lambda \left\|\widetilde{\boldsymbol{g}} \circledast \boldsymbol{x}\right\|_\infty\right)\right] &= \mathbb{E}\left[\max_{1\leq k\leq m} \exp\left(\lambda \widetilde{\boldsymbol{g}}_k^* \boldsymbol{x}\right)\right] \\ &\leq \sum_{k=1}^m \mathbb{E}\left[\exp\left(\lambda \widetilde{\boldsymbol{g}}_k^* \boldsymbol{x}\right)\right] \leq m \exp\left(\lambda^2\right), \end{aligned}$$

where we used the fact that the moment generating function of $\widetilde{\boldsymbol{g}}_k^* \boldsymbol{x}$ satisfies $\mathbb{E}\left[\exp\left(\lambda \widetilde{\boldsymbol{g}}_k^* \boldsymbol{x}\right)\right] \leq \exp\left(\lambda^2\right)$. Taking the logarithms on both sides, we have

$$\mathbb{E}\left[\left\|\widetilde{\boldsymbol{g}} \circledast \boldsymbol{x}\right\|_\infty\right] \leq \log m/\lambda + \lambda.$$

Taking $\lambda = \sqrt{\log m}$, the right hand side achieves the minimum, which is

$$\mathbb{E}\left[\left\|\widetilde{\boldsymbol{g}} \circledast \boldsymbol{x}\right\|_\infty\right] \leq 2\sqrt{\log m}.$$

Combining the results above, we obtain the desired result. ∎

B.3.2 Concentration of $\boldsymbol{M}(\boldsymbol{g})$

Given $\boldsymbol{M}(\boldsymbol{g})$ as in (B.3.2), let us define

$$\boldsymbol{H}(\boldsymbol{g}) = \boldsymbol{P}_{\boldsymbol{x}^\perp} \boldsymbol{M} \boldsymbol{P}_{\boldsymbol{x}^\perp} \tag{B.3.6}$$

and correspondingly its decoupled term

$$\mathcal{Q}_{dec}^{\boldsymbol{H}}(\boldsymbol{g}^1, \boldsymbol{g}^2) = \boldsymbol{P}_{\boldsymbol{x}^\perp} \mathcal{Q}_{dec}^{\boldsymbol{M}}(\boldsymbol{g}^1, \boldsymbol{g}^2) \boldsymbol{P}_{\boldsymbol{x}^\perp}, \tag{B.3.7}$$

and let

$$\eta_{\sigma^2}(t) = 1 - 2\pi\sigma^2 \xi_{\sigma^2 - \frac{1}{2}}(t), \qquad \nu_{\sigma^2}(t) = 1 - \frac{4\pi\sigma^4}{2\sigma^2 - 1} \xi_{\sigma^2 - \frac{1}{2}}(t), \tag{B.3.8}$$

where $\sigma^2 > 1/2$. In this subsection, we show the following result.

Theorem B.24 *For any $\delta > 0$, when $m \geq C\delta^{-2} \left\|\boldsymbol{C}_{\boldsymbol{x}}\right\|^2 n \log^4 n$, we have*

$$\left\|\boldsymbol{H}(\boldsymbol{g}) - \boldsymbol{P}_{\boldsymbol{x}^\perp}\right\| \leq \delta \tag{B.3.9}$$

$$\left\| \boldsymbol{M}(\boldsymbol{g}) - \boldsymbol{I} - \frac{2\sigma^2}{1+2\sigma^2}\boldsymbol{x}\boldsymbol{x}^* \right\| \le 3\delta \tag{B.3.10}$$

$$\|\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{M}(\boldsymbol{g}) - \boldsymbol{P}_{\boldsymbol{x}^\perp}\| \le 2\delta \tag{B.3.11}$$

holds with probability at least $1 - cm^{-c'\log^3 n}$, *where* c, c' *and* C *are some positive numerical constants depending only on* σ^2.

Proof Let $\mathcal{Q}^{\boldsymbol{H}}_{dec}(\boldsymbol{g}^1, \boldsymbol{g}^2)$ be defined as (B.3.7). By using Lemma B.29, we calculate its expectation with respect to $\boldsymbol{\delta}$, we observe

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\delta}}\left[\mathcal{Q}^{\boldsymbol{H}}_{dec}(\boldsymbol{g}+\boldsymbol{\delta}, \boldsymbol{g}-\boldsymbol{\delta})\right] \\
&= \frac{1+2\sigma^2}{m}\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{R}_{[1:n]}\boldsymbol{C}^*_{\boldsymbol{g}}\operatorname{diag}\left(\mathbb{E}_{\boldsymbol{\delta}}\left[\eta_{\sigma^2}\left(|(\boldsymbol{g}-\boldsymbol{\delta})\circledast\boldsymbol{x}|\right)\right]\right)\boldsymbol{C}_{\boldsymbol{g}}\boldsymbol{R}^\top_{[1:n]}\boldsymbol{P}_{\boldsymbol{x}^\perp} \\
&\quad + \frac{1+2\sigma^2}{m}\left\langle \boldsymbol{1}, \mathbb{E}_{\boldsymbol{\delta}}\left[\eta_{\sigma^2}\left((\boldsymbol{g}-\boldsymbol{\delta})\circledast\boldsymbol{x}\right)\right]\right\rangle \boldsymbol{P}_{\boldsymbol{x}^\perp} \\
&= \frac{1+2\sigma^2}{m}\left[\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{R}_{[1:n]}\boldsymbol{C}^*_{\boldsymbol{g}}\operatorname{diag}\left(\zeta_{\sigma^2}(|\boldsymbol{g}\circledast\boldsymbol{x}|)\right)\boldsymbol{C}_{\boldsymbol{g}}\boldsymbol{R}^\top_{[1:n]}\boldsymbol{P}_{\boldsymbol{x}^\perp} + \left\langle \boldsymbol{1}, \zeta_{\sigma^2}\left(|\boldsymbol{g}\circledast\boldsymbol{x}|\right)\right\rangle \boldsymbol{P}_{\boldsymbol{x}^\perp}\right].
\end{aligned}$$

Using the results above and Lemma B.30, for all integer $p \ge 1$, we observe

$$\begin{aligned}
&\left(\mathbb{E}\left[\|\boldsymbol{H} - \mathbb{E}\left[\boldsymbol{H}\right]\|^p\right]\right)^{1/p} \\
&= \left(\mathbb{E}_{\boldsymbol{g}}\left[\left\|\frac{1+2\sigma^2}{m}\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{R}_{[1:n]}\boldsymbol{C}^*_{\boldsymbol{g}}\operatorname{diag}\left(\zeta_{\sigma^2}(|\boldsymbol{g}\circledast\boldsymbol{x}|)\right)\boldsymbol{C}_{\boldsymbol{g}}\boldsymbol{R}^\top_{[1:n]}\boldsymbol{P}_{\boldsymbol{x}^\perp} - \boldsymbol{P}_{\boldsymbol{x}^\perp}\right\|^p\right]\right)^{1/p} \\
&= \left(\mathbb{E}_{\boldsymbol{g}}\left[\left\|\mathbb{E}_{\boldsymbol{\delta}}\left[\mathcal{Q}^{\boldsymbol{H}}_{dec}(\boldsymbol{g}+\boldsymbol{\delta}, \boldsymbol{g}-\boldsymbol{\delta})\right] - \boldsymbol{P}_{\boldsymbol{x}^\perp} - \frac{1+2\sigma^2}{m}\left\langle \boldsymbol{1}, \zeta_{\sigma^2}\left(|\boldsymbol{g}\circledast\boldsymbol{x}|\right)\right\rangle \boldsymbol{P}_{\boldsymbol{x}^\perp}\right\|^p\right]\right)^{1/p} \\
&\le \left(\mathbb{E}_{\boldsymbol{g}}\left[\left\|\mathbb{E}_{\boldsymbol{\delta}}\left[\mathcal{Q}^{\boldsymbol{H}}_{dec}(\boldsymbol{g}+\boldsymbol{\delta}, \boldsymbol{g}-\boldsymbol{\delta})\right] - 2\boldsymbol{P}_{\boldsymbol{x}^\perp}\right\|^p\right]\right)^{1/p} + \left(\mathbb{E}_{\boldsymbol{g}}\left[\left|1 - \frac{1+2\sigma^2}{m}\left\langle \boldsymbol{1}, \zeta_{\sigma^2}\left(|\boldsymbol{g}\circledast\boldsymbol{x}|\right)\right\rangle\right|^p\right]\right)^{1/p} \\
&\le \left(\mathbb{E}_{\boldsymbol{g}^1,\boldsymbol{g}^2}\left[\left\|\mathcal{Q}^{\boldsymbol{M}}_{dec}(\boldsymbol{g}^1, \boldsymbol{g}^2) - 2\boldsymbol{I}\right\|^p\right]\right)^{1/p} + \left(\mathbb{E}_{\boldsymbol{g}}\left[\left|1 - \frac{1+2\sigma^2}{m}\left\langle \boldsymbol{1}, \zeta_{\sigma^2}\left(|\boldsymbol{g}\circledast\boldsymbol{x}|\right)\right\rangle\right|^p\right]\right)^{1/p},
\end{aligned}$$

where $\mathcal{Q}^{\boldsymbol{M}}_{dec}(\boldsymbol{g}^1, \boldsymbol{g}^2)$ is defined as (B.3.2), and we have used the Minkowski's inequality and the Jensen's inequality, respectively. By Lemma B.25 and Lemma B.31, we obtain

$$\left(\mathbb{E}\left[\|\boldsymbol{H} - \mathbb{E}\left[\boldsymbol{H}\right]\|^p\right]\right)^{1/p} \le C_{\sigma^2}\left(\sqrt{\frac{n}{m}}\log^{3/2} n \log^{1/2} m + \sqrt{p}\sqrt{\frac{n}{m}} + p\frac{n}{m}\right),$$

where C_{σ^2} is some numerical constant depending only on σ^2. Thus, by using the tail bound in Lemma B.6, for any $t > 0$, we obtain

$$\mathbb{P}\left(\|\boldsymbol{H} - \mathbb{E}\left[\boldsymbol{H}\right]\| \ge C_1\sqrt{\frac{n}{m}}\log^{3/2} n \log^{1/2} m + t\right) \le 2\exp\left(-C_2\frac{mt^2}{n}\right)$$

for some constants $C_1, C_2 > 0$. This further implies that for any $\delta > 0$, if $m \ge C_3\delta^{-2}n\log^3 n \log m$ for some

positive numerical constant C_3, we have

$$\|\boldsymbol{H} - \mathbb{E}[\boldsymbol{H}]\| \;\leq\; \delta,$$

holds with probability at least $1 - 2m^{-C_4 \log^3 n}$, where $C_4 > 0$ is numerical constant. Next, we use this result to bound the term $\|\boldsymbol{M} - \mathbb{E}[\boldsymbol{M}]\|$, by Lemma B.30, notice that

$$\begin{aligned}\|\boldsymbol{M} - \mathbb{E}[\boldsymbol{M}]\| \;&\leq\; \|\boldsymbol{P}_{\boldsymbol{x}^\perp}(\boldsymbol{M} - \mathbb{E}[\boldsymbol{M}])\boldsymbol{P}_{\boldsymbol{x}^\perp}\| + 2\|\boldsymbol{P}_{\boldsymbol{x}^\perp}(\boldsymbol{M} - \mathbb{E}[\boldsymbol{M}]\boldsymbol{P}_{\boldsymbol{x}})\| + \|\boldsymbol{P}_{\boldsymbol{x}}(\boldsymbol{M} - \mathbb{E}[\boldsymbol{M}])\boldsymbol{P}_{\boldsymbol{x}}\| \\ &\leq\; \|\boldsymbol{H} - \mathbb{E}[\boldsymbol{H}]\| + 2\|\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{M}\boldsymbol{x}\| + |\boldsymbol{x}^*(\boldsymbol{M} - \mathbb{E}[\boldsymbol{M}])\boldsymbol{x}|.\end{aligned}$$

Hence, by using the results in Lemma B.26 and Lemma B.27, whenever $m \geq C\|\boldsymbol{C}_{\boldsymbol{x}}\|^2 \delta^{-2} n \log^4 n$ we obtain

$$\|\boldsymbol{M} - \mathbb{E}[\boldsymbol{M}]\| \leq 3\delta,$$

holds with probability at least $1 - cm^{-c' \log^3 n}$. Here $c, c' > 0$ are some numerical constants. Similarly, we have

$$\begin{aligned}\|\boldsymbol{P}_{\boldsymbol{x}^\perp}(\boldsymbol{M} - \mathbb{E}[\boldsymbol{M}])\| &\leq \|\boldsymbol{P}_{\boldsymbol{x}^\perp}(\boldsymbol{M} - \mathbb{E}[\boldsymbol{M}])\boldsymbol{P}_{\boldsymbol{x}^\perp}\| + \|\boldsymbol{P}_{\boldsymbol{x}^\perp}(\boldsymbol{M} - \mathbb{E}[\boldsymbol{M}])\boldsymbol{P}_{\boldsymbol{x}}\| \\ &= \|\boldsymbol{H} - \mathbb{E}[\boldsymbol{H}]\| + \|\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{M}\boldsymbol{x}\|.\end{aligned}$$

Again, by Lemma B.27, we have

$$\|\boldsymbol{P}_{\boldsymbol{x}^\perp}(\boldsymbol{M} - \mathbb{E}[\boldsymbol{M}])\| \leq 2\delta,$$

holds with probability at least $1 - cm^{-c' \log^3 n}$. By using Lemma B.30, we obtain the desired results. ∎

Lemma B.25 *Suppose $\boldsymbol{g}^1, \boldsymbol{g}^2$ are independent with $\boldsymbol{g}^1, \boldsymbol{g}^2 \sim \mathcal{CN}(\boldsymbol{0}, 2\boldsymbol{I})$, and let $\mathcal{Q}_{dec}^{\boldsymbol{M}}(\boldsymbol{g}^1, \boldsymbol{g}^2)$ be defined as (B.3.4), then for any integer $p \geq 1$, we have*

$$\left(\mathbb{E}_{\boldsymbol{g}^1,\boldsymbol{g}^2}\left[\left\|\mathcal{Q}_{dec}^{\boldsymbol{M}}(\boldsymbol{g}^1, \boldsymbol{g}^2) - 2\boldsymbol{I}\right\|^p\right]\right)^{1/p} \leq C_{\sigma^2}\left(\sqrt{\frac{n}{m}}\log^{3/2} n \log^{1/2} m + \sqrt{p}\sqrt{\frac{n}{m}} + p\frac{n}{m}\right), \tag{B.3.12}$$

where $C_{\sigma^2} > 0$ is some numerical constant only depending on σ^2.

Proof Let $\boldsymbol{b} = \left(2\sigma^2 + 1\right)\eta_{\sigma^2}\left(\boldsymbol{g}^2 \circledast \boldsymbol{x}\right)$, set $\boldsymbol{b} = (b_k)_{k=1}^m$, and write $\mathcal{Q}_{dec}^{\boldsymbol{M}}(\boldsymbol{g}^1, \boldsymbol{g}^2) = \frac{1}{m}\boldsymbol{R}_{[1:n]}\boldsymbol{C}_{\boldsymbol{g}}^* \operatorname{diag}(\boldsymbol{b})\boldsymbol{C}_{\boldsymbol{g}}\boldsymbol{R}_{[1:n]}^\top$. By Minkowski's inequality, we observe

$$\left(\mathbb{E}_{\boldsymbol{g}^1,\boldsymbol{g}^2}\left[\left\|\mathcal{Q}_{dec}^{\boldsymbol{M}}(\boldsymbol{g}^1, \boldsymbol{g}^2) - 2\boldsymbol{I}\right\|^p\right]\right)^{1/p} \leq \underbrace{\left(\mathbb{E}_{\boldsymbol{g}^1,\boldsymbol{g}^2}\left[\left\|\mathcal{Q}_{dec}^{\boldsymbol{M}}(\boldsymbol{g}^1, \boldsymbol{g}^2) - \frac{2}{m}\sum_{k=1}^m b_k\boldsymbol{I}\right\|^p\right]\right)^{1/p}}_{\mathcal{T}_1}$$

$$+2\underbrace{\left\|\frac{1+2\sigma^2}{m}\left\langle \mathbf{1}, \eta_{\sigma^2}\left(\left|\boldsymbol{g}^2 \circledast \boldsymbol{x}\right|\right)\right\rangle - 1\right\|_{L^p}}_{\mathcal{T}_2}.$$

For the term $\mathcal{T}_1$, conditioned $\boldsymbol{g}^2$ so that $\boldsymbol{b}$ is fixed, Theorem B.12 implies that for any integer $p \geq 1$,

$$\left(\mathbb{E}_{\boldsymbol{g}^1}\left[\left\|\mathcal{Q}_{dec}^{\boldsymbol{M}}(\boldsymbol{g}^1, \boldsymbol{g}^2) - \frac{2}{m}\sum_{k=1}^{m} b_k \boldsymbol{I}\right\|^p \mid \boldsymbol{g}^2\right]\right)^{1/p} \leq C_{\sigma^2}\left\|\boldsymbol{b}\right\|_\infty \left(\sqrt{\frac{n}{m}}\log^{3/2} n \log^{1/2} m + \sqrt{p}\sqrt{\frac{n}{m}} + p\frac{n}{m}\right),$$

where $C_{\sigma^2} > 0$ is some numerical constant depending only on σ^2. Given the fact that $\|\boldsymbol{b}\|_\infty \leq c_{\sigma^2}$ for some constant $c_{\sigma^2} > 0$, and for any choice of $\boldsymbol{g}^2$, we have

$$\mathcal{T}_1 \;\leq\; C_{\sigma^2}\left(\sqrt{\frac{n}{m}}\log^{3/2} n \log^{1/2} m + \sqrt{p}\sqrt{\frac{n}{m}} + p\frac{n}{m}\right).$$

For the term $\mathcal{T}_2$, Lemma B.31 implies that

$$\mathcal{T}_2 = \left\|\frac{1+2\sigma^2}{m}\left\langle \mathbf{1}, \eta_{\sigma^2}\left(\left|\boldsymbol{g}^2 \circledast \boldsymbol{x}\right|\right)\right\rangle - 1\right\|_{L^p} \leq \frac{C'_{\sigma^2}}{\sqrt{m}}\left\|\boldsymbol{C}_{\boldsymbol{x}}\right\|\sqrt{p},$$

for some constant $C'_{\sigma^2} > 0$. Combining the results above and use the fact that $\|\boldsymbol{C}_{\boldsymbol{x}}\| \leq \sqrt{n}$, we obtain

$$\left(\mathbb{E}_{\boldsymbol{g}^1,\boldsymbol{g}^2}\left[\left\|\mathcal{Q}_{dec}^{\boldsymbol{M}}(\boldsymbol{g}^1, \boldsymbol{g}^2) - 2\boldsymbol{I}\right\|^p\right]\right)^{1/p} \leq C''_{\sigma^2}\left(\sqrt{\frac{n}{m}}\log^{3/2} n \log^{1/2} m + \sqrt{p}\sqrt{\frac{n}{m}} + p\frac{n}{m}\right),$$

where $C''_{\sigma^2} > 0$ is some numerical constant only depending on σ^2. ∎

Lemma B.26 *Let $\boldsymbol{g} \in \mathbb{C}^m$ be a complex Gaussian random variable $\boldsymbol{g} \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{I})$. Let $\boldsymbol{M}(\boldsymbol{g})$ be defined as (B.3.2). For any $\delta \geq 0$, whenever $m \geq C_{\sigma^2}\delta^{-1}\left\|\boldsymbol{C}_{\boldsymbol{x}}\right\|^2 n \log m$, we have*

$$\left|\boldsymbol{x}^*\left(\boldsymbol{M} - \mathbb{E}\left[\boldsymbol{M}\right]\right)\boldsymbol{x}\right| \leq \delta$$

holds with $1 - m^{-C'_{\sigma^2}\|\boldsymbol{C}_{\boldsymbol{x}}\|^2 n}$. Here, C_{σ^2}, C'_{σ^2} are some numerical constants depending on σ^2.

Proof Let $h(\boldsymbol{g}) = \left|\boldsymbol{x}^*\boldsymbol{M}(\boldsymbol{g})\boldsymbol{x}\right|^{1/2} = \sqrt{\frac{2\sigma^2+1}{m}}\left\|\mathrm{diag}\left(\zeta_{\sigma^2}^{1/2}(\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{g})\right)\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{g}\right\|$. Then we have its Wirtinger gradient

$$\frac{\partial}{\partial \boldsymbol{z}}h(\boldsymbol{g}) = \frac{1}{2}\sqrt{\frac{2\sigma^2+1}{m}}\left\|\mathrm{diag}\left(\zeta_{\sigma^2}^{1/2}\left(\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{g}\right)\right)\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{g}\right\|^{-1}\left[\boldsymbol{C}_{\boldsymbol{x}}^*\,\mathrm{diag}\left(\zeta_{\sigma^2}(\boldsymbol{C}_x\boldsymbol{g})\right)\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{g} + \boldsymbol{C}_{\boldsymbol{x}}^*\,\mathrm{diag}\left(f(\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{g})\right)\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{g}\right],$$

where $g_1(t) = \frac{|t|^2}{2\sigma^2}\exp\left(-\frac{|t|^2}{2\sigma^2}\right)$, so that

$$\left\|\nabla_{\boldsymbol{g}}h(\boldsymbol{g})\right\| = \sqrt{\frac{2\sigma^2+1}{m}}\left\|\mathrm{diag}\left(\zeta_{\sigma^2}^{1/2}(\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{g})\right)\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{g}\right\|^{-1} \times$$

$$\left\|\boldsymbol{C}_{\boldsymbol{x}}^{*}\operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{g})\right)\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{g}+\boldsymbol{C}_{\boldsymbol{x}}^{*}\operatorname{diag}\left(g_1(\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{g})\right)\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{g}\right\|.$$

Thus, we have

$$\left\|\nabla_{\boldsymbol{g}}h(\boldsymbol{g})\right\|\leq\sqrt{\frac{2\sigma^2+1}{m}}\left\|\boldsymbol{C}_{\boldsymbol{x}}\right\|\left(\left\|\operatorname{diag}\left(\zeta_{\sigma^2}^{1/2}\left(\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{g}\right)\right)\right\|+\left\|\operatorname{diag}\left(g_2(\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{g})\right\|\right),$$

where $g_2(t)=g_1(t)\zeta_{\sigma^2}^{-1/2}(t)$. By using the fact that $\left\|\zeta_{\sigma^2}^{1/2}\right\|_{\ell^\infty}\leq 1$ and $\|g_2\|_{\ell^\infty}\leq C_1$ for some constant $C_1>0$, we have

$$\left\|\nabla_{\boldsymbol{g}}h(\boldsymbol{g})\right\|\leq C_2\sqrt{\frac{2\sigma^2+1}{m}}\left\|\boldsymbol{C}_{\boldsymbol{x}}\right\|,$$

for some constant $C_2>0$. Therefore, we can see that the Lipschitz constant L of $h(\boldsymbol{g})$ is bounded by $C_2\sqrt{\frac{2\sigma^2+1}{m}}\left\|\boldsymbol{C}_{\boldsymbol{x}}\right\|$. Thus, by Gaussian concentration inequality, we observe

$$\mathbb{P}\left(\left|h(\boldsymbol{g})-\mathbb{E}\left[h(\boldsymbol{g})\right]\right|\geq t\right)\leq 2\exp\left(-\frac{C_{\sigma^2}mt^2}{\left\|\boldsymbol{C}_{\boldsymbol{x}}\right\|^2}\right)\tag{B.3.13}$$

holds with some constant $C_{\sigma^2}>0$ depending only on σ^2. Thus, we have

$$-t\leq h(\boldsymbol{g})-\mathbb{E}\left[h(\boldsymbol{g})\right]\leq t\tag{B.3.14}$$

holds with probability at least $1-2\exp\left(-\frac{C_{\sigma^2}mt^2}{\left\|\boldsymbol{C}_{\boldsymbol{x}}\right\|^2}\right)$. By Lemma B.30, we know that

$$\mathbb{E}\left[h^2(\boldsymbol{g})\right]=\boldsymbol{x}^{*}\mathbb{E}\left[\boldsymbol{M}(\boldsymbol{g})\right]\boldsymbol{x}=\frac{4\sigma^2+1}{2\sigma^2+1}.$$

This implies that

$$h^2(\boldsymbol{g})\leq\left(\mathbb{E}\left[h(\boldsymbol{g})\right]+t\right)^2\implies h^2(\boldsymbol{g})-\mathbb{E}\left[h^2(\boldsymbol{g})\right]\leq 2t\sqrt{\mathbb{E}\left[h^2(\boldsymbol{g})\right]}+t^2\leq 2t\sqrt{\frac{1+4\sigma^2}{1+2\sigma^2}}+t^2,\tag{B.3.15}$$

holds with probability at least $1-2\exp\left(-\frac{C_{\sigma^2}mt^2}{\left\|\boldsymbol{C}_{\boldsymbol{x}}\right\|^2}\right)$. On the other hand, (B.3.13) also implies that $h(\boldsymbol{g})$ is subgaussian, Lemma B.5 implies that

$$\mathbb{E}\left[\left(h(\boldsymbol{g})-\mathbb{E}\left[h(\boldsymbol{g})\right]\right)^2\right]\leq\frac{C'_{\sigma^2}\left\|\boldsymbol{C}_{\boldsymbol{x}}\right\|^2}{m}\implies\mathbb{E}\left[h^2(\boldsymbol{g})\right]\leq\left(\mathbb{E}\left[h(\boldsymbol{g})\right]\right)^2+\frac{C'_{\sigma^2}\left\|\boldsymbol{C}_{\boldsymbol{x}}\right\|^2}{m}$$

for some constant $C'_{\sigma^2}>0$ only depending on σ^2. Suppose $m\geq C''_{\sigma^2}\left\|\boldsymbol{C}_{\boldsymbol{x}}\right\|^2$ for some large constant $C''_{\sigma^2}>0$ depending on $\sigma^2>0$, from (B.3.14), we have

$$h(\boldsymbol{g})\geq\mathbb{E}\left[h(\boldsymbol{g})\right]-t\geq\sqrt{\mathbb{E}\left[h^2(\boldsymbol{g})\right]-\frac{C'_{\sigma^2}\left\|\boldsymbol{C}_{\boldsymbol{x}}\right\|^2}{m}}-t.$$

Suppose $t \leq \sqrt{\mathbb{E}\left[h^2(\boldsymbol{g})\right] - \frac{C'_{\sigma^2}\|\boldsymbol{C}_{\boldsymbol{x}}\|^2}{m}}$, by squaring both sides, we have

$$h^2(\boldsymbol{g}) \geq \mathbb{E}\left[h^2(\boldsymbol{g})\right] - \frac{C'_{\sigma^2}\left\|\boldsymbol{C}_{\boldsymbol{x}}\right\|^2}{m} + t^2 - 2t\sqrt{\mathbb{E}\left[h^2(\boldsymbol{g})\right] - \frac{C'_{\sigma^2}\left\|\boldsymbol{C}_{\boldsymbol{x}}\right\|^2}{m}}.$$

This further implies that

$$h^2(\boldsymbol{g}) - \mathbb{E}\left[h^2(\boldsymbol{g})\right] \geq t^2 - 2t\sqrt{\frac{4\sigma^2+1}{2\sigma^2+1} - \frac{C'_{\sigma^2}\left\|\boldsymbol{C}_{\boldsymbol{x}}\right\|^2}{m}} - \frac{C'_{\sigma^2}\left\|\boldsymbol{C}_{\boldsymbol{x}}\right\|^2}{m}, \tag{B.3.16}$$

holds $1-2\exp\left(-\frac{C_{\sigma^2}mt^2}{\|\boldsymbol{C}_{\boldsymbol{x}}\|^2}\right)$. Therefore, combining the results in (B.3.15) and (B.3.16), for any $\delta \geq 0$, whenever $m \geq C_4\delta^{-1}\left\|\boldsymbol{C}_{\boldsymbol{x}}\right\|^2 n\log m$, choosing $t = C_5\delta$, we have

$$\left|h^2(\boldsymbol{g}) - \mathbb{E}\left[h^2(\boldsymbol{g})\right]\right| \leq \delta,$$

holds with probability at least $1 - m^{-C_6\|\boldsymbol{C}_{\boldsymbol{x}}\|^2 n}$. ∎

Lemma B.27 *Let $\boldsymbol{g} \in \mathbb{C}^m$ be a complex Gaussian random variable $\boldsymbol{g} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$, and let $\boldsymbol{M}(\boldsymbol{g})$ be defined as (B.3.2). For any $\delta > 0$, whenever $m \geq C_{\sigma^2}\delta^{-2}\left\|\boldsymbol{C}_{\boldsymbol{x}}\right\|^2 n\log^4 n$, we have*

$$\left\|\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{M}\boldsymbol{x}\right\| \leq \delta,$$

holds with probability at least $1 - 2m^{-c_{\sigma^2}\log^3 n}$. Here, $c_{\sigma^2}, C_{\sigma^2}$ are some positive constants only depending on σ^2.

Proof First, let us define decoupled terms

$$\mathcal{Q}_{dec}^{\boldsymbol{M}\boldsymbol{x}^\perp}(\boldsymbol{g}^1, \boldsymbol{g}^2) = \frac{2\sigma^2+1}{m}\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{R}_{[1:n]}\boldsymbol{C}^*_{\boldsymbol{g}^1}\operatorname{diag}\left(\nu_{\sigma^2}\left(\boldsymbol{g}^2 \circledast \boldsymbol{x}\right)\right)\boldsymbol{C}_{\boldsymbol{g}^2}\boldsymbol{R}^\top_{[1:n]}\boldsymbol{x}, \tag{B.3.17}$$

$$\mathcal{Q}_{dec}^{\boldsymbol{H}\boldsymbol{x}^\perp}(\boldsymbol{g}^1, \boldsymbol{g}^2) = \frac{2\sigma^2+1}{m}\boldsymbol{R}_{[1:n]}\boldsymbol{C}^*_{\boldsymbol{g}^1}\operatorname{diag}\left(\nu_{\sigma^2}\left(\boldsymbol{g}^2 \circledast \boldsymbol{x}\right)\right)\boldsymbol{C}_{\boldsymbol{g}^2}\boldsymbol{R}^\top_{[1:n]}\boldsymbol{x}, \tag{B.3.18}$$

where $\nu_{\sigma^2}(t)$ is defined in (B.3.8). Let $\boldsymbol{C}_{\boldsymbol{g}}\boldsymbol{R}^\top_{[1:n]} = \begin{bmatrix}\boldsymbol{g}_1^* \\ \dots \\ \boldsymbol{g}_m^*\end{bmatrix}$ and $\boldsymbol{C}_{\boldsymbol{\delta}}\boldsymbol{R}^\top_{[1:n]} = \begin{bmatrix}\boldsymbol{\delta}_1^* \\ \dots \\ \boldsymbol{\delta}_m^*\end{bmatrix}$, then by Lemma B.29, we observe

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\delta}}\left[\mathcal{Q}_{dec}^{\boldsymbol{M}\boldsymbol{x}^\perp}(\boldsymbol{g}+\boldsymbol{\delta}, \boldsymbol{g}-\boldsymbol{\delta})\right] &= \frac{2\sigma^2+1}{m}\mathbb{E}_{\boldsymbol{\delta}}\left[\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{R}_{[1:n]}\boldsymbol{C}^*_{\boldsymbol{g}+\boldsymbol{\delta}}\operatorname{diag}\left(\nu_{\sigma^2}\left((\boldsymbol{g}-\boldsymbol{\delta})\circledast\boldsymbol{x}\right)\right)\boldsymbol{C}_{\boldsymbol{g}-\boldsymbol{\delta}}\boldsymbol{R}^\top_{[1:n]}\boldsymbol{x}\right] \\ &= \frac{2\sigma^2+1}{m}\sum_{k=1}^m \mathbb{E}_{\boldsymbol{\delta}}\left[\nu_{\sigma^2}\left((\boldsymbol{g}_k-\boldsymbol{\delta}_k)^*\boldsymbol{x}\right)\boldsymbol{P}_{\boldsymbol{x}^\perp}(\boldsymbol{g}_k+\boldsymbol{\delta}_k)(\boldsymbol{g}_k-\boldsymbol{\delta}_k)^*\boldsymbol{x}\right]\end{aligned}$$

$$
\begin{aligned}
&= \frac{2\sigma^2+1}{m}\sum_{k=1}^{m} \boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{g}_k \mathbb{E}_{\boldsymbol{\delta}_k^*\boldsymbol{x}}\left[\nu_{\sigma^2}\left(\left(\boldsymbol{g}_k-\boldsymbol{\delta}_k\right)^*\boldsymbol{x}\right)\left(\boldsymbol{g}_k-\boldsymbol{\delta}_k\right)^*\boldsymbol{x}\right] \\
&= \frac{2\sigma^2+1}{m}\sum_{k=1}^{m} \zeta_{\sigma^2}\left(\boldsymbol{g}_k^*\boldsymbol{x}\right)\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{g}_k\boldsymbol{g}_k^*\boldsymbol{x} \\
&= \frac{2\sigma^2+1}{m}\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{R}_{[1:n]}\boldsymbol{C}_{\boldsymbol{g}}^* \operatorname{diag}\left(\zeta_{\sigma^2}\left(\boldsymbol{g}\circledast\boldsymbol{x}\right)\right)\boldsymbol{C}_{\boldsymbol{g}}\boldsymbol{R}_{[1:n]}^\top\boldsymbol{x}.
\end{aligned}
$$

Thus, for any integer $p \geq 1$, we have

$$
\begin{aligned}
&\left(\mathbb{E}_{\boldsymbol{g}}\left[\left\|\frac{2\sigma^2+1}{m}\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{R}_{[1:n]}\boldsymbol{C}_{\boldsymbol{g}}^* \operatorname{diag}\left(\zeta_{\sigma^2}\left(\boldsymbol{g}\circledast\boldsymbol{x}\right)\right)\boldsymbol{C}_{\boldsymbol{g}}\boldsymbol{R}_{[1:n]}^\top\boldsymbol{x}\right\|^p\right]\right)^{1/p} \\
&= \left(\mathbb{E}_{\boldsymbol{g}}\left[\left\|\mathbb{E}_{\boldsymbol{\delta}}\left[\mathcal{Q}_{dec}^{\boldsymbol{M}\boldsymbol{x}^\perp}(\boldsymbol{g}+\boldsymbol{\delta},\boldsymbol{g}-\boldsymbol{\delta})\right]\right\|^p\right]\right)^{1/p} \\
&\leq \left(\mathbb{E}_{\boldsymbol{g}^1,\boldsymbol{g}^2}\left[\left\|\mathcal{Q}_{dec}^{\boldsymbol{M}\boldsymbol{x}^\perp}(\boldsymbol{g}^1,\boldsymbol{g}^2)\right\|^p\right]\right)^{1/p} \leq \left(\mathbb{E}_{\boldsymbol{g}^1,\boldsymbol{g}^2}\left[\left\|\mathcal{Q}_{dec}^{\boldsymbol{H}\boldsymbol{x}^\perp}(\boldsymbol{g}^1,\boldsymbol{g}^2)\right\|^p\right]\right)^{1/p}.
\end{aligned}
$$

By Lemma B.28, we have

$$
\left(\mathbb{E}_{\boldsymbol{g}^1,\boldsymbol{g}^2}\left[\left\|\mathcal{Q}_{dec}^{\boldsymbol{H}\boldsymbol{x}^\perp}(\boldsymbol{g}^1,\boldsymbol{g}^2)\right\|^p\right]\right)^{1/p} \leq C_{\sigma^2}\left\|\boldsymbol{C}_{\boldsymbol{x}}\right\|\left[\sqrt{\frac{n}{m}}\left(1+\sqrt{\frac{n}{m}}\log^{3/2}n\log^{1/2}m\right)\sqrt{p}+\frac{n}{m}p\right].
$$

Therefore, by Lemma B.6, finally for any $\delta>0$, whenever $m \geq C\delta^{-2}\left\|\boldsymbol{C}_{\boldsymbol{x}}\right\|^2 n\log^4 n$ we obtain

$$
\mathbb{P}\left(\left\|\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{M}\boldsymbol{x}\right\| \geq \delta\right) \leq 2m^{-c\log^3 n},
$$

where $c, C>0$ are some positive constants. ∎

Lemma B.28 *Let $\boldsymbol{g}^1$ and $\boldsymbol{g}^2$ be random variables defined as in (B.3.5), and let $\mathcal{Q}_{dec}^{\boldsymbol{H}\boldsymbol{x}^\perp}(\boldsymbol{g}^1,\boldsymbol{g}^2)$ be defined as (B.3.18). Then for any integer $p \geq 1$, we have*

$$
\left(\mathbb{E}_{\boldsymbol{g}^1,\boldsymbol{g}^2}\left[\left\|\mathcal{Q}_{dec}^{\boldsymbol{H}\boldsymbol{x}^\perp}(\boldsymbol{g}^1,\boldsymbol{g}^2)\right\|^p\right]\right)^{1/p} \leq C_{\sigma^2}\left\|\boldsymbol{C}_{\boldsymbol{x}}\right\|\left[\sqrt{\frac{n}{m}}\left(1+\sqrt{\frac{n}{m}}\log^{3/2}n\log^{1/2}m\right)\sqrt{p}+\frac{n}{m}p\right],
$$

where C_{σ^2} is some positive constant only depending on σ^2.

Proof First, we fix $\boldsymbol{g}^1$, and let $h(\boldsymbol{g}^2)=\mathcal{Q}_{dec}^{\boldsymbol{H}\boldsymbol{x}^\perp}(\boldsymbol{g}^1,\boldsymbol{g}^2)$. Let $g(t)=t\nu_{\sigma^2}(t)$, for which the Lipschitz constant $L_f \leq C_{\sigma^2}$ for some positive constant C_{σ^2} only depending on σ^2. Then given an independent copy $\widetilde{\boldsymbol{g}^2}$ of $\boldsymbol{g}^2$, we observe

$$
\begin{aligned}
\left\|h(\boldsymbol{g}^2)-h(\widetilde{\boldsymbol{g}^2})\right\| &\leq \frac{2\sigma^2+1}{m}\left\|\boldsymbol{R}_{[1:n]}\boldsymbol{C}_{\boldsymbol{g}^1}^*\right\|\left\|g(\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{g}^2)-g(\boldsymbol{C}_{\boldsymbol{x}}\widetilde{\boldsymbol{g}^2})\right\| \\
&\leq \underbrace{\frac{C'_{\sigma^2}}{m}\left\|\boldsymbol{R}_{[1:n]}\boldsymbol{C}_{\boldsymbol{g}^1}^*\right\|\left\|\boldsymbol{C}_{\boldsymbol{x}}\right\|}_{L_h}\left\|\boldsymbol{g}^2-\widetilde{\boldsymbol{g}^2}\right\|
\end{aligned}
$$

where L_h is the Lipschitz constant of $h(\boldsymbol{g}^2)$. Given the fact that $\mathbb{E}_{\boldsymbol{g}^2}\left[h(\boldsymbol{g}^2)\right] = \boldsymbol{0}$, by Lemma B.4, for any $t > \sqrt{n}L_h$ we have

$$\mathbb{P}\left(\left\|h(\boldsymbol{g}^2)\right\| \geq t\right) \leq e\mathbb{P}\left(\|\boldsymbol{v}\| \geq \frac{t}{L_h}\right) \leq e\exp\left(-\frac{1}{2}\left(\frac{t}{L_h} - \sqrt{n}\right)^2\right),$$

where $\boldsymbol{v} \in \mathbb{R}^n$ with $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, and we used the Gaussian concentration inequality for the tail bound of $\|\boldsymbol{v}\|$. By a change of variable, we obtain

$$\mathbb{P}\left(\left\|h(\boldsymbol{g}^2)\right\| \geq t + \sqrt{n}L_h\right) \leq e\exp\left(-\frac{1}{2L_h^2}t^2\right)$$

holds for all $t > 0$. By using the tail bound above, we obtain

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{g}^2}\left[\left\|\mathcal{Q}_{dec}^{\boldsymbol{H}\boldsymbol{x}^\perp}(\boldsymbol{g}^1, \boldsymbol{g}^2)\right\|^p\right] \\
=& \int_{t=0}^{\infty} \mathbb{P}\left(\left\|h(\boldsymbol{g}^2)\right\|^p \geq t\right) dt \\
=& \int_{t=0}^{\left(\sqrt{n}L_h\right)^p} \mathbb{P}\left(\left\|h(\boldsymbol{g}^2)\right\|^p \geq t\right) dt + \int_{t=\left(\sqrt{n}L_h\right)^p}^{\infty} \mathbb{P}\left(\left\|h(\boldsymbol{g}^2)\right\| \geq t^{1/p}\right) dt \\
\leq& \left(\sqrt{n}L_h\right)^p + p\int_{u=\sqrt{n}L_h}^{\infty} \mathbb{P}\left(\left\|h(\boldsymbol{g}^2)\right\| \geq u\right) u^{p-1} du \\
=& \left(\sqrt{n}L_h\right)^p + p\int_{u=0}^{\infty} \mathbb{P}\left(\left\|h(\boldsymbol{g}^2)\right\| \geq u + \sqrt{n}L_h\right) \left(u + \sqrt{n}L_h\right)^{p-1} du \\
\leq& \left(\sqrt{n}L_h\right)^p + 2^{p-2}p\left(\sqrt{n}L_h\right)^{p-1} e\int_{u=0}^{\infty} \exp\left(-\frac{u^2}{2L_h^2}\right) du + 2^{p-2}pe\int_{u=0}^{\infty} \exp\left(-\frac{u^2}{2L_h^2}\right) u^{p-1} du \\
=& \left(\sqrt{n}L_h\right)^p + \sqrt{\frac{\pi}{2}}2^{p-2}p\sqrt{n}^{p-1}L_h^p e + 2^{3p/2-3}pL_h^p e\int_{\tau=0}^{\infty} e^{-\tau}\tau^{\frac{p}{2}-1} d\tau \\
\leq& 3\sqrt{n}^p L_h^p\left(1 + \sqrt{\frac{\pi}{2}}2^{p-1}p + 2^{3p/2-3}p\Gamma(p/2)\right) \leq 3\left(4\sqrt{n}L_h\right)^p p\max\left\{(p/2)^{p/2}, \sqrt{2\pi}\right\},
\end{aligned}$$

where we used the fact that $\Gamma(p/2) \leq \max\left\{(p/2)^{p/2}, \sqrt{2\pi}\right\}$ for any integer $p \geq 1$. By Corollary B.13, we know that

$$\mathbb{E}_{\boldsymbol{g}^1}\left[\left\|\boldsymbol{R}_{[1:n]}\boldsymbol{C}_{\boldsymbol{g}^1}^*\right\|^p\right] \leq c_{\sigma^2}^p\sqrt{m}^p\left(1 + \sqrt{\frac{n}{m}}\log^{3/2} n\log^{1/2} m + \sqrt{\frac{n}{m}}\sqrt{p}\right)^p,$$

where c_{σ^2} is some constant only depending only on σ^2. Therefore, using the fact that $L_h = C'_{\sigma^2}\left\|\boldsymbol{R}_{[1:n]}\boldsymbol{C}_{\boldsymbol{g}^1}^*\right\| \|\boldsymbol{C}_{\boldsymbol{x}}\|/m$ and $p^{1/p} \leq e^{1/e}$, we obtain

$$\begin{aligned}
\left(\mathbb{E}_{\boldsymbol{g}^1,\boldsymbol{g}^2}\left[\left\|\mathcal{Q}_{dec}^{\boldsymbol{H}\boldsymbol{x}^\perp}(\boldsymbol{g}^1, \boldsymbol{g}^2)\right\|^p\right]\right)^{1/p} &\leq C''_{\sigma^2}\|\boldsymbol{C}_{\boldsymbol{x}}\|\sqrt{\frac{n}{m}}\left(1 + \sqrt{\frac{n}{m}}\log^{3/2} n\log^{1/2} m + \sqrt{\frac{n}{m}}\sqrt{p}\right)\sqrt{p} \\
&= C''_{\sigma^2}\|\boldsymbol{C}_{\boldsymbol{x}}\|\left[\sqrt{\frac{n}{m}}\left(1 + \sqrt{\frac{n}{m}}\log^{3/2} n\log^{1/2} m\right)\sqrt{p} + \frac{n}{m}p\right],
\end{aligned}$$

where $C''_{\sigma^2} > 0$ is some constant depending only on σ^2. ■

Auxiliary Results. The following are some auxiliary results used in the main proof.

Lemma B.29 *Let $\xi_{\sigma^2}, \zeta_{\sigma^2}, \eta_{\sigma^2}$ and ν_{σ^2} be defined as (24.2.1) and (B.3.8), for $t \in \mathbb{C}$, we have*

$$\begin{aligned}
\mathbb{E}_{s\sim\mathcal{CN}(0,1)}\left[\xi_{\sigma^2}(t+s)\right] &= \xi_{\sigma^2+\frac{1}{2}}(t)\\
\mathbb{E}_{s\sim\mathcal{CN}(0,1)}\left[\eta_{\sigma^2}(t+s)\right] &= \zeta_{\sigma^2}(t)\\
\mathbb{E}_{s\sim\mathcal{CN}(0,1)}\left[\zeta_{\sigma^2}(s)\right] &= \frac{1}{2\sigma^2+1}\\
\mathbb{E}_{s\sim\mathcal{CN}(0,1)}\left[|t|^2\,\zeta_{\sigma^2}(s)\right] &= \frac{4\sigma^2+1}{(2\sigma^2+1)^2}\\
\mathbb{E}_{s\sim\mathcal{CN}(0,2)}\left[\eta_{\sigma^2}(s)\right] &= \frac{1}{2\sigma^2+1}\\
\mathbb{E}_{s\sim\mathcal{CN}(0,1)}\left[(t+s)\nu_{\sigma^2}(t+s)\right] &= t\zeta_{\sigma^2}(t).
\end{aligned}$$

Proof Let $s_r = \Re(s)$, $s_i = \Im(s)$ and $t_r = \Re(t)$, $t_i = \Im(t)$, by definition, we observe

$$\begin{aligned}
&\mathbb{E}_{s\sim\mathcal{CN}(0,1)}\left[\xi_{\sigma^2}(t+s)\right]\\
&= \frac{1}{2\pi\sigma^2}\frac{1}{\pi}\int_s \exp\left(-\frac{|t+s|^2}{2\sigma^2}\right)\exp\left(-|s|^2\right)ds\\
&= \frac{1}{2\pi^2\sigma^2}\int_{s_r=-\infty}^{+\infty}\exp\left(-\frac{(s_r+t_r)^2}{2\sigma^2}-s_r^2\right)ds_r\int_{s_i=-\infty}^{+\infty}\exp\left(-\frac{(s_i+t_i)^2}{2\sigma^2}-s_i^2\right)ds_i\\
&= \frac{1}{2\pi\left(\sigma^2+1/2\right)}\exp\left(-\frac{|t|^2}{2\left(\sigma^2+1/2\right)}\right) \;=\; \xi_{\sigma^2+\frac{1}{2}}(t).
\end{aligned}$$

Thus, by definition of η_{σ^2} and ζ_{σ^2}, we have

$$\mathbb{E}_{s\sim\mathcal{CN}(0,1)}\left[\eta_{\sigma^2}(t+s)\right] = 1-2\pi\sigma^2\mathbb{E}_{s\sim\mathcal{CN}(0,1)}\left[\xi_{\sigma^2-1/2}(t+s)\right] = 1-2\pi\sigma^2\xi_{\sigma^2}(t) = \zeta_{\sigma^2}(t).$$

For $\mathbb{E}_{t\sim\mathcal{CN}(0,1)}\left[\zeta_{\sigma^2}(t)\right]$, we have

$$\begin{aligned}
\mathbb{E}_{t\sim\mathcal{CN}(0,1)}\left[\zeta_{\sigma^2}(t)\right] &= 1-2\pi\sigma^2\mathbb{E}_{t\sim\mathcal{CN}(0,1)}\left[\xi_{\sigma^2}(t)\right]\\
&= 1-\mathbb{E}_{t\sim\mathcal{CN}(0,1)}\left[\exp\left(-\frac{|t|^2}{2\sigma^2}\right)\right]\\
&= 1-\frac{2\sigma^2}{2\sigma^2+1} = \frac{1}{1+2\sigma^2}.
\end{aligned}$$

For $\mathbb{E}_{t\sim\mathcal{CN}(0,1)}\left[|t|^2\zeta_{\sigma^2}(t)\right]$, we observe

$$\begin{aligned}\mathbb{E}_{t\sim\mathcal{CN}(0,1)}\left[|t|^2\zeta_{\sigma^2}(t)\right] &= \frac{1}{\pi}\int_t |t|^2\left[1-\exp\left(-\frac{|t|^2}{2\sigma^2}\right)\right]\exp\left(-|t|^2\right)dt\\ &= \mathbb{E}_{t\sim\mathcal{CN}(0,1)}\left[|t|^2\right]-\frac{1}{\pi}\int_t |t|^2\exp\left(-\frac{2\sigma^2+1}{2\sigma^2}|t|^2\right)dt\\ &= 1-\frac{2\sigma^2}{2\sigma^2+1}\mathbb{E}_{t\sim\mathcal{CN}\left(0,\frac{2\sigma^2}{2\sigma^2+1}\right)}\mathbb{E}\left[|t|^2\right]\\ &= 1-\left(\frac{2\sigma^2}{2\sigma^2+1}\right)^2 \;=\; \frac{4\sigma^2+1}{(2\sigma^2+1)^2}.\end{aligned}$$

In addition, by using the fact that $\mathbb{E}_{s\sim\mathcal{CN}(0,1)}\left[\xi_{\sigma^2}(t+s)\right]=\xi_{\sigma^2+\frac{1}{2}}(t)$, we have

$$\mathbb{E}_{t\sim\mathcal{CN}(0,2)}\left[\eta_{\sigma^2}(t)\right]=\mathbb{E}_{t_1,t_2\sim_{i.i.d.}\mathcal{CN}(0,1)}\left[\eta_{\sigma^2}(t_1+t_2)\right]=\mathbb{E}_{t_1\sim\mathcal{CN}(0,1)}\left[\zeta_{\sigma^2}(t_1)\right]=\frac{1}{1+2\sigma^2}$$

For the last equality, first notice that

$$\begin{aligned}\mathbb{E}_{s\sim\mathcal{CN}(0,1)}\left[s\xi_{\sigma^2}(t+s)\right] &= \frac{1}{\pi}\int_s s\frac{1}{2\pi\sigma^2}\exp\left(-\frac{|t+s|^2}{2\sigma^2}\right)\exp\left(-|s|^2\right)ds\\ &= \frac{1}{2\pi^2\sigma^2}\exp\left(-\frac{|t|^2}{1+2\sigma^2}\right)\int_s s\exp\left(-\frac{1+2\sigma^2}{2\sigma^2}\left|s+\frac{t}{1+2\sigma^2}\right|^2\right)ds\\ &= \frac{1}{2\pi^2\sigma^2}\exp\left(-\frac{|t|^2}{1+2\sigma^2}\right)\times 2\pi\frac{\sigma^2}{1+2\sigma^2}\times\frac{-t}{1+2\sigma^2}\\ &= \frac{-t}{\pi\left(1+2\sigma^2\right)^2}\exp\left(-\frac{|t|^2}{1+2\sigma^2}\right) \;=\; \frac{-t}{1+2\sigma^2}\xi_{\sigma^2+\frac{1}{2}}(t).\end{aligned}$$

Therefore, we have

$$\begin{aligned}\mathbb{E}_{s\sim\mathcal{CN}(0,1)}\left[(t+s)\xi_{\sigma^2-\frac{1}{2}}(t+s)\right] &= t\mathbb{E}_{s\sim\mathcal{CN}(0,1)}\left[\xi_{\sigma^2-\frac{1}{2}}(t+s)\right]+\mathbb{E}_{s\sim\mathcal{CN}(0,1)}\left[s\xi_{\sigma^2-\frac{1}{2}}(t+s)\right]\\ &= t\xi_{\sigma^2}(t)-\frac{t}{2\sigma^2}\xi_{\sigma^2}(t) \;=\; \frac{2\sigma^2-1}{2\sigma^2}t\xi_{\sigma^2}(t).\end{aligned}$$

Using the result above, we observe

$$\begin{aligned}\mathbb{E}_{s\sim\mathcal{CN}(0,1)}\left[(t+s)\nu_{\sigma^2}(t+s)\right] &= t-\frac{4\pi\sigma^4}{2\sigma^2-1}\mathbb{E}_{s\sim\mathcal{CN}(0,1)}\left[(t+s)\xi_{\sigma^2-\frac{1}{2}}(t+s)\right]\\ &= t\left(1-2\pi\sigma^2\xi_{\sigma^2}(t)\right) \;=\; t\zeta_{\sigma^2}(t).\end{aligned}$$

■

Lemma B.30 *Let $\boldsymbol{g} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$, and $\boldsymbol{M}(\boldsymbol{g})$, $\boldsymbol{H}(\boldsymbol{g})$ be defined as (B.3.2) and (B.3.6), we have*

$$\mathbb{E}_{\boldsymbol{g}}\left[\boldsymbol{M}(\boldsymbol{g})\right] = \boldsymbol{P}_{\boldsymbol{x}^\perp} + \frac{1+4\sigma^2}{1+2\sigma^2}\boldsymbol{x}\boldsymbol{x}^*, \qquad \mathbb{E}_{\boldsymbol{g}}\left[\boldsymbol{H}(\boldsymbol{g})\right] = \boldsymbol{P}_{\boldsymbol{x}^\perp}.$$

Proof By Lemma B.29 and suppose $\boldsymbol{C}_{\boldsymbol{g}}\boldsymbol{R}_{[1:n]}^\top = \begin{bmatrix} \boldsymbol{g}_1^* \\ \dots \\ \boldsymbol{g}_m^* \end{bmatrix}$, we observe

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{M}\right] &= \frac{2\sigma^2+1}{m}\sum_{k=1}^m \mathbb{E}\left[\zeta_{\sigma^2}(\boldsymbol{g}_k^*\boldsymbol{x})\boldsymbol{g}_k\boldsymbol{g}_k^*\right] \\
&= \frac{2\sigma^2+1}{m}\sum_{k=1}^m \left\{\mathbb{E}\left[\zeta_{\sigma^2}(\boldsymbol{g}_k^*\boldsymbol{x})\right]\mathbb{E}\left[\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{g}_k\boldsymbol{g}_k^*\boldsymbol{P}_{\boldsymbol{x}^\perp}\right] + \mathbb{E}\left[\zeta_{\sigma^2}(\boldsymbol{g}_k^*\boldsymbol{x})\boldsymbol{P}_{\boldsymbol{x}}\boldsymbol{g}_k\boldsymbol{g}_k^*\boldsymbol{P}_{\boldsymbol{x}}\right]\right\} \\
&= \boldsymbol{P}_{\boldsymbol{x}^\perp} + \frac{2\sigma^2+1}{m}\boldsymbol{x}\boldsymbol{x}^*\sum_{k=1}^m \mathbb{E}\left[\zeta_{\sigma^2}\left(\boldsymbol{g}_k^*\boldsymbol{x}\right)\left|\boldsymbol{g}_k^*\boldsymbol{x}\right|^2\right] \\
&= \boldsymbol{P}_{\boldsymbol{x}^\perp} + \frac{4\sigma^2+1}{2\sigma^2+1}\boldsymbol{x}\boldsymbol{x}^*.
\end{aligned}$$

Thus, we have

$$\mathbb{E}\left[\boldsymbol{H}\right] = \boldsymbol{P}_{\boldsymbol{x}^\perp}\mathbb{E}\left[\boldsymbol{M}\right]\boldsymbol{P}_{\boldsymbol{x}^\perp} = \boldsymbol{P}_{\boldsymbol{x}^\perp}\left[\boldsymbol{P}_{\boldsymbol{x}^\perp} + \frac{4\sigma^2+1}{2\sigma^2+1}\boldsymbol{x}\boldsymbol{x}^*\right]\boldsymbol{P}_{\boldsymbol{x}^\perp} = \boldsymbol{P}_{\boldsymbol{x}^\perp}$$

■

Lemma B.31 *Let $\boldsymbol{g} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$ and $\widetilde{\boldsymbol{g}} \sim \mathcal{CN}(\boldsymbol{0}, 2\boldsymbol{I})$, for any positive integer $p \geq 1$, we have*

$$\begin{aligned}
\left\|1 - \frac{1+2\sigma^2}{m}\left\langle \boldsymbol{1}, \zeta_{\sigma^2}\left(\boldsymbol{g} \circledast \boldsymbol{x}\right)\right\rangle\right\|_{L^p} &\leq \frac{3}{\sqrt{m}}\frac{\left(2\sigma^2+1\right)}{\sigma}\left\|\boldsymbol{C}_{\boldsymbol{x}}\right\|\sqrt{p}, \\
\left\|1 - \frac{1+2\sigma^2}{m}\left\langle \boldsymbol{1}, \eta_{\sigma^2}\left(\widetilde{\boldsymbol{g}} \circledast \boldsymbol{x}\right)\right\rangle\right\|_{L^p} &\leq \frac{3}{\sqrt{m}}\frac{\sigma^2\left(2\sigma^2+1\right)}{\left(\sigma^2-\frac{1}{2}\right)^{3/2}}\left\|\boldsymbol{C}_{\boldsymbol{x}}\right\|\sqrt{p}.
\end{aligned}$$

Proof Let $h(\boldsymbol{g}) = \frac{1+2\sigma^2}{m}\left\langle \boldsymbol{1}, \zeta_{\sigma^2}\left(\boldsymbol{g} \circledast \boldsymbol{x}\right)\right\rangle - 1$ and let $h'(\widetilde{\boldsymbol{g}}) = \frac{1}{m}\left\langle \boldsymbol{1}, \eta_{\sigma^2}\left(|\widetilde{\boldsymbol{g}} \circledast \boldsymbol{x}|\right)\right\rangle$, by Lemma B.29, we know that

$$\mathbb{E}_{\boldsymbol{g}}\left[h(\boldsymbol{g})\right] = 0, \qquad \mathbb{E}_{\widetilde{\boldsymbol{g}}}\left[h'(\widetilde{\boldsymbol{g}})\right] = 0.$$

And for an independent copy $\boldsymbol{g}'$ of $\boldsymbol{g}$, we have

$$\begin{aligned}
\left|h(\boldsymbol{g}) - h(\boldsymbol{g}')\right| &\leq \frac{1+2\sigma^2}{m}\left|\left\langle \boldsymbol{1}, \exp\left(-\frac{1}{2\sigma^2}\left|\boldsymbol{g} \circledast \boldsymbol{x}\right|^2\right) - \exp\left(-\frac{1}{2\sigma^2}\left|\boldsymbol{g}' \circledast \boldsymbol{x}\right|^2\right)\right\rangle\right| \\
&\leq \frac{1+2\sigma^2}{m}\left\|\exp\left(-\frac{1}{2\sigma^2}\left|\boldsymbol{g} \circledast \boldsymbol{x}\right|^2\right) - \exp\left(-\frac{1}{2\sigma^2}\left|\boldsymbol{g}' \circledast \boldsymbol{x}\right|^2\right)\right\|_1
\end{aligned}$$

$$\leq \frac{1+2\sigma^2}{\sqrt{m}\sigma} \left\| \boldsymbol{C}_{\boldsymbol{x}}(\boldsymbol{g}-\boldsymbol{g}') \right\| \leq \frac{1+2\sigma^2}{\sqrt{m}\sigma} \left\| \boldsymbol{C}_{\boldsymbol{x}} \right\| \left\| \boldsymbol{g}-\boldsymbol{g}' \right\|,$$

where we used the fact that $\exp\left(-\frac{x^2}{2\sigma^2}\right)$ is $\frac{1}{\sigma}e^{-1/2}$-Lipschitz. By applying Gaussian concentration inequality in Lemma B.3, we have

$$\mathbb{P}\left(|h(\boldsymbol{g})| \geq t\right) = \mathbb{P}\left(\left| \frac{1+2\sigma^2}{m} \left\langle \mathbf{1}, \zeta_{\sigma^2}\left(|\boldsymbol{g} \circledast \boldsymbol{x}|\right)\right\rangle - 1 \right| \geq t\right) \leq \exp\left(-\frac{\sigma^2 m t^2}{2\left(2\sigma^2+1\right)^2 \left\|\boldsymbol{C}_{\boldsymbol{x}}\right\|^2}\right),$$

for any scalar $t \geq 0$. Thus, we can see that $h(\boldsymbol{g})$ is a centered $\frac{\left(\sigma^2+1\right)^2\|\boldsymbol{C}_{\boldsymbol{x}}\|^2}{\sigma^2 m}$-subgaussian random variable, by Lemma B.5, we know that for any positive integer $p \geq 1$

$$\|h(\boldsymbol{g})\|_{L^p} \leq 3\frac{\left(2\sigma^2+1\right)\|\boldsymbol{C}_{\boldsymbol{x}}\|}{\sigma\sqrt{m}}\sqrt{p},$$

as desired. For $h'(\widetilde{\boldsymbol{g}})$, we can obtain the result similarly. ■