Structured Signal Recovery from Nonlinear Measurements with Applications in Phase Retrieval and Linear Classification

Thesis by
Fariborz Salehi

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

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2021
Defended March 15th, 2021

Fariborz Salehi
ORCID: 0000-0002-9679-1016

To my beloved parents.

Humans think in stories, and we try to make sense of the world by telling stories.

- Yuval Noah Harari

ACKNOWLEDGEMENTS

As I look back on the journey that brought me to where I am standing today, remembering both the joyful moments and challenges along the way, I take it upon myself to recognize the support of individuals who accompanied me in this path. I would like to thank companions whose presence, help, and guidance through this process allowed me to continue working despite difficulties. Without them, reaching this stage and seeing years of work coming to fruition would have been much more difficult or perhaps altogether infeasible.

I would like to extend my deepest gratitude to my advisor, professor Babak Hassibi. His extensive knowledge, technical intuition, and problem-solving skills have played an important role in the progress of research projects presented in this thesis. Babak taught me how to always have the big picture in mind while working on the technical details of a mathematical problem. I would definitely miss our one-on-one and group meetings and discussions that opened my eyes and greatly enhanced my knowledge of topics ranging from control theory and signal processing, to machine learning, and information theory.

I would also like to thank other members of my candidacy and defense examinations' committees: professors Venkat Chandrasekaran, Palghat P. Vaidyanathan, Victoria Kostina, Joel Tropp, and Anima Anandkumar. I learned greatly from Joel and Venkat as I took their courses on advanced optimization, high-dimensional probability and statistics, which provided me with a solid base and profound understanding of these fundamental subjects. I also had the opportunity to serve as the teaching assistant for the graduate course in optimization under Venkat, which I always remember as a pleasing and valuable teaching experience.

Among the valuable lessons that I learned in grad school, is the significance of being a team player and collaborating with other people. I was fortunate to have amazing collaborators who provided me with opportunities to learn and grow as a researcher. Kishore Jaganathan mentored me in my first research project and introduced me to the the phase retrieval problem and its literature. I especially appreciate how he also patiently helped in writing and publishing my first research paper as a graduate student. My greatest number of collaborations was done with a dear friend, Ehsan Abbasi. Together we published several papers, and won the Qualcomm Innovation Fellowship (QIF'18). Ehsan has always generously shared his knowledge with me, and has had significant contributions to advancement of my research projects with his problem-solving skills. As a brother in the academic world, Ehsan inspired me to work with perseverance and present high-caliber results. We share many great memories from these collaborations and beyond, and I am delighted to have

had a collaborator who made this journey a very smooth and joyful one. Last but not least, I am happy to have had the opportunity to work with my friend, Ahmed Douik, on the application of Riemannian optimization techniques to the phase retrieval problem.

Beyond the research topics presented in this thesis, I had opportunities to work on a number of projects with some of the greatest researchers and engineers from Caltech, and also form other institutions as I did my internships. I am very thankful to Kallista A. Bonawitz, Ali Hajimiri, Wael Halbawi, Marc Joye, Taylan Kargin, Parham Khial, Jakub Konecny, Madison Lee, H. Brendan McMahan, and Alexander White, who collaborated with me on these projects. Being a part of such strong teams gave me the chance to learn about new subjects and expand my knowledge beyond my own research focus, communicate better, enhance my programming skills, and overall become a more well-rounded researcher and engineer.

The staff and offices at Caltech have always been present to facilitate processes for students. I am thankful to Laura Flower Kim, Alice Sogomonian, and Katheryn G. McAnulty who advised me on how to resolve issues and overcome challenges as a graduate student. I also want to acknowledge the service and help provided by the administrative crew in the Electrical Engineering Department, Tanya M. Owen, Shirley Slattery, Catherine L. Pichotta, and Liliana Chavarria.

Family and friends have undoubtedly been a great source of inspiration and support for me. While visiting my home country was difficult due to travel and visa limitations, I was privileged to have relatives who treated me like immediate family. My aunts and uncles, Fati, Leila, Hossein, and Mansour, and my cousins, Reza, Mona, Faraz, Maryam, and Alborz, have constantly provided me with their love and guidance, and warmly welcomed me into their homes during the past six years. I am grateful that I never felt like a stranger in my new home because of their immense kindness.

A unique aspect of being at Caltech was meeting some of my brightest colleagues who became dear friends. They not only provided much emotional support, but also taught me invaluable lessons. Arian's generosity and kindness, Pooya's modesty, Parham's determination to success, Pooria's dedication to research, Omid's great sense of humor, and Peyman's optimism and composure, are examples of distinguished qualities that I will remember as exemplary to apply to my own life.

My very special appreciation is reserved for my fiance, Yeganeh Amini, who brightened my life by generously extending her kindness, love, and support towards me, and helping me become a better person. I am extremely excited about the new chapter of life that we recently started together.

Finally, I can not be more thankful to anyone but my family who have been the greatest source of motivation and support for me from 8000 miles away. My mother, Mehri, set an example for me

to passionately pursue my dreams and know that determination is the key to success. My father, Firouz, gave me the courage to take steps with confidence and stay strong in the face of challenges. My siblings, Maral and Farzin, never hesitated to show their support and care throughout these years. Despite the distances, I see them as greatest blessings of my life and cherish their emotional presence every second.

Nonlinear models are widely used in signal processing, statistics, and machine learning to model real-world applications. A popular class of such models is the single-index model where the response variable is related to a linear combination of dependent variables through a *link function*. In other words, if $\mathbf{x} \in \mathbb{R}^p$ denotes the input signal, the posterior mean of the generated output y has the form, $\mathbb{E}[y|\mathbf{x}] = \rho(\mathbf{x}^T\mathbf{w})$, where $\rho : \mathbb{R} \to \mathbb{R}$ is a known function (referred to as the link function), and $\mathbf{w} \in \mathbb{R}^p$ is the vector of unknown parameters. When $\rho(\cdot)$ is invertible, this class of models is called generalized linear models (GLMs). GLMs are commonly used in statistics and are often viewed as flexible generalizations of linear regression. Given n measurements (samples) from this model, $\mathcal{D} = \{(\mathbf{x}_i, y_i)|1 \leq i \leq n\}$, the goal is to estimate the parameter vector $\mathbf{w}$. While the model parameters are assumed to be unknown, in many applications these parameters follow certain structures (sparse, low-rank, group-sparse, etc.) The knowledge on this structure can be used to form more accurate estimators.

The main contribution of this thesis is to provide a precise performance analysis for convex optimization programs that are used for parameter estimation in two important classes of single-index models. These classes are: (1) phase retrieval in signal processing, and (2) binary classification in statistical learning.

The first class of models studied in this thesis is the phase retrieval problem, where the goal is to recover a discrete complex-valued signal from amplitudes of its linear combinations. Methods based on convex optimization have recently gained significant attentions in the literature. The conventional convex-optimization-based methods resort to the idea of lifting which makes them computationally inefficient. In addition to providing an analysis of the recovery threshold for the semidefinite-programming-based methods, this thesis studies the performance of a new convex relaxation for the phase retrieval problem, known as *phasemax*, which is computationally more efficient as it does not lift the signal to higher dimensions. Furthermore, to address the case of structured signals, regularized phasemax is introduced along with a precise characterization of the conditions for its perfect recovery in the asymptotic regime.

The next important application studied in this thesis is the binary classification in statistical learning. While classification models have been studied in the literature since 1950's, the understanding of their performance has been incomplete until very recently. Inspired by the maximum likelihood (ML) estimator in logistic models, we analyze a class of optimization programs that attempts to find the model parameters by minimizing an objective that consists of a loss function (which is often inspired by the ML estimator) and an additive regularization term that enforces our knowledge on

the structure. There are two operating regimes for this problem depending on the separability of the training data set $\mathcal{D}$. In the asymptotic regime, where the number of samples and the number of parameters grow to infinity, a phase transition phenomenon is demonstrated that happens at a certain over-parameterization ratio. We compute this phase transition for the setting where the underlying data is drawn from a Gaussian distribution.

In the case where the data is non-separable, the ML estimator is well-defined, and its attributes have been studied in the classical statistics. However, these classical results fail to provide reasonable estimate in the regime where the number of data points is proportional to the number of samples. One contribution of this thesis is to provide an exact analysis on the performance of the regularized logistic regression when the number of training data is proportional to the number of samples. When the data is separable (a.k.a. the interpolating regime), there exist multiple linear classifiers that perfectly fit the training data. In this regime, we introduce and analyze the performance of *"extended margin maximizers"* (EMMs). Inspired by the max-margin classifier, EMM classifiers simultaneously consider maximizing the margin and the structure of the parameter. Lastly, we discuss another generalization to the max-margin classifier, referred to as the robust max-margin classifier, that takes into account the perturbations by an adversary. It is shown that for a broad class of loss functions, gradient descent iterates (with proper step sizes) converge to the robust max-margin classifier.

[1] F. Salehi et al. "The Performance Analysis of Generalized Margin Maximizer (GMM) on Separable Data". In: *International Conference on Machine Learning (ICML)* (2020), pp. 8417–8426. URL: http://proceedings.mlr.press/v119/salehi20a.html.
F.S. was the primary contributor in the conception of the project, analysis of the mathematical problem, conducting numerical simulations, writing the manuscript, and correspondence with reviewers and editors.

[2] F. Salehi and B. Hassibi. "Robustifying Binary Classification to Adversarial Perturbation". In: *arXiv preprint* (2020). URL: https://arxiv.org/abs/2010.15391.
F.S. was the primary contributor in the conception of the project, analysis of the mathematical problem, conducting numerical simulations, writing the manuscript, and correspondence with reviewers and editors.

[3] A. Douik et al. "A Novel Riemannian Optimization Approach and Algorithm for Solving the Phase Retrieval Problem". In: *53rd Asilomar Conference on Signals, Systems, and Computers* (2019). DOI: 10.1109/IEEECONF44664.2019.9049040.
F.S. has participated in the analysis of the mathematical problem, conducting numerical simulations, and writing the manuscript.

[4] E. Abbasi et al. "Universality in Learning from Linear Measurements". In: *Advances in Neural Information Processing Systems (NeurIPS)* (2019). URL: https://proceedings.neurips.cc/paper/2019/hash/dffbb6efd376d8dbb22cdf491e481edc-Abstract.html.
F.S. has participated in the analysis of the mathematical problem, conducting numerical simulations, and writing the manuscript.

[5] F. Salehi et al. "The Impact of Regularization on High-dimensional Logistic Regression". In: *Advances in Neural Information Processing Systems (NeurIPS)* (2019). URL: https://proceedings.neurips.cc/paper/2019/hash/ab49ef78e2877bfd2c2bfa738e459bf0-Abstract.html.
F.S. was the primary contributor in the conception of the project, analysis of the mathematical problem, conducting numerical simulations, writing the manuscript, and correspondence with reviewers and editors.

[6] F. Salehi et al. "A Precise Analysis of PhaseMax in Phase Retrieval". In: *2018 IEEE International Symposium on Information Theory (ISIT)* (2018), pp. 976–980. DOI: 10.1109/ISIT.2018.8437494.
F.S. was the primary contributor in the conception of the project, analysis of the mathematical problem, conducting numerical simulations, writing the manuscript, and correspondence with reviewers and editors.

[7] F. Salehi et al. "Learning without the Phase: Regularized Phasemax Achieves Optimal Sample Complexity". In: *Advances in Neural Information Processing Systems (NeurIPS)*

(2018), pp. 8641–8652. URL: https://proceedings.neurips.cc/paper/2018/hash/b91f4f4d36fa98a94ac5584af95594a0-Abstract.html.
F.S. was the primary contributor in the conception of the project, analysis of the mathematical problem, conducting numerical simulations, writing the manuscript, and correspondence with reviewers and editors.

[8] F. Salehi et al. "Multiple Illumination Phaseless Super-resolution (MIPS) with Applications to Phaseless DoA Estimation and Diffraction Imaging". In: *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on.* (2017), pp. 3949–3953. DOI: 10.1109/ICASSP.2017.7952897.
F.S. was the primary contributor in the conception of the project, analysis of the mathematical problem, conducting numerical simulations, writing the manuscript, and correspondence with reviewers and editors.

Number *Page*

Number *Page*

INTRODUCTION

In this chapter, we provide an overview of the results presented in the thesis. The recovery of discrete signals from a number of their samples (or measurements) has become the main challenge in various disciplines, including communication theory and signal processing, parameter estimation in statistics and machine learning, analysis of financial data, and genome sequencing. This challenge mainly arises due to the unknown factors in the measurement system as well as the presence of loss and distortions. To address this challenge, there have been many attempts to understand, design, and even simplify the measurement systems, with the goal of having an analyzable model that captures the main aspects of the real-world phenomenon.

Once this mathematical description is available, the recovery problem reduces to tuning the parameters of the model such that it generates the best (possible) output when compared to the measured values. The latter problem has been studied in the optimization theory, where the best choice is translated into minimizing (or maximizing) an objective function that takes its values in an ordered field[1].

This thesis focuses on a specific class of nonlinear models where the model output is related to a linear combination of its inputs through a link function. Let $\mathbf{x} \in \mathbb{R}^p$ (or $\mathbf{x} \in \mathbb{C}^p$ for phase retrieval) denote the input signal, and y be the model output. The posterior mean of the generated output takes the following form,

$$\mathbb{E}[y|\mathbf{x}] = \rho(\mathbf{x}^T\mathbf{w}), \tag{1.1}$$

where $\rho : \mathbb{R} \to \mathbb{R}$ is a known function (referred to as the link function), and $\mathbf{w} \in \mathbb{R}^p$ (or $\mathbf{w} \in \mathbb{C}^p$ for phase retrieval) is the vector of unknown parameters. The goal is to estimate the unknown parameters, $\mathbf{w}$, Given n measurements (samples) from this model, $\mathcal{D} = \{(\mathbf{x}_i, y_i)|1 \le i \le n\}$.

When the link function is invertible, resulting models are called generalized linear models (GLMs). In statistics literature, GLMs are viewed as generalizations of linear regression with an additional flexibility of choosing a link function. The underlying assumptions for classical linear regression are normality, homoscedasticity, and linearity, i.e., the errors are normally distributed, the error

[1] an ordered field is a field together with a total ordering

variances are constant and independent of the mean, and the systematic effects combine additively. However, there are many situations where these assumptions are far from being satisfied. Therefore, GLMs have been introduced to extend the scope of linear models. The term was coined in 1972 by Nelder and Wedderburn [96], and their main idea was to formulate linear models for a transformation of the mean value while keeping the observations untransformed [90].

In this thesis, we investigate the performance of the solutions of optimization problems to recover the underlying parameters. Our main focus will be on the class of convex programs where the objective function and the constraint set are both convex. Due to certificates of optimality that accompany their solutions, these programs are often very appealing for theoretical analyses. Moreover, assuming the convex constraint set is efficiently described, there are numerical methods that can find the optimal solution with a total number of computations that is bounded by a polynomial function of the input dimension (see e.g. [77]).

In general, to recover a p-dimensional signal, one needs to acquire at least p pieces of information, i.e., the number of measurements needed is at least $O(p)$. However, in many applications in machine learning, signal processing, statistics, etc., the underlying signal has certain structure (sparse, low-rank, finite alphabet, etc.), opening of up the possibility of recovering it from a number of measurements smaller than the ambient dimension, i.e., $n < p$. Understanding the role of this structure and finding ways to incorporate it into the optimization framework has been a very active area of research in the past two decades. The conventional methods add a penalty (regularization) function which enforces our prior knowledge on the structure [132]. These methods have been successfully used in various applications. However, the theoretical understanding of their success has been quite a challenge, and it was achieved years later mainly by the emergence of the field "Compressed Sensing" [43].

Initial theoretical results focused on analyzing the performance of optimization programs for the recovery of sparse signals [26, 46, 30], where ℓ_1 norm has been used to induce the sparse solution. Also, efficient numerical algorithms, such as orthogonal matching pursuit [134], have been introduced for the problem of sparse signal recovery. Later on, using analogous techniques the problem of low-rank matrix recovery has been analyzed [107, 29], where nuclear (trace) norm ($\|\cdot\|_*$) was used as a convex surrogate for the rank minimization. Chandrasekaran et al. [33] introduced a unified framework of atomic norm minimization where structured signals can be written in terms of a linear combination of a few simple building blocks (the so-called "atoms".) As a consequence of such theoretical understandings, using (non-smooth) regularization functions became very common in numerous applications.

While these initial results provided great theoretical insights on the required number of measurements for signal recovery in linear inverse problems, the resulting upper bounds (i.e., sufficient recovery conditions) were often not tight. Hence, finding lower bounds (i.e., necessary conditions) on the required number of measurements has become the next challenge. Sharp results on the recovery threshold of structured signal recovery in linear inverse problems have been first derived by Stojnic [121] and Amelunxen et al. [7]. Also, around the same time sharp theoretical results on least-sqaures with ℓ_1 regularization (a.k.a. LASSO) have been studied [13, 45]. More recently, sharp analyses for more general class of loss functions and regularizers (M-estimators for linear measurements) have been provided [42, 129].

Inspired by these results, in this thesis we study the precise performance of convex optimization problems for signal recovery in two specific examples of nonlinear models. Our theoretical assumptions are similar to the ones used in the analyses of linear inverse problems. However, the loss functions and the resulting optimizations have been formulated based on the measurement schemes in each application. The two classes of problems that will be extensively studied in the remaining of this thesis are: (1) structured signal recovery in phase retrieval, and (2) linear classification with structured parameters. Each of these applications falls into the category of single-index models introduced earlier. For the phase retrieval problem, it is often assumed that the input vector $\mathbf{x}$ and the parameter vector $\mathbf{w}$ are both complex-valued, and the link function is the absolute value function, $\rho(z) = |z|$. For binary classification the output, y, is the class label, and the link function determines the probability of the output being +1. In this case there are multiple choices for the link function, the most popular of which is the sigmoid function $\rho(u) = \frac{1}{1+e^{-u}}$. We provide more detailed explanation of the results presented in the thesis in the next section.

1.1 Contributions and Organization

The technical contents of the thesis are divided into two main parts, where in each part we study one of the problem classes introduced above.

Structured signal recovery in phase retrieval

The fundamental problem of recovering a signal from magnitude-only measurements is known as *phase retrieval*. It has a rich history and occurs in many areas in engineering and applied sciences such as medical imaging [6], X-ray crystallography [93], astronomical imaging [52], and optics [144]. In most of these cases, measuring the phase is either expensive or even infeasible. For instance, in some optical settings, detection devices like CCD cameras and photosensitive films cannot measure the phase of a light wave and instead measure the photon flux. Due to the loss of important phase

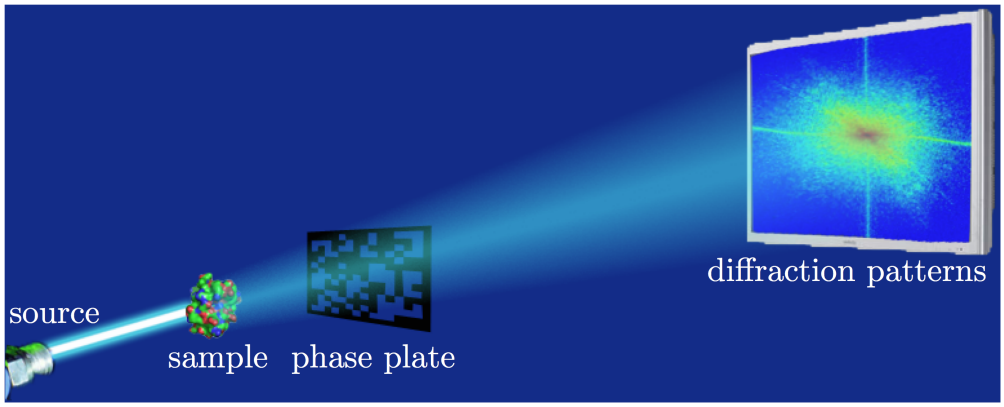


Figure 1.1: An example of a setup for phase retrieval using masks (courtesy of [23]). The phase plate applied after the sample modulates the spectrum.

information, signal reconstruction from magnitude-only measurements can be quite challenging. Therefore, despite a variety of proposed methods and analysis frameworks, phase retrieval still faces fundamental theoretical and algorithmic challenges.

In **Chapter 2**, the phase retrieval problem is introduced mathematically. Consequently, we provide discussions on challenges (due to ill-posedness) in signal recovery as well as commonly-used methods to solve this problem. While the conventional methods mainly focus on solving the original non-convex formulation of the phase retrieval, recently convex methods have gained significant attention to solve this problem. The first convex-relaxation-based methods were based on semidefinite programs (SDPs) [27, 25] and resorted to the idea of *lifting* [8, 22, 68, 115] the signal from a vector to a matrix to linearize the quadratic constraints. After introducing this convex formulation, known as PhaseLift, we focus on a more efficient optimization algorithm for solving the phase retrieval problem. For this, we define a Riemannian manifold of the points that satisfy phaseless Fourier measurements (this manifold is referred to fixed norms manifold). By analyzing the first and second order geometry of this manifold, a novel approach based on Riemannian gradient is proposed. Numerical simulations demonstrate that this approach outperforms the others in speed and accuracy.

Chapter 3 investigates the performance of signal recovery by solving an SDP for the Fourier phase retrieval, where the measurement vectors are rows of the discrete Fourier transform. We further assume that only low-frequency measurements are available to us (the problem of signal recovery from low-frequency measurements is known as super-resolution). The results presented in this chapter provide a flexible measurement scheme using masks, under which the signal recovery is guaranteed through solving the SDP. The flexible masks design can actually be implemented in

real-world applications. Figure 1.1, courtesy of [23], shows an example of modulating the signal in an X-ray imaging setting. We provide a discussion in this chapter on how to implement the proposed masking scheme in two applications, coherent diffraction imaging (CDI) and direction of arrival estimation.

Chapter 4 investigates the recovery threshold for the (real-valued) signal through solving an SDP (PhaseLift) in a setting where the measurements are drawn from a sub-Gaussian distribution. We analyze this problem as a special example of low-rank matrix recovery from quadratic measurements. The recovery threshold is established via a universality result that demonstrates equivalence to another problem where the measurements are independently drawn from an isotropic Gaussian distribution.

While the convex nature of their formulation makes them appealing for theoretical analysis, semidefinite relaxation squares the number of unknowns which makes these algorithms computationally inefficient, especially in large systems. Therefore, multiple researchers attempted to find other alternatives to these methods. We should also note that methods based on non-convex optimization are often complex for precise theoretical analysis and recovery guarantees.

In **Chapter 5**, we focus on analyzing a recently proposed convex-optimization-formulation for the complex phase retrieval problem known as *PhaseMax* where the constraint set is obtained by relaxing the non-convex equality constraints in the original phase retrieval problem to inequality constraints. Our results in this chapter provide the first exact analysis of the phase transition of (complex-valued) PhaseMax. Consequently, **Chapter 6** addresses the problem of structured signal recovery by introducing *regularized PhaseMax* and analyzing its performance. When the measurement matrix has i.i.d. Gaussian entries, it is shown that this method is indeed order-wise optimal, allowing perfect recovery from a number of phaseless measurements that is only a constant factor away from the optimal number of measurements required when phase information is available.

Binary classification with structured parameters

Machine learning models have been very successful in many applications, ranging from spam detection, face and pattern recognition, to the analysis of genome sequencing and financial markets. However, despite this indisputable success, our knowledge on why the various machine learning methods exhibit the performances they do is still at a very early stage. To make this gap between the theory and the practice narrower, researchers have recently begun to revisit simple machine learning models with the hope that understanding their performance will lead the way to understanding the performance of more complex machine learning methods.

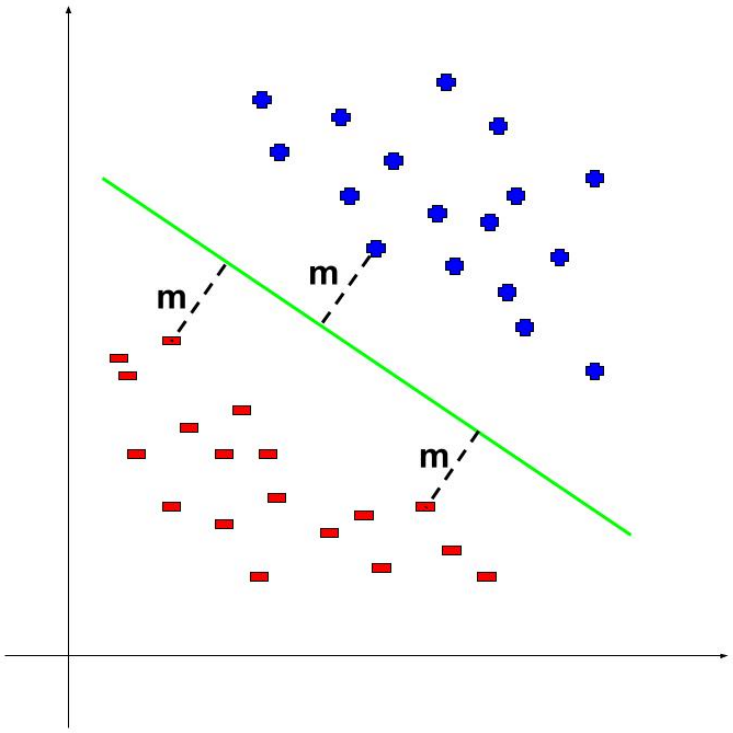


Figure 1.2: An example of a separable data set with the hyperplane corresponding to the max-margin classifier. The points that are closest to the hyperplane are called *support vectors*.

Linear classification is an important building block for most modern machine learning models. Researchers have studied this problem since the 1950's, with the goal of finding "optimal" parameters of the model that separates the two classes of data. In **Chapter 7**, we mathematically set up the problem by showing that the parameters of the classifier can be derived by solving an optimization problem consisting of a loss function and an additive regularization term, where the loss function is often inspired by the maximum-likelihood estimator, and the regularization term enforces the structure of the vector of parameters. This optimization problem exhibits two different behaviors depending on the separability of the training data set. Our results in this chapter give the asymptotic condition for the separability of the training data set when the data points are drawn from a Gaussian distribution.

After characterizing the exact phase transition which separates the problem into two different regimes of operations, we investigate the performance in each case. In **Chapter 8**, we investigate the performance of the solution to the optimization problem when the data set in inseparable. In this regime we form the regularized logistic regression and characterize the performance of its unique solution.

In **Chapter 9**, we study the behavior of the optimization when the data is separable. In this interpolating regime, there are multiple classifiers that perfectly fit the training data. By studying their generalization error, Vapnik has provided an upper bound which is inversely proportional to the minimum distance of the points to the separating hyperplane (a.k.a. the margin). Therefore, the max-margin classifier has been introduced as the "optimal" classifier. Figure 1.2 shows an example of a max-margin classifier on a separable 2-D data set. Inspired by the max-margin classifier, we introduce the **E**xtended **M**argin **M**aximizer (**EMM**) which takes into account the structure of the underlying parameter as well as the minimum distance of the data points to the separating hyperplane (a.k.a. the margin). We provide sharp asymptotic results on various performance measures (such as the generalization error) of EMMs and show that an appropriate choice of the potential function can in fact improve the resulting estimator.

Finally, in **Chapter 10** we introduce a new classifier, referred to as the *robust max-margin* classifier which incorporates the presence of adversarial perturbations. We show that the proposed classifier is the solution to a saddle-point optimization problem. Our main result in this chapter establishes that for a broad class of loss functions, gradient descent algorithms (with properly-tuned step sizes) converge to the robust max-margin classifier.

1.2 Notations

We gather here the basic notations that are used throughout this writing.

Bold face lower case letters are reserved for vectors and bold face upper case letters are used for matrices. For a vector $\mathbf{v}$, $\mathbf{v}^T$ is its transpose, v_i denotes its i^{th} entry and $\|\mathbf{v}\|_p$ is its l_p norm, where we often drop the subscript for $p = 2$. For a scalar $t \in \mathbb{R}$, $(t)_+ = \max(t, 0)$ denotes its positive part, and $\text{SIGN}(t)$ indicates its sign. The set of symmetric (or Hermitian) matrices are denoted by $\mathbb{S}^n$, and $\text{tr}(\cdot)$ denotes the trace of a square matrix (i.e., sum of its diagonal entries). $\mathbf{I}_d$ represents the identity matrix in dimension d. $\sigma_{\max}(\mathbf{M})$ denotes the maximum singular value of the matrix $\mathbf{M}$. $\mathbf{0}_d$ and $\mathbf{1}_d$ respectively represent the all-one and all-zero vectors in dimension d. We use calligraphy letters for sets. For set $\mathcal{S}$, $\text{cone}(\mathcal{S})$ is the closed conical hull of $\mathcal{S}$.

For a complex number $c \in \mathbb{C}$, the notation $\Re(c) = \frac{1}{2}(c + c^*)$ represents the real part of c. Similarly, the symbol $\Im(c) = \frac{1}{2\mathrm{i}}(c - c^*)$ refers to the imaginary part of c wherein i denotes the imaginary unit, i.e., $\mathrm{i}^2 = -1$. We also have $|z| = \sqrt{z_\Re{}^2 + z_\Im{}^2}$ and $\angle(z)$ denotes the phase of the complex scalar z. For a complex scalar, $z \in \mathbb{Z}$, $\bar{z}$ denotes its conjugate, and $(\cdot)^*$ is used to denote the conjugate transpose of a vector.

$X \sim p_X$ implies that the random variable X has a density p_X, and $\mathbb{E}\, X$ denotes its expected value. $\mathcal{N}(\mu, \sigma^2)$ denotes real Gaussian distribution with mean μ and variance σ^2. Likewise, $\mathcal{N}_{\mathbb{C}}(\mu, \sigma^2)$ refers to a *complex* Gaussian distribution with real and imaginary parts drawn independently from $\mathcal{N}_{\mathbb{C}}(\mu_{\Re}, \sigma^2/2)$ and $\mathcal{N}_{\mathbb{C}}(\mu_{\Im}, \sigma^2/2)$, respectively. $\mathcal{R}(2\sigma^2)$ denotes the Rayleigh distribution with the second moment equals to $2\sigma^2$. $\textsc{Rad}(p)$, for $p \in [0, 1]$, is the symmetric Bernoulli random variable which takes the value $+1$ with probability p and -1 with probability $1 - p$. $\xrightarrow{D}$ and $\xrightarrow{P}$ represent convergence in distribution and in probability, respectively.

A function $f(\cdot)$ is said to be L-smooth if its derivative, $f'(\cdot)$, is L-Lipschitz. $f : \mathbb{R}^d \to \mathbb{R}$ is called (invariantly) separable, when for all $\mathbf{w} \in \mathbb{R}^d$, $f(\mathbf{w}) = \sum_{i=1}^{d} \tilde{f}(w_i)$, for a real-valued function $\tilde{f}$. For a function $\Phi : \mathbb{R}^d \to \mathbb{R}$, the Moreau envelope associated with $\Phi(\cdot)$ is defined as,

$$M_{\Phi}(\mathbf{v}, t) = \min_{\mathbf{x} \in \mathbb{R}^d} \ \frac{1}{2t} ||\mathbf{v} - \mathbf{x}||^2 + \Phi(\mathbf{x}) \ , \tag{1.2}$$

and the proximal operator is the solution to this optimization, i.e.,

$$\text{Prox}_{t\Phi(\cdot)}(\mathbf{v}) = \arg\min_{\mathbf{x} \in \mathbb{R}^d} \ \frac{1}{2t} ||\mathbf{v} - \mathbf{x}||^2 + \Phi(\mathbf{x}) \ . \tag{1.3}$$

The function $\Phi(\cdot)$ is said to be locally-Lipschitz if for any $M > 0$, there exists a constant L_M, such that,

$$\forall \mathbf{u}, \mathbf{v} \in [-M, +M]^d, \ \ |\Phi(\mathbf{u}) - \Phi(\mathbf{v})| \leq L_M \, \|\mathbf{u} - \mathbf{v}\| \, . \tag{1.4}$$

Finally, for any vector $\mathbf{w} \in \mathbb{R}^p$, the binary classifier associated with $\mathbf{w}$ is defined as: $C_{\mathbf{w}} : \mathbb{R}^p \to \{\pm 1\}$, such that $C_{\mathbf{w}}(\mathbf{x}) = \text{Sign}(\mathbf{w}^T \mathbf{x})$.

Part I:

Structured Signal Recovery in Phase Retrieval

PHASE RETRIEVAL: CHALLENGES AND ALGORITHMS

In this chapter, we introduce and study the first application of the single-index models known as the phase retrieval. Phase retrieval emerges in many applications in engineering and applied sciences, where measuring the phase is expensive or altogether infeasible. We start by introducing the phase retrieval problem in Section 2.1. After mathematically setting up the problem, we discuss its ill-posedness and present the modern approaches to solve the problem based on imposing a prior (e.g. sparsity) or exploiting additional measurements. Consequently, in section 2.2 we discuss the recovery algorithms by first explaining the Gerchberg-Saxton (GS) algorithm [56] which is a conventional method based on alternating minimization. Consequently, we shift our attention to convex programs by introducing the PhaseLift method which is a convex-optimization formulation of the phase retrieval based on semidefinite programming [25].

In Sections 2.3, we suggest a novel Riemannian optimization approach for solving the Fourier phase retrieval problem by studying and exploiting the geometry of the problem to reduce the ambient dimension and derive extremely fast and accurate algorithms. We reformulate the problem as a constrained problem on novel Riemannian manifold, referred to as the fixed-norms manifold. Deriving the first-order geometry of this manifold in closed form allows the design of a highly efficient optimization algorithm which is presented in Section 2.4. Numerical simulations in Section 2.5 suggests that the proposed approach outperforms conventional optimization-based methods both in accuracy and convergence speed. The results presented in this section are available in the research paper [47] by Douik et al.[1], and some of the texts appear as it is in the publication.

2.1 Motivations and Problem Setup

The fundamental problem of recovering a signal from magnitude-only measurements is known as *phase retrieval*. This problem has a rich history and appears in many areas in engineering and applied physics, such as astronomical imaging [52], X-ray crystallography [93], medical imaging [6], and optics [144]. In most of these cases, measuring the phase is either expensive or even infeasible. For instance, in some optical settings, detection devices like CCD cameras and photosensitive films cannot measure the phase of a light wave and instead measure the photon flux.

[1] A. Douik, F. Salehi, and B. Hassibi. "A Novel Riemannian Optimization Approach and Algorithm for Solving the Phase Retrieval Problem." In: Proc. of the 53rd Asilomar Conference on Signals, Systems, and Computers, Asilomar, CA, USA. Vol. 1. 1. Nov. 2019, pp. 1962–1966.

Let $\mathbf{x}_0 \in \mathbb{C}^n$ denote the underlying signal. We consider the phase retrieval problem with the goal of recovering $\mathbf{x}_0$ from m magnitude-only measurements of the form,

$$b_i = |\mathbf{a}_i^\star \mathbf{x}_0|, \quad i = 1, \dots, m, \tag{2.1}$$

where we assume that $\{\mathbf{a}_i \in \mathbb{C}^n\}_{i=1}^m$ is the set of known measurement vectors. Originally, the phase retrieval problem has been introduced in applications such as coherent diffraction imaging and optics, where the measurements correspond to the Fourier transform of the underlying signal, i.e., the measurement vectors are the rows of the DFT matrix. In more recent applications, more general settings become feasible for the measurement vectors. As an example, [82] designed a measurement framework using a random dielectric metasurface diffuser (MD) where the MD can be designed to have scattering matrix, with certain properties.

Given the measurements, the phase retrieval problem can be formalized as the following optimization:

$$\begin{aligned} &\text{find} && \mathbf{x} \\ &\text{subject to:} && |\mathbf{a}_i^\star \mathbf{x}| = b_i \ , \quad 1 \le i \le m. \end{aligned} \tag{2.2}$$

Identifiability

We first note that there is a trivial ambiguity due to global phase change, which cannot be identified in the phase retrieval problem. To resolve this, one can assume, without loss of generality, that the first entry of the signal is real-valued.

When the number of measurements, m, is the same as the number of unknowns n (e.g. Fourier phase retrieval), the available data is highly incomplete. In fact, for any given Fourier magnitude, the Fourier phase can be chosen in an n-dimensional set, and distinct phases result in distinct signals. For Fourier phase retrieval, it is well known that the phase often contains more information than the Fourier magnitude. Therefore, the Fourier phase retrieval is a highly-illposed problem.

To compensate on this ill-posedness of the phase retrieval problem, and inspired by the recent developments in measurement technologies, researchers have investigated new approaches for the phase retrieval problem which can be categorized into the following two main streams:

(i) *Imposing prior information:* When the number of measurements are not enough to uniquely determine the underlying signal, one can often enforce certain structure(s) on the underlying signals. Imposing such structures reduces the "effective dimension" (a.k.a. degrees of freedom) of the underlying signal and reduces the ill-posedness of the phase retrieval problem. The most

popular example of a structure is the sparse structure, when it is assumed that the underlying signal is sparse (with few non-zero entries). Inspired by the recent advances in the area of compressed sensing [26, 43], researchers have recently analyzed recovery algorithms for the sparse phase retrieval.

(ii) *Additional measurements:* The abovementioned ill-posedness arises due to the fact that the number of Fourier measurements (n) is less than the number of $2n - 1$. Therefore, when no structure is present, having additional measurements is inevitable to uniquely identify the signal. However, when $m > n$, many of the measurement vectors can no longer form an orthonormal basis. There are two main approaches to introduce additional measurements:

a) Fourier measurements: A widely-used method to acquire additional Fourier measurements is to use multiple masks and measure the Fourier transform of the masked signal.

b) Random measurements: Another popular approach is to consider the setting in which the measurment vectors are drawn randomly from a distribution.

We will see examples of both of these approaches in the remaining of this chapter.

To conclude this section, we state the following result from Jaganathan et al. [68] on the identifiability of the signal in sparse phase retrieval.

Theorem 1 (Theorem 2.1 in [68])**.** *Let $\mathcal{S}_k$ represent the set of all k-sparse signals with aperiodic support, where $3 \leq k \leq n-1$. Almost all signals in $\mathcal{S}_k$ can be uniquely recovered.*

2.2 Recovery Algorithms

Phase Retrieval is a classical problem, and various algorithms have been proposed to tackle this problem in the identifiable regime. The conventional algorithms focus on the Fourier setting where the measurements are the magnitude of the Fourier transform of the underlying signal. The conventional methods focus on solving the non-convex optimization via iterative updates. The Gerchberg-Saxton [56] is a widely-used algorithm for Fourier phase retrieval in practice based on the alternating projections.

GS starts by adding n non-zero entries to the underlying signal $\mathbf{x}$. Define $\tilde{\mathbf{x}} \in \mathbb{C}^{2n}$ such that $\tilde{\mathbf{x}}_i = \mathbf{x}_i$ for $i = 1, 2, \ldots, n$, and $\tilde{\mathbf{x}}_i = 0$ for $i > n$. The GS considers the $2n$-DFT measurements of $\tilde{\mathbf{x}}$.

In a consequent work, Fienup [53] has improved the Gerchberg-Saxton algorithm by imposing additional constraint on the signal. Despite great success in practical application, rigorous theoretical analyses are not available for these methods.

Algorithm 1 GS Algorithm

Require:
$\mathbf{F} \in \mathbb{C}^{2n\times 2n}$: $2n$-DFT matrix
$\mathbf{b} = |\mathbf{F}\tilde{\mathbf{x}}|$: Magnitude of the discrete Fourier measurements
Ensure: $\mathbf{x} \in \mathbb{C}^n$: Estimate of the underlying signal
Initialize $\mathbf{x}^{(0)}$ randomly, $t \leftarrow 0$
while $t < T$ **do**
Compute the Fourier transform $\mathbf{y}^{(t)} = \mathbf{F}\mathbf{x}^{(t)}$
Impose measurement constraints $\tilde{\mathbf{y}}_i^{(t)} = \text{SIGN}(\mathbf{y}_i^{(t)})\ \mathbf{b}_i$, for $i = 1, 2, \ldots, 2n$.
Compute the inverse Fourier transform, $\mathbf{x}^{(t+1)} = \mathbf{F}^{-1}\tilde{\mathbf{y}}^{(t)}$
Set values equal to zero: $\mathbf{x}_i^{(t+1)} = 0$ for $i = n+1, \ldots, 2n$.
$t \leftarrow t + 1$
end while
return $\mathbf{x}^{(T)}[1:n]$

SDP-based methods

More recently, methods based on convex optimization have gained significant attentions to solve the phase retrieval problem. Due to the convex nature of their formulation, these algorithms usually have rigorous theoretical guarantees. These methods are mainly based on semidefinite programming by linearizing the resulting quadratic constraints using the idea of lifting [27, 57, 142, 143, 25, 11, 101, 8, 67].

By lifting the optimization variable, $\mathbf{x}$, one can rewrite the measurement in terms of the lifted variable, $\mathbf{X} = \mathbf{x}^\star\mathbf{x}$ as follows:

$$\mathbf{b}_i^2 = |\mathbf{a}_i^\star\mathbf{x}|^2 = \mathbf{a}_i^\star\mathbf{x}\mathbf{x}^\star\mathbf{a}_i = \text{tr}\big(\mathbf{a}_i^\star\mathbf{X}\mathbf{a}_i\big) = \text{tr}\big(\mathbf{X}(\mathbf{a}_i\mathbf{a}_i^\star)\big) = \text{tr}(\mathbf{X}\mathbf{A}_i),$$

where $\mathbf{A}_i \in \mathbb{S}^n$ is defined as $\mathbf{A}_i = \mathbf{a}_i\mathbf{a}_i^\star$. Using this lifted formulation, one can rewrite the phase retrieval problem as,

$$\begin{aligned} \underset{\mathbf{X}\in\mathbb{S}^n}{\text{find}} \quad & \mathbf{X} \\ \text{s.t.} \quad & \text{tr}\big(\mathbf{X}\mathbf{A}_i\big) = \mathbf{b}_i^2\ , \quad \text{for} \quad i = 1, 2, \ldots, m \\ & \text{rank}(\mathbf{X}) = 1 \\ & \mathbf{X} \succeq \mathbf{0}\ . \end{aligned} \tag{2.3}$$

Or, equivalently, it can be written as,

$$\begin{aligned} \min_{\mathbf{X}\in\mathbb{S}^n} \quad & \operatorname{rank}\big(\mathbf{X}\big) \\ \text{s.t.} \quad & \operatorname{tr}\big(\mathbf{X}\mathbf{A}_i\big) = \mathbf{b}_i^2 \ , \quad \text{for} \quad i = 1, 2, \ldots, m \\ & \mathbf{X} \succeq \mathbf{0} \ . \end{aligned} \tag{2.4}$$

Note that (2.3) and (2.4) are still non-convex optimization programs since $\operatorname{rank}(\cdot)$ is a non-convex function. The problem of finding a minimum rank solution among symmetric (Hermitian) matrices that satisfy linear constraints has been studied extensively. A promising approach considers a convex-surrogate for the $\operatorname{rank}(\cdot)$ function which is known as nuclear (or trace norm), and defined as:

$$||\mathbf{X}||_* = \sum_{i=1}^{n} \sigma_i(\mathbf{X}), \tag{2.5}$$

and for hermitian matrices $||\mathbf{X}||_* = \operatorname{tr}(\mathbf{X})$. Therefore, the following semidefinite program is derived by replacing the $\operatorname{rank}(\cdot)$ function with the nuclear norm.

$$\begin{aligned} \min_{\mathbf{X}\in\mathbb{S}^n} \quad & \operatorname{tr}(\mathbf{X}) \\ \text{s.t.} \quad & \operatorname{tr}\big(\mathbf{X}\mathbf{A}_i\big) = \mathbf{b}_i^2 \ , \quad \text{for} \quad i = 1, 2, \ldots, m \\ & \mathbf{X} \succeq \mathbf{0} \ . \end{aligned} \tag{2.6}$$

Finding the solution to the phase retrieval by solving the optimization problem (2.6) is often known as the *PhaseLift* method [25].

2.3 A Novel Approach Based on Riemannian Optimization for Solving the Phase Retrieval Problem

Despite the success of semidefinite programs in solving the phase retrieval problem and the theoretical guarantees and recovery thresholds that follow, these methods are often computationally expensive. Semidefinite relaxation squares the number of unknowns which makes these algorithms computationally complex, especially in large systems. This caveat makes these approaches intractable in real-world applications.

In many applications of the phase retrieval problem, a subset of phaseless measurements is obtained from an orthonormal basis. For example, the Fourier phase retrieval problem reconstructs a signal from phaseless measurements of its discrete Fourier transform. This particular structure of the phase retrieval problem allows its reformulation as a constrained optimization problem wherein the constraint set is represented by an orthonormal basis. In this section, we suggest exploiting the

problem structure to reduce the dimension of the problem and design fast recovery algorithms using Riemannian optimization techniques. To this end, we introduce a new manifold, referred to as the "fixed norms" manifold, which generalizes the complex sphere $\mathbb{S}^{n-1}$. The results presented in this section have been published in the research paper [47] by Douik et al.[2]

Some of the concepts in Riemannian geometry as well as optimization algorithms that are related to our analysis will be reviewed. For a thorough introduction to these concepts, readers are referred to the standard texts [16] for differential geometry, [83] for abstract manifold, [105] and [74] for Riemannian geometry, and [4] for optimization on matrix manifolds.

Setup

Let $\mathbf{x} \in \mathbb{C}^n$ be a complex vector of dimension n and assume that the m observations are obtained by $\sqrt{b_i} = |\mathbf{a}_i^* \mathbf{x}|$, $1 \leq i \leq m$ with the sensing vectors $\mathbf{a}_i \in \mathbb{C}^n$. Considering a smooth loss function ℓ, the phase retrieval problem can be formulated as

$$\min_{\mathbf{x} \in \mathbb{C}^n} \sum_{i=1}^{m} \ell(|\mathbf{a}_i^* \mathbf{x}|, \sqrt{b_i}) \ . \tag{2.7}$$

Without loss of generality, assume that the first k observations are obtained from an orthogonal basis, say the discrete Fourier transform for $k = n$. In other words, matrices $\mathbf{A}_i = \mathbf{a}_i \mathbf{a}_i^*, \ 1 \leq i \leq k$ are non-negative orthogonal projection matrices that collectively span the whole ambient space $\mathbb{C}^n$, i.e., $\mathbf{A}_i = \mathbf{A}_i^*$, $\mathbf{A}_i \mathbf{A}_j = \delta_{ij} \mathbf{A}_i$ and $\sum_{i=1}^{k} \mathbf{A}_i = \mathbf{I}_n$.

The Riemannian optimization formulation for the phase retrieval problem

The unconstrained optimization of the phase retrieval problem in (2.7) can be formulated as a constrained optimization as follows:

$$\min_{\mathbf{x} \in \mathbb{C}^n} \sum_{i=k+1}^{m} \ell(|\mathbf{a}_i^* \mathbf{x}|, \sqrt{b_i}) \tag{2.8a}$$

$$\text{s.t.} \ \ |\mathbf{a}_i^* \mathbf{x}| = \sqrt{b_i}, \ 1 \leq i \leq k \ . \tag{2.8b}$$

Clearly, the modulus equality constraint $|\mathbf{a}_i^* \mathbf{x}| = \sqrt{b_i}$ of (2.8) is equivalent to the quadratic constraint $\mathbf{x}^* \mathbf{a}_i \mathbf{a}_i^* \mathbf{x} = b_i^2$. Define the matrices $\mathbf{A}_i = \mathbf{a}_i \mathbf{a}_i^*, \ 1 \leq i \leq k$. From the previous assumptions on the system model, the set of matrices $\{\mathbf{A}_i\}_{i=1}^{k}$ are non-negative orthogonal projection matrices

[2] A. Douik, F. Salehi, and B. Hassibi. "A Novel Riemannian Optimization Approach and Algorithm for Solving the Phase Retrieval Problem." In: Proc. of the 53rd Asilomar Conference on Signals, Systems, and Computers, Asilomar, CA, USA. Vol. 1. 1. Nov. 2019, pp. 1962–1966.

that collectively span the whole ambient space $\mathbb{C}^n$. In other words, for all $1 \leq i, j \leq k$, we have $\mathbf{A}_i = \mathbf{A}_i^*$, $\mathbf{A}_i\mathbf{A}_j = \delta_{ij}\mathbf{A}_i$ and $\sum_{i=1}^k \mathbf{A}_i = \mathbf{I}_n$. Let $\mathcal{M}$ denote the set of solutions to the optimization problem (2.8), i.e., $\mathcal{M} = \{\mathbf{x} \in \mathbb{C}^n \mid \mathbf{x}^*\mathbf{A}_i\mathbf{x} = b_i,\ 1 \leq i \leq k\}$, called herein the *fixed norms* manifold. The optimization problem can then be expressed as,

$$\min_{\mathbf{x}\in\mathcal{M}} \sum_{i=k+1}^{m} \ell(|\mathbf{a}_i^*\mathbf{x}|, \sqrt{b_i}) \ . \tag{2.9}$$

Background on Riemannian manifold and optimization

A Riemannian manifold $(\mathcal{M}, g)$ is a real smooth manifold $\mathcal{M}$ embedded in the Euclidean space $\mathcal{E}$ and equipped with a Riemannian metric g. For a point $\mathbf{x} \in \mathcal{M}$, the tangent space is denoted by $\mathcal{T}_\mathbf{x}\mathcal{M}$. For tangent vectors $\xi_\mathbf{x}$ and $\eta_\mathbf{x}$ in $\mathcal{T}_\mathbf{x}\mathcal{M}$, the restriction of the Riemannian metric $g_\mathbf{x}$ to $\mathcal{T}_\mathbf{x}\mathcal{M}$ is denoted by $g_\mathbf{x}(\xi_\mathbf{x}, \eta_\mathbf{x}) = \langle\xi_\mathbf{x}, \eta_\mathbf{x}\rangle_\mathbf{x}$. Similarly, the norm of $\xi_\mathbf{x} \in \mathcal{T}_\mathbf{x}\mathcal{M}$ is defined and denoted by $||\xi_\mathbf{x}||_\mathbf{x} = \sqrt{\langle\xi_\mathbf{x}, \xi_\mathbf{x}\rangle_\mathbf{x}}$.

For a real and smooth function $f : \mathcal{M} \to \mathbb{R}$, the directional derivative of f at the point $\mathbf{x} \in \mathcal{M}$ in the direction $\xi_\mathbf{x} \in \mathcal{T}_\mathbf{x}\mathcal{M}$ is denoted by $\mathbf{D}f(\mathbf{x})[\xi_\mathbf{x}]$. The function that associates to each $\xi_\mathbf{x}$ the directional derivative $\mathbf{D}f(\mathbf{x})[\xi_\mathbf{x}]$ is called the indefinite directional derivative of f at $\mathbf{x}$. The Euclidean and Riemannian gradients of f at $\mathbf{x} \in \mathcal{M}$ are denoted by $\mathrm{Grad}\, f(\mathbf{x})$ and $\mathrm{grad}\, f(\mathbf{x})$, respectively. Similarly, the Euclidean and Riemannian Hessian of f at the point $\mathbf{x} \in \mathcal{M}$ in the direction $\xi_\mathbf{x} \in \mathcal{T}_\mathbf{x}\mathcal{M}$ are denoted by $\mathrm{Hess}\, f(\mathbf{x})[\xi_\mathbf{x}]$ and $\mathrm{hess}\, f(\mathbf{x})[\xi_\mathbf{x}]$, respectively. For a single variable function $\gamma(t)$, we use the shorthand notation $\dot{\gamma}(t)$ to denote the first order derivative $\frac{\delta\gamma(t)}{\delta t}$.

Given a Riemannian connection ∇ on $\mathcal{M}$ and an interval $\mathcal{I} \subseteq \mathbb{R}$ containing 0, a geodesic curve $\gamma : \mathcal{I} \to \mathcal{M}$ going through $\mathbf{x} \in \mathcal{M}$ in the direction $\xi_\mathbf{x} \in \mathcal{T}_\mathbf{x}\mathcal{M}$, i.e., $\gamma(0) = \mathbf{x}$ and $\dot{\gamma}(0) = \xi_\mathbf{x}$, is denoted by $\gamma_{\mathbf{x},\xi_\mathbf{x}}(t)$. The geodesic $\gamma_{\mathbf{x},\xi_\mathbf{x}}(t)$ defines the Exponential map $\mathrm{Exp}_\mathbf{x} : \mathcal{T}_\mathbf{x}\mathcal{M} \to \mathcal{M}$ by $\mathrm{Exp}_\mathbf{x}(\xi_\mathbf{x}) = \gamma_{\mathbf{x},\xi_\mathbf{x}}(1)$.

Properties of fixed norms manifold

Given a set of k non-negative and orthogonal $n \times n$ projection matrices $\{\mathbf{A}_i\}_{i=1}^k$ over the complex field $\mathbb{C}$, i.e., $\mathbf{A}_i \succeq \mathbf{0}$ and $\mathbf{A}_i\mathbf{A}_j = \delta_{ij}\mathbf{A}_i$ for all $1 \leq i, j \leq k$, satisfying $\sum_{i=1}^k \mathbf{A}_i = \mathbf{I}_n$ and k positive real numbers $\{b_i\}_{i=1}^k \in \mathbb{R}_{++}$, the fixed norms manifold is defined by

$$\mathcal{M} = \{\mathbf{x} \in \mathbb{C}^n \mid \mathbf{x}^*\mathbf{A}_i\mathbf{x} = b_i,\ 1 \leq i \leq k\} \ . \tag{2.10}$$

The linear approximation of the manifold at each point is known as the *tangent space*. The following result characterizes the tangent space of $\mathcal{M}$.

Lemma 1. *The set $\mathcal{M}$ is a well-defined real manifold of dimension $2n-k$ embedded in $\mathbb{R}^n \times \mathbb{R}^n$, which is isomorphic to $\mathbb{C}^n$, and whose tangent space at $\mathbf{x} \in \mathcal{M}$ is given by*

$$\mathcal{T}_{\mathbf{x}}\mathcal{M} = \left\{ \xi_{\mathbf{x}} \in \mathbb{C}^n \mid \Re\left(\xi_{\mathbf{x}}^* \mathbf{A}_i \mathbf{x}\right) = 0, 1 \leq i \leq k \right\}. \tag{2.11}$$

The restriction of the *real* Riemannian metric g to $\mathcal{T}_{\mathbf{x}}\mathcal{M}$ is defined by

$$\langle \xi_{\mathbf{x}}, \eta_{\mathbf{x}} \rangle_{\mathbf{x}} = \Re(\xi_{\mathbf{x}}^* \eta_{\mathbf{x}}) = \frac{1}{2}(\xi_{\mathbf{x}}^* \eta_{\mathbf{x}} + \eta_{\mathbf{x}}^* \xi_{\mathbf{x}}), \tag{2.12}$$

which turns $(\mathcal{M}, g)$ into a real smooth Riemannian manifold.

The result of Lemma 1 will be used to derive an efficient first-order iterative optimization method to solve the optimization problem (2.9). Tangent spaces play an important role in Riemannian optimization in the same fashion that derivatives of smooth functions play a crucial role in numerical optimization.

The *normal space* is the orthogonal complement of the tangent space with respect to the Riemannian metric. For the fixed norms manifold, the normal space has the form

$$\mathcal{N}_{\mathbf{x}}\mathcal{M} = \left\{ \eta_{\mathbf{x}} \in \mathbb{C}^n \mid \eta_{\mathbf{x}} = \sum_{i=1}^{k} \alpha_i \mathbf{A}_i \mathbf{x}, \ \{\alpha_i\}_{i=1}^{k} \in \mathbb{R} \right\}. \tag{2.13}$$

This is due to the fact that for any tangent vector $\xi_{\mathbf{x}} \in \mathcal{T}_{\mathbf{x}}\mathcal{M}$, we have the following,

$$\langle \xi_{\mathbf{x}}, \eta_{\mathbf{x}} \rangle_{\mathbf{x}} = \frac{1}{2} \sum_{i=1}^{k} \alpha_i \left(\xi_{\mathbf{x}}^* \mathbf{A}_i \mathbf{x} + \mathbf{x}^* \mathbf{A}_i \xi_{\mathbf{x}} \right) = \sum_{i=1}^{k} \alpha_i \Re\left(\xi_{\mathbf{x}}^* \mathbf{A}_i \mathbf{x}\right) = 0 \,,$$

where the last equality is from the (2.11). Proceeding onwards, we can now derive a closed-form for the orthogonal projections onto the normal space and the tangent space.

For an aribitrary vector $\mathbf{y} \in \mathbb{C}^n$, the projections onto the tangent space $\mathcal{T}_{\mathbf{x}}\mathcal{M}$ and the normal space $\mathcal{N}_{\mathbf{x}}\mathcal{M}$, respectively denoted by $\Pi_{\mathbf{x}}(\mathbf{y})$ and $\Pi_{\mathbf{x}}^{\perp}(\mathbf{y})$, are given by

$$\Pi_{\mathbf{x}}(\mathbf{y}) = \mathbf{y} - \sum_{i=1}^{k} \frac{1}{2b_i} (\mathbf{y}^* \mathbf{A}_i \mathbf{x} + \mathbf{x}^* \mathbf{A}_i \mathbf{y}) \mathbf{A}_i \mathbf{x}, \tag{2.14}$$

$$\Pi_{\mathbf{x}}^{\perp}(\mathbf{y}) = \sum_{i=1}^{k} \frac{1}{2b_i} (\mathbf{y}^* \mathbf{A}_i \mathbf{x} + \mathbf{x}^* \mathbf{A}_i \mathbf{y}) \mathbf{A}_i \mathbf{x} \,. \tag{2.15}$$

Another important concept in Riemannian manifolds is geodesics, which generalizes the concept of straight lines in a Euclidean space. Here we state the following lemma without proof which represents the geodesic curve of the fixed norms manifold.

Lemma 2. *The geodesic curve $\gamma_{\mathbf{x},\xi_\mathbf{x}} : \mathbb{R} \to \mathcal{M}$ going through $\mathbf{x} \in \mathcal{M}$ in the direction $\xi_\mathbf{x} \in \mathcal{T}_\mathbf{x}\mathcal{M}$ is given by*

$$\gamma_{\mathbf{x},\xi_\mathbf{x}}(t) = \sum_{i=1}^{k} \left[\cos\left(\sqrt{\frac{\xi_\mathbf{x}^* \mathbf{A}_i \xi_\mathbf{x}}{b_i}} t\right) \mathbf{A}_i \mathbf{x} + \sqrt{\frac{b_i}{\xi_\mathbf{x}^* \mathbf{A}_i \xi_\mathbf{x}}} \sin\left(\sqrt{\frac{\xi_\mathbf{x}^* \mathbf{A}_i \xi_\mathbf{x}}{b_i}} t\right) \mathbf{A}_i \xi_\mathbf{x} \right] . \tag{2.16}$$

The reader can refer to Lemma 2 in [47] for the proof.

The exponential map is a function from a subset of a tangent space $\mathcal{T}_\mathbf{x}\mathcal{M}$ to the manifold $\mathcal{M}$ that associates to each tangent direction $\xi_\mathbf{x}$ in the neighborhood of $\mathbf{0}_\mathbf{x}$ a geodesic curve $\gamma : \mathbb{R} \to \mathcal{M}$ going through $\mathbf{x} \in \mathcal{M}$ in the direction $\xi_\mathbf{x} \in \mathcal{T}_\mathbf{x}\mathcal{M}$, i.e., $\gamma(0) = \mathbf{x}$ and $\dot{\gamma}(0) = \xi_\mathbf{x}$. A manifold is said to be *geodesically complete* if the domain of its exponential map is the whole tangent space.

2.4 Proposed Algorithms for Phase Retrieval

Now that we characterized the tangent space, the only remaining ingredient for our Riemannian optimization algorithm is a retraction, a mapping from the vectors in the tangent space to the points on the manifold.

Mapping from the tangent space to the manifold

After obtaining the descent direction $-\text{grad} f(\mathbf{x}^t)$, the unconstrained optimization algorithms update the point by $\mathbf{x}^{t+1} = \mathbf{x}^t - \mu^t \text{grad} f(\mathbf{x}^t)$, μ^t is a (time-varying) step size. However, on Riemannian manifolds, such point may lie outside the manifold. The natural approach is to move along the geodesic in $\mathbf{x}^t$ in the direction $-\mu^t \text{grad} f(\mathbf{x}^t)$, i.e., to use the Exponential map for the update by setting $\mathbf{x}^{t+1} = \text{Exp}_{\mathbf{x}^t}(-\mu^t \text{grad} f(\mathbf{x}^t))$. However, evaluating this map for the fixed norms manifold is computationally intensive to calculate. Therefore, to improve the efficiency of the algorithm, instead of moving along the geodesic, one needs to move on a curve that only preserves the gradient at $\mathbf{x}^t$. This is accomplished by the concept of a retraction defined below

Definition 1 (Retraction)**.** *A retraction $R_\mathbf{x}$ is a mapping from $\mathcal{T}_\mathbf{x}\mathcal{M}$ to $\mathcal{M}$ that satisfies the following properties:*

1. *Centering Property:* $R_\mathbf{x}(\mathbf{0}_\mathbf{x}) = \mathbf{x}$
2. *Local Rigidity Property:* $\left.\dfrac{\delta R_\mathbf{x}(t\xi_\mathbf{x})}{\delta t}\right|_{t=0} = \xi_\mathbf{x}$

The choice of a computationally efficient retraction is a crucial step in designing highly efficient Riemannian optimization algorithms. [4] provides a way of constructing a retraction by exploiting

the Euclidean structure of the embedding space. Following the same approach, we design a highly efficient first-order retraction on the fixed norms manifold:

$$R_{\mathbf{x}}(\xi_{\mathbf{x}}) = \sum_{i=1}^{k} \sqrt{\frac{b_i}{b_i + \xi_{\mathbf{x}}^* \mathbf{A}_i \xi_{\mathbf{x}}}} \mathbf{A}_i \left(\mathbf{x} + \xi_{\mathbf{x}}\right) \ . \tag{2.17}$$

Riemannian steepest descent algorithm

After introducing the efficient retraction in (2.17), we are now ready to use this to present the Riemannian steepest descent optimization algorithm on the embedded fixed norms manifold where the steps are summarized in Algorithm 2.

Algorithm 2 Gradient Descent on the Fixed Norms Manifold

Require:

$\mathcal{M}$: Fixed norms manifold
$\ell(\cdot)$: The loss function
$\nabla\ell$: Gradient of the loss

- Initialize $\mathbf{x} \in \mathcal{M}$.
while $\nabla\ell^*(\mathbf{x})\nabla\ell(\mathbf{x}) \geq \epsilon$ **do**
- Compute search direction

$$\xi_{\mathbf{x}} = \nabla\ell(\mathbf{x}) - \sum_{i=1}^{k} \frac{1}{2b_i} (\nabla\ell^*(\mathbf{x})\mathbf{A}_i\mathbf{x} + \mathbf{x}^*\mathbf{A}_i\nabla\ell(\mathbf{x}))\mathbf{A}_i\mathbf{x}$$

- Find Armijo step size α using Backtracking.
- Update $\mathbf{x} = \sum_{i=1}^{k} \sqrt{\frac{b_i}{b_i + \alpha^2 \xi_{\mathbf{x}}^* \mathbf{A}_i \xi_{\mathbf{x}}}} \mathbf{A}_i \left(\mathbf{x} + \alpha\xi_{\mathbf{x}}\right)$.
end while
return x

2.5 Numerical Simulations

In this section, the convergence time and accuracy of the proposed Riemannian gradient descent and conjugate gradient algorithms on the fixed-norms manifold (Algorithm 2) are compared to state-of-the-art unconstrained, e.g., trust-region, and constrained, e.g., interior point and active set, optimization methods. For the empirical simulations, we generate the underlying signal $\mathbf{x}_{\text{opt}}$ as a random complex Gaussian vector and all algorithms are initialized with $\mathbf{x}$ such that $\mathbb{E}||\mathbf{x} - \mathbf{x}_{\text{opt}}||_2^2 = 2n\sigma^2$.

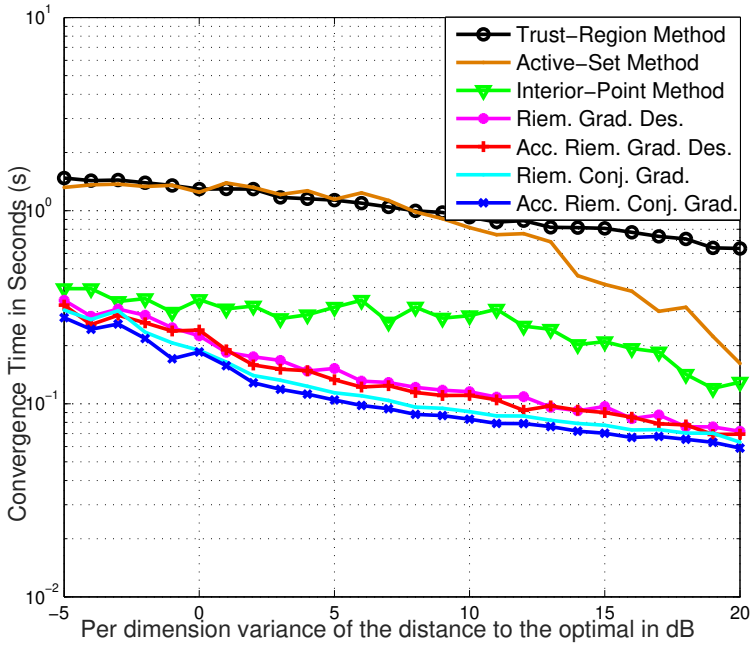


Figure 2.1: Comparison of the running time (in seconds) of different constrained and unconstrained optimization algorithms for solving the Fourier phase retrieval. The horizontal axis represents the inverse of the standard deviation (σ) in dB.

Figure 2.1 depicts a comparison of the running time of the different algorithms in solving the Fourier phase retrieval problem. It can be seen from the figure that the proposed algorithms on the fixed norms manifold systematically run faster than all other tested algorithms with an average of $50 - 100$ fold gain. Furthermore, it can be observed in Figure 2.2 that optimizing over the fixed norms manifold provides significantly higher accuracy, or equivalently a lower loss. As an example, for a σ equal to 5-dB, the achieved accuracy by the conjugate-gradient on the fixed-norm manifold is 7 order of magnitude higher than the best accuracy achieved by other tested algorithms. Therefore, properly exploiting the geometry of the problem, the proposed algorithm outperforms traditional optimization-based methods both in accuracy and convergence speed.

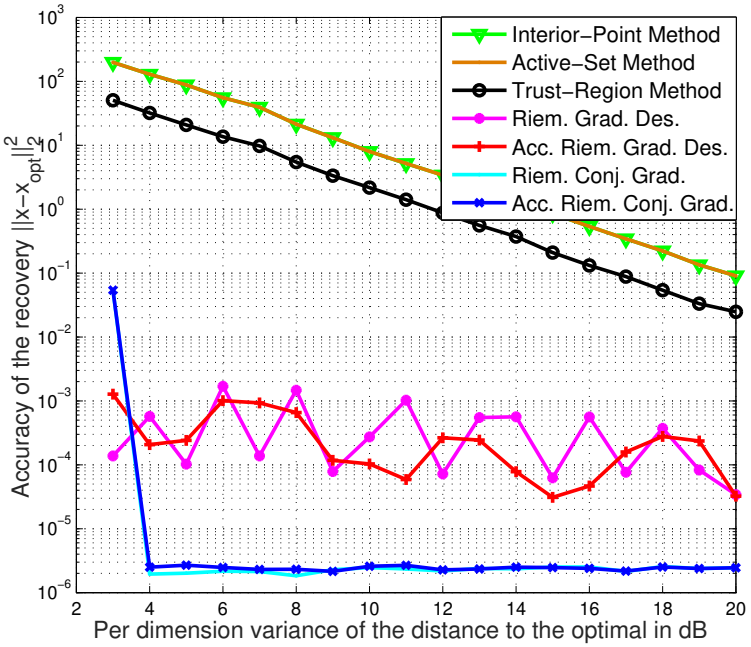


Figure 2.2: Comparison of accuracy of different constrained and unconstrained optimization algorithms in reconstructing the solution of the Fourier phase retrieval. The horizontal axis represents the inverse of the standard deviation (σ) in dB.

MULTIPLE ILLUMINATIONS PHASELESS SUPER-RESOLUTION

Phaseless super-resolution is the problem of recovering an unknown signal from measurements of the "magnitudes" of the "low frequency" Fourier transform of the signal. This problem arises in applications where measuring the phase and making high-frequency measurements are either too costly or altogether infeasible. The problem is especially challenging because it combines the difficult problems of phase retrieval and classical super-resolution. It has been shown that by appropriately "masking" the signal, and obtaining measurements of the masked signals, one can uniquely and robustly identify the phase using semidefinite programming. This is particularly useful as, upon recovering the phase, the problem will reduce to the classical super-resolution problem for which the performance has been analyzed (see e.g. [28]).

In this section, we broadly extend the class of masks that can be used to recover the phase and show how their effect can be emulated in coherent diffraction imaging using multiple illuminations, as well as in direction-of-arrival (DoA) estimation using multiple sources to excite the environment. We provide numerical simulations to demonstrate the efficacy of the method and approach. The results presented in this chapter are available in the research paper [115][1], and some of the texts appear as it is in the publication.

3.1 Background and Motivation

It is often difficult to obtain high-frequency measurements in sensing systems due to physical limitations on the highest possible resolution a system can achieve. As an example, the fundamental resolution limit in optical systems caused by diffraction is an obstacle to observe sub-wavelength structures. *Super-resolution* is the problem of recovering the high-frequency features of the signal using low-frequency Fourier measurements. In addition, as discussed in the previous chapter, many measurement systems can only measure the magnitude of the Fourier transform of the underlying signal, and the fundamental problem of recovering a signal from the magnitude of its Fourier transform is known as *phase retrieval*.

Both of the aforementioned reconstruction problems have rich histories and occur in many areas in

[1] F. Salehi, K. Jaganathan, and B. Hassibi. "Multiple Illumination Phaseless Super-resolution (MIPS) with Applications to Phaseless DoA Estimation and Diffraction Imaging." In: Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE. 2017, pp. 3949–3953.

engineering and applied physics such as astronomical imaging [106, 52], X-ray crystallography [93], medical imaging [59, 6, 5], and optics [144]. A wide variety of techniques have been proposed for super-resolution [116, 109, 28, 127] and phase retrieval [53, 66, 117] problems.

Here we consider the *phaseless super-resolution* problem, which is the problem of reconstructing a signal using its low-frequency Fourier magnitude measurements. Our work is inspired by [69] where it was shown that using three phaseless low frequency measurements, obtained by appropriately "masking" the signal, one can uniquely and robustly identify the phase using convex programming and obtain the same super-resolution performance reported in [28]. While this is a significant result, due to physical limitations in measuring systems, it is not always possible to generate the mask matrices required in [69]. The main contribution of this paper is to broadly extend the class of masks that can be used to recover the phase using convex programming. In addition, we show how these masks can be implemented in **coherent diffraction imaging**, using multiple illuminations, and **direction of arrival estimation**, using multiple sources to excite the environment.

The remaining of this chapter is organized as follows. In Section 2, we mathematically set up the reconstruction problem and present our main result. In Section 3, we describe the practical significance of our result. Section 4 contains the details of the proof. The results of the various numerical simulations are presented in Section 5.

3.2 Main Result

Let $\mathbf{x} = \begin{bmatrix} x_0, x_1, \ldots, x_{n-1} \end{bmatrix}^T$ be a complex-valued signal of length n and sparsity s. Suppose we have a device that can only measure the magnitude-squares of the $2k+1$ low frequency terms of the n-point discrete Fourier transform (DFT) of $\mathbf{x}$ (one DC term and k lowest frequencies on either side of it). Clearly, this is not sufficient to generally recover $\mathbf{x}$. The idea of masked phaseless measurements is to obtain additional information by first masking the signal and then measuring the magnitude-squares of the $2k+1$ low frequency terms of its n-point DFT. Mathematically, masking a signal is equivalent to multiplying it by a diagonal "mask" matrix, say $\mathbf{D}$ [23, 65].

In fact, more than one mask is necessary if one wishes to recover general signals from such measurements. Assuming we have r masks, for $0 \leq l \leq r-1$, we will represent them by $\mathbf{D}_l = \mathrm{diag}(d_l[0], d_l[1], \ldots, d_l[n-1])$. The problem we are interested in is recovering $\mathbf{x} \in \mathbb{C}^n$ from the resulting collection of low frequency masked phaseless measurements, viz.,

$$\begin{aligned} \text{find} \quad & \mathbf{x} \\ \text{s.t.} \quad & Z[m,l] = |\langle \mathbf{f}_m,\ \mathbf{D}_l \mathbf{x} \rangle|^2 \\ & \text{for} \quad -k \leq m \leq k \ , \quad \text{and} \quad 0 \leq l \leq r-1, \end{aligned} \tag{3.1}$$

where $\langle \cdot, \cdot \rangle$ is the standard inner product operator, $\mathbf{f}_m$ is the conjugate of the m^{th} column of the n-point DFT matrix, and $Z[m, l]$ denotes the magnitude-square of the m^{th} term of the n-point DFT for the l^{th} mask. The index m is to be understood modulo n, due to the nature of the n-point DFT.

There are two main issues that arise with the above problem: (1) designing a set of masks for which one can (up to a global phase) uniquely, efficiently, and stably identify the signal, and (2) developing an algorithm that can provably do so. Both these issues were resolved in [69] where it is shown that, under appropriate conditions, the following three masks

$$\mathbf{D}_0 = \mathbf{I}_n \quad , \quad \mathbf{D}_1 = \mathbf{I}_n + \mathbf{D}^{(1)} \quad , \quad \mathbf{D}_2 = \mathbf{I}_n - \mathrm{i}\, \mathbf{D}^{(1)}, \tag{3.2}$$

where $\mathbf{D}^{(1)}$ is a diagonal matrix whose diagonal entries are given by

$$d^{(1)}[u] = e^{\mathrm{i}2\pi \frac{u}{n}}, \quad u = 0, 1, \ldots, n-1, \tag{3.3}$$

are sufficient to uniquely identify the rank-one Hermitian matrix, $\mathbf{X} = \mathbf{x}\mathbf{x}^*$ by solving the following convex (semidefinite) program,

$$\begin{aligned} \min_{X \in \mathbb{S}^n} \quad & \|\mathbf{X}\|_1 \\ \text{s.t.} \quad & Z[m, l] = \mathrm{tr}\big(\mathbf{D}_l^\star \mathbf{f}_m \mathbf{f}_m^\star \mathbf{D}_l \mathbf{X}\big) \\ & \qquad \text{for} \quad -k \le m \le k \quad , \quad \text{and} \quad 0 \le l \le r-1, \\ & \mathbf{X} \succeq \mathbf{0}. \end{aligned} \tag{3.4}$$

The above convex program is obtained by the standard method of linearizing a quadratic-constrained problem by *lifting* the problem to the rank-one matrix $\mathbf{X} = \mathbf{x}\mathbf{x}^*$, and consequently convexifying it by relaxing the rank one constraint to a non-negativity constraint. This method has been explained earlier in Chapter 2.2. Here, since the matrix we want to recover is sparse, the ℓ_1 norm is used as the objective function.

While the result of [69] is very nice, in many applications, the masking matrix $\mathbf{D}^{(1)}$ is difficult to implement. Therefore, it is desirable to have more flexibility in the mask designs so as to permit more applications. We herein propose a set of 5 flexible masks. The building blocks of these masks are the diagonal matrices denoted by $\mathbf{D}^{(q)}$, for $0 \le q \le n-1$, whose diagonal entries are,

$$d^{(q)}[u] = e^{\mathrm{i}2\pi \frac{qu}{n}}, \quad u = 0, 1, \ldots, n-1. \tag{3.5}$$

We are now in a position to state the main result of this chapter.

Theorem 2. *The convex program* (3.4) *has a unique optimizer, namely* $\mathbf{X} = \mathbf{x}\mathbf{x}^\star$*, and thus* $\mathbf{x}$ *can be uniquely identified (up to a global phase), if*

1. $\Delta = \min_{0 \le i,j \le s-1, i \ne j} (t_i - t_j) \bmod n \ge \frac{Cn}{k}$*, where* t_i*s for* $0 \le i \le k-1$ *are the positions of the non-zero entries of* x*, and* C *is a numerical constant.*

2. $\mathbf{y}_{-k}, \ldots, \mathbf{y}_0, \ldots, \mathbf{y}_k \ne 0$*, where* $\mathbf{y}$ *denotes the n-point DFT of* $\mathbf{x}$*.*

3. *The following mask matrices are used:*

$$\begin{aligned} \mathbf{D}_0 = \mathbf{D}^{(0)} = \mathbf{I},\ \mathbf{D}_1 = \mathbf{I} + \mathbf{D}^{(q_1)},\ \mathbf{D}_2 = \mathbf{I} - \mathrm{i}\,\mathbf{D}^{(q_1)} \\ \mathbf{D}_3 = \mathbf{I} + \mathbf{D}^{(q_2)}, \quad \mathbf{D}_4 = \mathbf{I} - \mathrm{i}\,\mathbf{D}^{(q_2)}. \end{aligned} \tag{3.6}$$

4. q_1 *and* q_2 *are integers that satisfy*

$$gcd(q_1, q_2) = 1, \quad |q_1| + |q_2| \le 2k. \tag{3.7}$$

As we shall presently see, the masks used in the theorem are easy to implement in both DoA Estimation and Coherent Diffraction Imaging setups.

3.3 Applications

Phaseless Direction of Arrival Estimation

Consider the planar direction of arrival estimation setup described in Fig. 3.1. Suppose there are M objects which can reflect waves, with the m^{th} object, for $0 \le m \le M-1$, located at distance r_m and angle θ_m from the origin. A transmitter positioned at location $-\frac{l\lambda}{2}$ on the x-axis, where λ is the transmission wavelength, is used to transmit narrow-band waves with center frequency $f_c = \frac{c}{\lambda}$, and a uniform linear array (ULA) consisting of $2k+1$ receivers located along the x-axis at $(-\frac{k\lambda}{2}, \ldots, 0, \frac{\lambda}{2}, \ldots, \frac{k\lambda}{2})$ is used for signal detection. The direction of arrival estimation problem deals with estimating θ_m, for $0 \le m \le M-1$, from the received signal.

If $\mathbf{y}$ denotes the narrow-band vector impinging on the receivers in the frequency domain, then we can write:

$$y_k \quad \propto \sum_{m=0}^{M-1} (\rho_m e^{\frac{-\mathrm{i}2\omega_c r_m}{c}}) e^{\mathrm{i}\pi(k-l)sin\theta_m}, \tag{3.8}$$

where ρ_m is the reflectivity of object m and $\omega_c = 2\pi f_c$ [136]. We refer the reader to section 6.1 of [64] to follow details of this formulation. If $l = 0$, then the vector $\mathbf{y}$ represents the $2k+1$

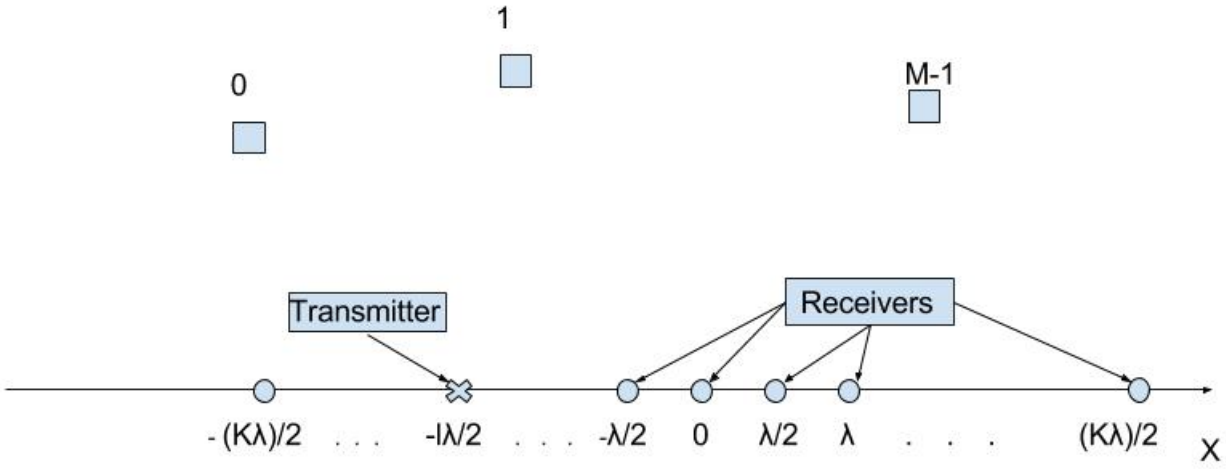


Figure 3.1: Direction of arrival estimation using a uniform linear array.

low-frequency terms of the Fourier series of a signal having amplitudes $\rho_m e^{\frac{-i2\omega_c r_m}{c}}$ at locations $\frac{sin\theta_m}{2}$. Hence, direction of arrival estimation involves solving the classic super-resolution problem.

Observe that, for an integer q, the vector $\mathbf{y}$ represents the $2k+1$ low-frequency measurements of the same signal which is masked by the matrix $\mathbf{D}^{(q)}$. Theorem 2, coupled with this critical observation, enables phaseless direction of arrival estimation.

The mask $\mathbf{D}_0$ in Theorem 2 can be implemented by putting an in-phase transmitter at the origin, $\mathbf{D}_1$ and $\mathbf{D}_3$ by using additional in-phase transmitters at $-\frac{q_1\lambda}{2}$ and $-\frac{q_2\lambda}{2}$, respectively, and $\mathbf{D}_2$ and $\mathbf{D}_4$ by using additional transmitters that have $\pi/2$ phase difference at those very locations. As a result, if 5 strategically placed transmitters are used for transmission, then there is no need to measure phase during reception and the angles can be provably recovered by solving (3.4). This is particularly useful in scenarios where measuring phase reliably is either impractical or too costly.

Remark 1. *This idea also extends to the nested array and co-prime array setups described in [102] and [137], respectively.*

Coherent Diffraction Imaging (CDI)

Consider the planar CDI setup described in Figure 3.2. Let the object and the detector be perpendicular to the x-axis, located at $x = 0$ and $x = d$, respectively, and $\psi(z)$ denote the one-dimensional object which we wish to determine. The object is illuminated using a coherent source incident at an angle θ with respect to the x-axis.

Detection devices cannot measure the phase of the incoming light waves (the frequency is too high),

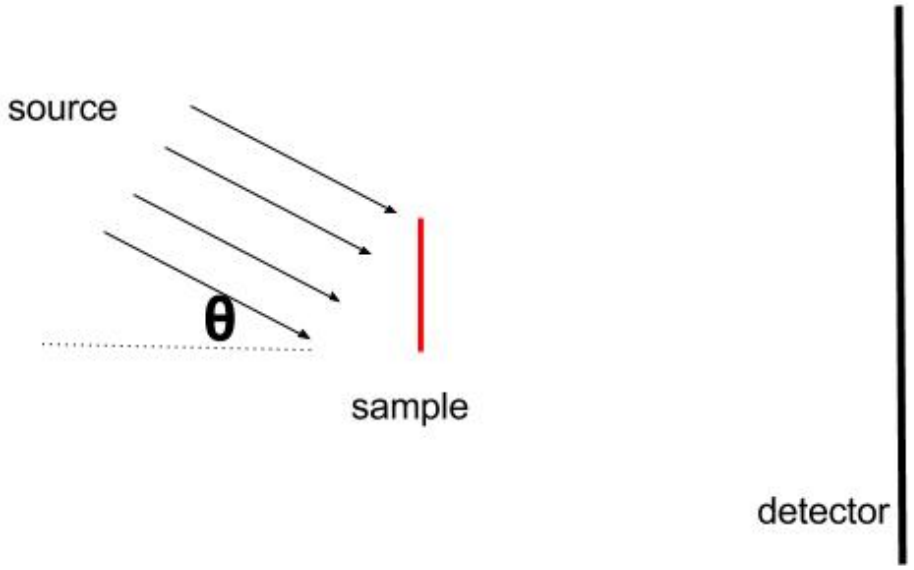


Figure 3.2: A typical Coherent Diffraction Imaging setup.

and instead measure the photon flux. The flux measurements at position z' on the detector, denoted by $I(z')$, are well approximated by:

$$I(z') \propto \left| \int_z \psi(z) e^{\mathrm{i}\frac{2\pi z}{\lambda}(-\frac{z'}{d}+\theta)} dz \right|^2 . \tag{3.9}$$

If $\theta = 0$, then the measurements provide the knowledge of the Fourier magnitude-square of $\psi(z)$. Section 6.2 in [64] presents details of the above formulation. Therefore, diffraction imaging involves solving the phase retrieval problem. Quite often, the approximation (3.9) only applies to positions closer to $z = 0$. As a result, one needs to solve phaseless super-resolution in order to recover the underlying object.

If $\theta = \frac{q}{d}$, then the measurements correspond to the Fourier magnitude-square of $\psi(z)$ masked by the matrix $\mathbf{D}^{(q)}$. The equations are identical to those in the direction of arrival setup. Hence, by using 5 strategic illuminations (using sources placed at $\theta = 0, \frac{q_1}{d}, \frac{q_2}{d}$), one can provably recover the object from the low-frequency Fourier magnitude measurements by solving (3.4).

3.4 Proof of Theorem 2

Let $\mathbf{F}$ denote the n-point DFT matrix and $\mathbf{F}_k$ be the $(2k+1) \times N$ sub-matrix of $\mathbf{F}$, consisting of the rows $-k \leq m \leq k$ (understood modulo N). Define $\mathbf{y}_k = \mathbf{F}_k \mathbf{x}$ which simply denote the $2k+1$ low frequency terms in the n-point DFT of $\mathbf{x}$. The proof involves two key steps:

1. The matrix $\mathbf{y}_k\mathbf{y}_k^*$ is uniquely determined by the set of constraints described in (3.4).

2. Given $\mathbf{y}_k\mathbf{y}_k^*$, the matrix $\mathbf{x}\mathbf{x}^*$ can be uniquely reconstructed by minimizing $\|\mathbf{X}\|_1$ under the conditions specified in Theorem 2.

Proceeding onward, we now provide the details for the first step. Consider the following affine transformation $\mathbf{Y} = \mathbf{F}_k\mathbf{X}\mathbf{F}_k^*$. When measurements are obtained using the masks proposed in Condition 3 of Theorem 2, the affine constraints of (3.4) can be rewritten in terms of the variable $\mathbf{Y}$ as follows,

$$\begin{aligned}
\mathbf{Y}[m,m] &= |\mathbf{y}_k[m]|^2, \qquad \text{for } -k \le m \le k\ , \\
\mathbf{Y}[m,m+q_1] &= \mathbf{y}_k[m]\mathbf{y}_k^*[m+q_1], \qquad \text{for } -k \le m \le k-q_1\ , \\
\mathbf{Y}[m,m+q_2] &= \mathbf{y}_k[m]\mathbf{y}_k^\star[m+q_2], \qquad \text{for } -k \le m \le k-q_2.
\end{aligned} \tag{3.10}$$

where $\mathbf{Y}[r,c]$ denotes the entry in row r and column c of the matrix $\mathbf{Y}$. For the sake of brevity, we omit some of the details here. We refer the interested readers to the proof of Theorem 3.1 in [69].

As a result of (3.10), the set of constraints in (3.4) can be viewed as a matrix completion problem in $\mathbf{Y}$. Next, we define a graph $G = (\mathcal{V}, \mathcal{E})$ on the vertices $\mathcal{V} = \{-k, -k+1, \ldots, k-1, k\}$ with the edge set $\mathcal{E}$ defined such that, for $-k \le m \le k$, $(m, m-q_1) \in \mathcal{E}$ and $(m, m-q_2) \in \mathcal{E}$. In other words, the graph G contains an edge between vertices i and j if the $(i,j)^{\text{th}}$ entry of Y is fixed by the measurements. Since l_1 and l_2 are co-prime (Condition 4 in Theorem 2), the graph G is connected. Additionally, every vertex has an edge with itself (i.e., all the diagonal entries are fixed by the measurements).

The following lemma establishes that under the conditions specified in Theorem 2, the matrix $\mathbf{Y}_k = \mathbf{y}_k\mathbf{y}_k^*$ is *the unique feasible solution*.

Lemma 3. *Suppose $G = (\mathcal{V}, \mathcal{E})$ is an undirected graph on $\mathcal{V} = \{v_0, v_1, \ldots, v_{n-1}\}$. For $e = (v_i, v_j) \in \mathcal{E}$, define $\mathbf{A}_e \in \mathbb{C}^{n\times n}$ as the matrix with all entries zero except for $\mathbf{A}[i,j]$, which is equal to 1. Also, for $i = 0, 1, \ldots, n-1$, define the matrix $\mathbf{A}_i \in \mathbb{C}^{n\times n}$ as the matrix that is zero everywhere except for $\mathbf{A}[i,i]$, which is equal to 1. Suppose $\mathbf{z} \in \mathbb{C}^n$ is a vector with all entries being non-zero.*

The matrix $\mathbf{Z} = \mathbf{z}\mathbf{z}^*$ *is the unique solution of*

$$\begin{aligned} \underset{\mathbf{X}\in\mathbb{S}^n}{\textit{find}} \quad & \mathbf{X} \\ \textit{s.t.} \quad & tr\big(\mathbf{A}_i\mathbf{X}\big) = |\mathbf{z}[i]|^2, \textit{ for } \ i = 0, 1, \ldots, n-1 \\ & tr(\mathbf{A}_e\mathbf{X}) = \mathbf{z}^*[j]\mathbf{z}[i], \textit{ for } \ e = (v_i, v_j) \in \mathcal{E} \\ & \mathbf{X} \succeq \mathbf{0} \end{aligned} \tag{3.11}$$

if and only if G *is connected.*

The proof of Lemma 3 is based on the method of dual certificates. We postpone the detailed proof to Section 3.6.

Now that we established the uniqueness of $\mathbf{Y}_k$ from the result of Lemma 3, we can explain the second key ingredient of the proof. Note that after finding $\mathbf{Y}_k$, the problem would reduce to the classical super-resolution problem.

To apply the theoretical result from the theory of super-resolution, we exploit the fact that $\mathbf{Y}$ corresponds to the $2k + 1$ two-dimensional low frequencies of the two-dimensional signal $\mathbf{X}$. Consequently, this step would be a direct consequence of the two-dimensional super-resolution theorem in [28]. When Condition 1 in Theorem 2, also known as the minimum separation condition, holds the optimization program (3.4) uniquely identifies the signal.

3.5 Numerical Results

In this section, the performance of the solution to the optimization problem (3.4) is demonstrated through numerical simulations. We provide simulation results for both noiseless and noisy settings.

Noiseless setting

We choose $n = 40$, $q_1 = 2$, and $q_2 = 3$. The masks $\{\mathbf{D}_0, \mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3, \mathbf{D}_4\}$ defined in (3.6) are used to obtain phaseless low frequency measurements. Using parser YALMIP and solver SeDuMi, we simulate 20 trials for various choices of k and Δ. We generate the indices of the support of the signal in such a way that the minimum separation condition (i.e. condition 1 in Theorem 2) is satisfied. Signal values in the support are drawn independently from a standard normal distribution. The probability of successful reconstruction of the signal by the semidefinite program (3.4) as a function of k and Δ is depicted in Figure 3.3. The white region corresponds to a success probability of 1 and the black region corresponds to a success probability of 0. The plot shows that (3.4) successfully reconstructs signals with high probability when $k \geq \frac{n}{\Delta}$.

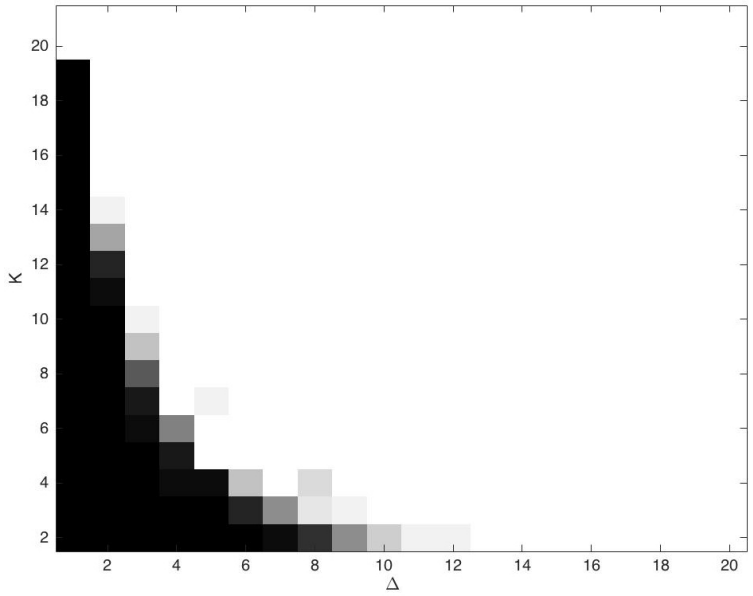


Figure 3.3: Probability of successful reconstruction of the signal by solving the semidefinite program 3.4. For the numerical simulations, we set $n = 20$, $q_1 = 2$, and $q_2 = 3$. The empirical result is based on 20 trials for various choices of k and Δ, using the masks defined in (3.6).

Noisy setting

A major advantage of semidefinite programming-based reconstruction is robustness to noise. In this part, we demonstrate the performance of the solution of the optimization program (3.4) in the noisy setting.

To test the robustness, here we add an i.i.d. standard normal noise (with appropriate variance) to each measurement, $Z[m, r]$. We first solve the program (3.4) by replacing the equality constraints with appropriate inequality constraints, and obtain the optimizer $\hat{\mathbf{X}}$. Then, we find its best rank-one approximation, say $\hat{\mathbf{x}}\hat{\mathbf{x}}^\star$. The estimate $\hat{\mathbf{x}}$ is then compared with the true solution $\mathbf{x}$ in terms of the mean-squared error.

We set $n = 40$, $q_1 = 2$, $q_2 = 3$, $k = 14$, and $\Delta = 8$. By varying the signal-to-noise ratio (SNR), we simulate 20 trials and compute the mean-squared error $\mathbb{E}[\frac{\|\hat{\mathbf{x}}-\mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2}]$. The results are depicted in Figure 3.4.

In the logarithmic scale, we see a linear relationship between the mean-squared error and SNR. This clearly indicates that the reconstruction is stable in the noisy setting.

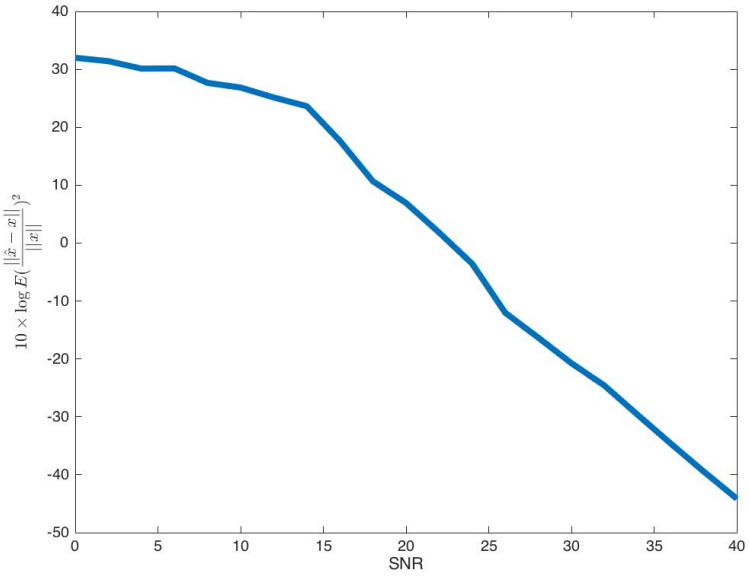


Figure 3.4: Mean-squared error (MSE) as a function of SNR for $n = 40$, $q_1 = 2$, $q_2 = 3$, $k = 14$, and $\Delta = 8$.

3.6 Proof of Lemma 3

The proof is based on the method of dual certificates. We define matrix $\mathbf{W} \in \mathbb{C}^{n \times n}$ as follows:

$$\mathbf{W} = \sum_{i,j:(v_i,v_j)\in\mathcal{E}} \mathbf{W}_{ij} \tag{3.12}$$

where $\mathbf{W}_{ij}$, for $0 \leq i, j \leq n-1$, is a matrix defined as follows,

$$\mathbf{W}_{ij} = \mathbf{w}_{ij}\mathbf{w}_{ij}^*, \quad \mathbf{w}_{ij} = \mathbf{z}[j]^*\mathbf{e}_i - \mathbf{z}^*[i]\mathbf{e}_j, \tag{3.13}$$

where $\mathbf{e}_i \in \mathbb{C}^n$ denotes the i^{th} vector in the standard basis. We will show that $\mathbf{W}$ satisfies the following properties:

1. $\mathbf{W} \succeq \mathbf{0}$,
2. $\mathbf{WZ} = \mathbf{0}$,
3. $\text{rank}(\mathbf{W}) = n - 1$.

$\mathbf{W}$ is a positive semidefinite matrix as it is the sum of $\mathbf{W}_{ij}$ where $\mathbf{W}_{ij} = \mathbf{w}_{ij}\mathbf{w}_{ij}^* \succeq \mathbf{0}$. In order to show that properties 2 and 3 are satisfied, we show the following:

$$\mathbf{y}^*\mathbf{W}\mathbf{y} = 0 \Leftrightarrow \mathbf{y} = \alpha\mathbf{z} \;\text{ for some }\; \alpha \in \mathbb{C}, \tag{3.14}$$

which simply means that the null-space of $\mathbf{W}$ is $\text{span}(\mathbf{z})$. One can write:

$$\mathbf{y}^*\mathbf{W}\mathbf{y} = \sum_{i,j:(v_i,v_j)\in\mathcal{E}} \mathbf{y}^*\mathbf{W}_{ij}\mathbf{y} = \sum_{i,j:(v_i,v_j)\in\mathcal{E}} |\mathbf{y}[i]\mathbf{z}[j] - \mathbf{y}[j]\mathbf{z}[i]|^2 \tag{3.15}$$

which gives the following,

$$\mathbf{y}^*\mathbf{W}\mathbf{y} = 0 \Leftrightarrow \mathbf{y}[i]\mathbf{z}[j] - \mathbf{y}[j]\mathbf{z}[i] = 0, \quad \forall (i,j) \in \mathcal{E}\ . \tag{3.16}$$

If G is connected and the entries of $\mathbf{z}$ are all non-zero, (3.16) is valid *if and only if* $\mathbf{y} = \alpha\mathbf{z}$ for some $\alpha \in \mathbb{C}$. This shows that $\text{rank}(\mathbf{W}) = n-1$. In addition, we have

$$\text{tr}(\mathbf{W}\mathbf{Z}) = \text{tr}(\mathbf{W}\mathbf{z}\mathbf{z}^*) = \mathbf{z}^*\mathbf{W}\mathbf{z} = 0. \tag{3.17}$$

Proceeding onwards, we use the above properties to prove Lemma 3. To this end, we need to show that the matrix $\mathbf{Z}$ is the unique feasible point of the following optimization problem,

$$\begin{aligned} \underset{\mathbf{X}\in\mathbb{S}^n}{\text{find}} \quad & \mathbf{X} \\ \text{s.t.} \quad & \text{tr}(\mathbf{A}_i\mathbf{X}) = |\mathbf{z}[i]|^2, \;\text{ for }\; i = 0,1,\ldots,n-1 \\ & \text{tr}(\mathbf{A}_e\mathbf{X}) = \mathbf{z}^*[j]\mathbf{z}[i], \;\text{ for }\; e = (v_i,v_j) \in \mathcal{E} \\ & \mathbf{X} \succeq \mathbf{0}\ . \end{aligned} \tag{3.18}$$

The dual of this optimization problem is

$$\begin{aligned} \max_{\lambda\in\mathbb{C}^n,\mu\in\mathbb{C}^{|\mathcal{E}|}} \quad & -\sum_{i=0}^{n-1} \lambda_i|\mathbf{z}[i]|^2 - \sum_{i,j:(v_i,v_j)\in\mathcal{E}} (\mu_{i,j}\mathbf{z}^*[j]\mathbf{z}[i] + \bar{\mu}_{i,j}\mathbf{z}^*[i]\mathbf{z}[j]) \\ \text{s.t.} \quad & \sum_{i=0}^{n-1} \lambda_i\mathbf{A}_i + \sum_{i,j:e=(v_i,v_j)\in\mathcal{E}} (\mu_{i,j}\mathbf{A}_e + \bar{\mu}_{i,j}\mathbf{A}_e^*) \;\succeq \mathbf{0}\ . \end{aligned} \tag{3.19}$$

For $0 \leq i \leq n-1$, define $\mathcal{N}(i)$ as the set of neighbors of node v_i in G. By choosing $\lambda_i^\star = \sum_{j:j\in N(i)} |\mathbf{z}[j]|^2$ and $\mu_{ij}^\star = \mathbf{z}^*[j]\mathbf{z}[i]$, we can define the following matrix,

$$\mathbf{W} = \sum_{i=0}^{n-1} \lambda_i^\star \mathbf{A}_i + \sum_{i,j:e=(v_i,v_j)\in\mathcal{E}} (\mu_{i,j}^\star \mathbf{A}_e + \bar{\mu}_{i,j}^\star \mathbf{A}_e{}^*) \ . \tag{3.20}$$

Property 1 of the matrix $\mathbf{W}$ ensures that $\mathbf{W} \succeq \mathbf{0}$ which is the dual feasibility. Property 2 is the complimentary slackness. These two properties prove that $\mathbf{Z} = \mathbf{z}\mathbf{z}^*$ is an optimal solution for (3.11).

Now suppose there is another solution, namely $\mathbf{Z} + \mathbf{H}$, where $\mathbf{H} \in \mathbb{S}^n$ is an $n \times n$ Hermitian matrix. Let $\mathcal{T}_\mathbf{z}$ denote the set of Hermitian matrices of the form,

$$\mathcal{T}_\mathbf{z} = \{\mathbf{z}\mathbf{h}^* + \mathbf{h}\mathbf{z}^* : \mathbf{h} \in \mathbb{C}^n\}, \tag{3.21}$$

and $\mathcal{T}_\mathbf{z}^\perp$ be its orthogonal complement. In other words, $\mathcal{T}_\mathbf{z}$ is the tangent space at $\mathbf{z}\mathbf{z}^*$ to the manifold of Hermitian matrices of rank one. $\mathbf{H}$ can be decomposed as two parts $\mathbf{H}_{\mathcal{T}_\mathbf{z}}$ and $\mathbf{H}_{\mathcal{T}_\mathbf{z}^\perp}$, that are the projections of $\mathbf{H}$ onto the subspaces $\mathcal{T}_\mathbf{z}$ and $\mathcal{T}_\mathbf{z}^\perp$, respectively. In order to be an optimal solution, $\mathbf{H}$ should satisfy

$$\mathrm{tr}\big(\mathbf{W}\mathbf{H}\big) = \mathrm{tr}\big(\mathbf{W}\mathbf{H}_{\mathcal{T}_\mathbf{z}}\big) + \mathrm{tr}\big(\mathbf{W}\mathbf{H}_{\mathcal{T}_\mathbf{z}^\perp}\big) = 0. \tag{3.22}$$

Property 2 ensures that $\mathrm{tr}\big(\mathbf{W}\mathbf{H}_{\mathcal{T}_\mathbf{z}}\big) = 0$; therefore, we have $\mathrm{tr}(\mathbf{W}\mathbf{H}_{\mathcal{T}_\mathbf{z}^\perp}) = 0$. Since $\mathbf{H}$ is a positive semidefinite matrix, its projection onto $\mathcal{T}_\mathbf{z}^\perp$ is also positive semidefinite. $\mathbf{H}_{\mathcal{T}_\mathbf{z}^\perp} \succeq \mathbf{0}$ together with properties 2 and 3 lead to,

$$\mathrm{tr}(\mathbf{W}\mathbf{H}_{\mathcal{T}_\mathbf{z}^\perp}) = 0 \Leftrightarrow \mathbf{H}_{\mathcal{T}_\mathbf{z}^\perp} = \mathbf{0}, \tag{3.23}$$

where for the last equality, we used the fact that the matrix $\mathbf{W}$ has rank $n-1$ and its null space lies on the line spanned by $\mathbf{z}$.

To conclude the proof, it remains to show that $\mathbf{H}_{\mathcal{T}_\mathbf{z}} = \mathbf{0}$. In order to be a feasible point, $\mathbf{H}_{\mathcal{T}_\mathbf{z}}$ must satisfy the following conditions:

$$\begin{aligned} &\mathrm{tr}\big(\mathbf{A}_i\mathbf{H}_{\mathcal{T}_\mathbf{z}}\big) = 0, \ \text{for} \ \ i = 0, 1, \ldots, n-1, \ \ \text{and,} \\ &\mathrm{tr}\big(\mathbf{A}_e\mathbf{H}_{\mathcal{T}_\mathbf{z}}\big) = 0, \ \text{for} \ \ e = (v_i, v_j) \in \mathcal{E}. \end{aligned} \tag{3.24}$$

It is easy to check that the only matrix in $\mathcal{T}_{\mathbf{z}}$ which satisfies the above conditions is the zero matrix. Therefore, $\mathbf{H} = \mathbf{0}$ and $\mathbf{Z} = \mathbf{z}\mathbf{z}^*$ is the unique solution of (3.11).

RECOVERY THRESHOLD OF PHASELIFT IN REAL PHASE RETRIEVAL

In this chapter, we study the recovery threshold of the semidefinite program (the method known as the *PhaseLift*) for the phase retrieval problem. Our result provides a precise analysis on the required number of measurements in order to perfectly recover the underlying signal, and is valid for a broad class of sub-Gaussian distributions. To analyze this problem, we first formulate the phase retrieval as the problem of finding a rank-1 matrix from its quadratic measurements. Consequently, we consider the problem of low-rank matrix recovery from its quadratic measurements, where the goal is to recover a low-rank positive semidefinite matrix. The presented recovery threshold is valid in the asymptotic regime when the dimension of the underlying signal, n, and the number of measurements, m, go to infinity at a proportional rate, $\delta := \frac{m}{n}$. We show that the minimum rate of random quadratic measurements (also known as rank-one projections) required to recover a low rank positive semidefinite matrix is $3r$, where r denotes the rank of the PSD matrix. As a consequence, we settle the long standing open question of determining the minimum number of measurements required for perfect signal recovery in phase retrieval using the celebrated PhaseLift algorithm, and show it to be $3n$. The results presented in this section are available in the research paper [2] [1], and some of the texts appear as it is in the publication. This research paper provides a novel universality result for the setting where the entries of measurement vectors are sub-Gaussian that can be (slightly) correlated. This is an upgrade compared to the previous results in the literature [100, 103].

4.1 Matrix Recovery from Quadratic Measurements

In this section we consider the problem of recovering a matrix from (so-called) *quadratic measurements*. Our approach here is similar to that described earlier in Section 2.2. The goal is to reconstruct a PSD matrix $\in\mathbb{R}^{n\times n}$ in a convex set $\mathcal{S}$, given m measurements of the form,

$$b_i^2 = \mathbf{a}_i^T \mathbf{X}_0 \mathbf{a}_i = \mathrm{tr}\left(\mathbf{X}_0(\mathbf{a}_i\mathbf{a}_i^T)\right), \quad i = 1, \ldots, m \ . \tag{4.1}$$

Here, $\{\mathbf{a}_i\}_{i=1}^n$ denotes the set of measurement vectors. Depending on the application, the matrix $\mathbf{X}$ may exhibit various structures. To enforce this structure, we use a convex penalty function

[1] E. Abbasi, F. Salehi, and B. Hassibi. "Universality in Learning from Linear Measurements." In: arXiv preprint arXiv:1906.08396(2019).

$f : \mathbb{S}^n \to \mathbb{R}$, to enforce this structure via the following convex estimator,

$$\begin{aligned} \hat{\mathbf{X}} = \arg\min_{\mathbf{X} \in \mathcal{S}} \; & f(\mathbf{X}) \\ \text{subject to:} \quad & \mathbf{a}_i^T \mathbf{X} \mathbf{a}_i = b_i^2, \quad i = 1, \ldots, m \; . \end{aligned} \tag{4.2}$$

Note that the measurements in (4.1) are linear with respect to the matrix $\mathbf{X}$, yet quadratic with respect to the measurement vectors $\mathbf{a}_i$.

For our result to hold, we require the measurement vectors to satisfy the following assumption:

Assumption 1. *We say vectors $\{\mathbf{a}_i\}_{i=1}^m$ satisfy Assumption 1, if*

1. *$\mathbf{a}_i$'s are drawn independently from a sub-Gaussian distribution.*
2. *For each i, the entries of $\mathbf{a}_i$ are independent, zero-mean and unit-variance.*

In particular, this assumption is valid when $\{\mathbf{a}_i\}$'s have i.i.d. standard normal entries. We also impose the following assumptions on the objective function $f(\cdot)$.

Assumption 2. *We say the function $f(\cdot)$ satisfies Assumption 2, if the followings hold true.*

1. *[Separability] $f(\cdot)$ is continuous, convex, and separable, where $f(\mathbf{x}) = \sum_{i=1}^n f_i(x_i)$.*
2. *[Smoothness] The functions $\{f_i(\cdot)\}$ are three times differentiable everywhere, except for a finite number of points.*
3. *[Bounded Third Derivative] For any $C > 0$, there exists a constant $c_f > 0$, such that for all i, we have $|\frac{\partial^3 f_i(x)}{\partial x^3}| \leq c_f$, for all smooth points in the domain of $f_i(\cdot)$ such that $|x| < C$.*

As observed in the Assumption 2, we only consider the special (yet popular) case of separable penalty functions. Common choices include $\|\mathbf{X}\|_{\ell_1}$, $\|\mathbf{X}\|_F$, and $\mathrm{tr}(\mathbf{X})$ (which is equivalent to the nuclear norm for PSD matrices) for matrices.

Our main result establishes that, when the measurement vectors satisfy Assumption 1, the recovery threshold of the optimization program is equal to the recovery threshold of the following optimization program:

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{X}\in\mathcal{S}} \; f(\mathbf{X})$$
$$\text{subject to:} \quad \operatorname{tr}\left((\mathbf{H}_i + \mathbf{I})\mathbf{X}\right) = b_i^2, \quad i = 1, \ldots, m \,, \tag{4.3}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix, and $\mathbf{H}$ is a random *Wigner* matrix, that is a *symmetric* matrix whose upper-diagonal entries are drawn independently from $\mathcal{N}(0, 1)$ and its diagonals entries are drawn independently from $\mathcal{N}(0, 2)$. The following proposition presents a formal statement of our argument.

Proposition 1. *Consider the problem of recovering the matrix $\mathbf{X}_0 \in \mathcal{S} \subseteq \mathbb{S}^n$, from m quadratic measurements of the form* (4.1)*, using the estimator* (4.2)*. Assume,*

- *The measurement vectors $\{\mathbf{a}_i\}_{i=1}^m$ satisfy Assumption 1, and,*
- *$\mathcal{S}$ is a convex set, and $f(\cdot)$ is a convex function that satisfies Assumption 2,*
- *$\{\mathbf{H}_i \in \mathbf{S}^n\}_{i=1}^m$ is a set of independent Wigner matrices.*

Then, as m and n grow to infinity at a fixed rate $m = \theta(n)$, the estimator (4.2) *perfectly recovers $\mathbf{X}_0$ with probability approaching one if and only if the estimator* (4.3) *perfectly recovers $\mathbf{X}_0$ with probability approaching one.*

Low-rank Matrix Recovery

Assume the unknown matrix $\mathbf{X}_0 \succeq \mathbf{0}$ has rank r, where r is a constant r (i.e., r does not grow with problem dimensions n, m). Such matrices appear in many applications such as traffic data monitoring, array signal processing, and phase retrieval. The nuclear norm, $\|\cdot\|_\star$, is often used as the convex surrogate for low-rank matrix recovery [107].

Here, we are interested in analyzing the following optimization,

$$\hat{\mathbf{X}} = \arg\min \; \operatorname{tr}(\mathbf{X})$$
$$\text{subject to:} \quad \mathbf{a}_i^T \mathbf{X} \mathbf{a}_i = b_i^2, \quad i = 1, \ldots, m \,.$$
$$\mathbf{X} \succeq \mathbf{0} \,. \tag{4.4}$$

Note that $\text{tr}(\cdot) = \|\cdot\|_\star$ in the cone of PSD matrices. According to Proposition 1, the perfect recovery in (4.4) is equivalent to perfect recovery in the following optimization with Gaussian measurements,

$$\begin{aligned} \hat{\mathbf{X}} = \arg\min \ & \text{tr}(\mathbf{X}) \\ \text{subject to}, \quad & \text{tr}\left((\mathbf{H}_i + \mathbf{I})\mathbf{X}\right) = b_i^2, \quad i = 1, \dots, m \ . \\ & \mathbf{X} \succeq \mathbf{0} \ , \end{aligned} \tag{4.5}$$

where $\mathbf{H}_i$'s are i.i.d. Wigner matrices as defined in Proposition 1. The following corollary provides the required number of measurements for the recovery of the true matrix $\mathbf{X}_0$.

Corollary 1. *Consider the optimization program* (4.4)*, where the matrix* $\mathbf{X}_0 \succeq 0$ *has rank* r *and the measurement vectors* $\{\mathbf{a}_i\}_{i=1}^m$ *satisfy Assumption 1. Assume* $m, n \to \infty$ *at the proportional rate* $\delta := \frac{m}{n} \in (0, +\infty)$*. The estimator perfectly recovers* $\mathbf{X}_0$ *iff* $\delta > 3r$*.*

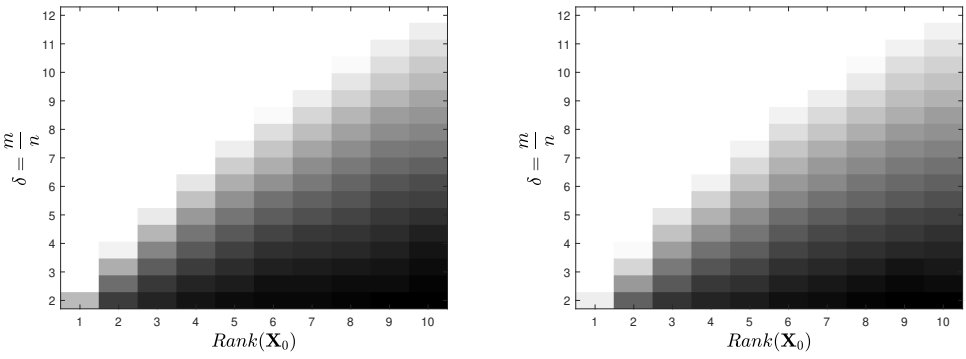
Corollary 1 indicates that $3rn$ measurements are needed to perfectly recover a rank-r PSD matrix $\mathbf{X}_0$, from quadratic measurements. To the extent of our knowledge, this is the first result that precisely computes the phase transition of low-rank matrix recovery from quadratic measurement.

Figure 4.1 depicts the result of numerical simulations. For different values of r and δ, the Frobenius norm of the error of the estimators (4.4) and (4.5) has been computed. As observed in this Figure, the empirical phase transition matches the result of Corollary 1, that is $\delta > 3r$.

Phase transition of PhaseLift in phase retrieval

We are now ready to settle the main question of the chapter, that is the required number of Gaussian measurements for perfect recovery of the signal in Phase retrieval. Recall from Chapter 2 that the PhaseLift optimization is defined as the optimization (2.6) which has exactly the same form as (4.4), with $\mathbf{X}_0 = \mathbf{x}_0\mathbf{x}_0^T$ being a rank-1 matrix.

Therefore, the recovery threshold for PhaseLift can be viewed as an important application for the result of Corollary 1, when the underlying matrix $\mathbf{X}_0$ is of rank 1. Corollary 1 states that the phase transition of the PhaseLift algorithm happens at $\delta^\star = 3$, i.e., $m > 3n$ measurements are needed for the perfect signal reconstruction in PhaseLift. We should emphasize the significance of this result as establishing the exact phase transition of the PhaseLift algorithm was an open problem.



(a) (b)

Figure 4.1: Phase transition regimes for both estimators (4.4) and (4.5), in terms of the oversampling ratio $\delta = \frac{m}{n}$ and $r = \text{rank}(\mathbf{X}_0)$, for the cases of (a) estimator (4.4) with quadratic measurements and (b) estimator (4.5) with Gaussian measurements. In the numerical simulations, we used matrices of size $n = 40$. The data is averaged over 20 independent realizations of the measurements.

4.2 A Universality Result

The result of Proposition 1 and Corollary 1 are justified via a more general universality result by Abbasi et al [2]. In this section, we state this universality results without proof. The interested reader is referred to the Appendix of [2] for a more detailed discussion as well as the technical proofs.

Motivation and background

Recovering a structured signal from a set of linear observations appears in many applications in areas ranging from finance to biology, and from imaging to signal processing. More formally, the goal is to recover an unknown vector $\mathbf{x}_0 \in \mathbb{R}^n$, from observations of the form $y_i = \mathbf{a}_i^T \mathbf{x}_0$, for $i = 1, \ldots, m$. In many modern applications, the ambient dimension of the signal, n, is often larger than the number of observations, m, which results in infinitely many solutions that satisfy the linear equations arising from the observations, and therefore to obtain a unique solution where one must assume some prior structure on the unknown vector. Therefore, the following estimator is used to recover $\mathbf{x}_0$,

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \ f(\mathbf{x}) \quad \text{subject to,} \quad y_i = \mathbf{a}_i^T \mathbf{x}, \ \ i = 1, \ldots, m \ , \tag{4.6}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex penalty function that captures the *structure* of the structured signal.

A canonical question in this area is "how many measurements are needed to recover $\mathbf{x}_0$ via this estimator?" This question has been extensively studied in the literature (see [122, 7, 33] and the references therein). The answer depends on the $\mathbf{a}_i$ and is very difficult to determine for any given set of measurement vectors. As a result, it is common to assume that the measurement vectors are drawn randomly from a given distribution and to ask whether the unknown vector can be recovered

with high probability. In the special case where the entries of the measurement matrix are drawn i.i.d. from a Gaussian distribution, the minimum number of measurements for the recovery of $\mathbf{x}_0$ with high probability is known and is related to the concept of the Gaussian width (see Chapter 6 for a more detailed discussion on this subject). For instance, it has been shown that $2k\log(n/k)$ linear measurements are required to recover a $k-$sparse signal, and $3rn$ measurements suffice for the recovery of a symmetric $n \times n$ rank-r matrix. Recently, Oymak et al. [100] showed that these thresholds remain unchanged, as long as the entries of each $\mathbf{a}_i$ are *i.i.d* and drawn from a "well-behaved" distribution. It has also been shown that similar universality holds in the case of noisy measurements [103]. Although these works are of great interest, the independence assumption on the entries of the measurement vectors can be restrictive. Here, we discuss a stronger universality result which holds for a broader class of measurement distributions. One important ramification of this result is to establish the precise recovery threshold for the low-rank matrix recovery from quadratic measurements. Such measurement schemes appear in a variety of problems [35, 21, 147, 87].

Universality theorem

Here we state the main Theorem that is known as the universality result. Before stating the result, we provide some definitions on the perfect recovery in convex estimators.

Definition 2. *Let $\mathbf{x}_0 \in \mathcal{S}$ where $\mathcal{S} \subseteq \mathbb{R}^n$ is a convex set. For a convex function $f : \mathbb{R}^n \to \mathbb{R}$ and a measurement matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, we define the convex estimator $\mathcal{E}\{\mathbf{x}_0, \mathbf{A}, \mathcal{S}, f(\cdot)\}$ as follows,*

$$\hat{\mathbf{x}} = \arg \min_{\substack{\mathbf{x}\in\mathcal{S} \\ \mathbf{A}\mathbf{x}=\mathbf{A}\mathbf{x}_0}} f(\mathbf{x}) \ . \tag{4.7}$$

We say $\mathcal{E}\{\mathbf{x}_0, \mathbf{A}, \mathcal{S}, f(\cdot)\}$ has perfect recovery iff $\hat{\mathbf{x}} = \mathbf{x}_0$.

In the main result, presented in Theorem 3, we show universality for a wide range of distributions on the measurement vector as well as a broad class of convex penalties. Here, we first explain the conditions needed for the measurement matrix,

Assumption 3. ***[The Measurement Vectors]*** *We say that the measurement matrix $\mathbf{A} = \left[\mathbf{a}_1, \ldots, \mathbf{a}_m\right]^T \in \mathbb{R}^{m\times n}$ satisfies Assumption 3 with parameters $\boldsymbol{\mu} \in \mathbf{R}^n$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$, if the following holds true.*

1. *[Sub-Exponential Tails] The vectors $\mathbf{a}_i$'s are independently drawn from a random sub-exponential distribution, with mean $\boldsymbol{\mu}$ and covariance $\bar{\kappa}\mathbf{I} \succ \boldsymbol{\Sigma} \succ \underline{\kappa}\mathbf{I}$, for some positive constants $\bar{\kappa}, \underline{\kappa} > 0$.*

2. *[Bounded Mean] For some constants $c_1, \tau_1 > 0$, we have $\frac{\|\mu\|_2^2}{\mathbb{E}[\|\mathbf{a}_i - \boldsymbol{\mu}\|^2]} \leq c_1 \cdot n^{-\tau_1}$ for all i.*

3. *[Bounded Power] For some constants $c_2, \tau_2 > 0$, we have $\frac{Var(\|\mathbf{a}_i\|^2)}{\mathbb{E}^2[\|\mathbf{a}_i - \boldsymbol{\mu}\|^2]} \leq c_2 \cdot n^{-\tau_2}$ for all i .*

Assumption 3 summarizes the technical conditions that are essential in the proof of our main theorem. The first assumption on the tail of the distribution would enable us to exploit concentration inequalities for sub-exponential distributions. We allow the vector $\mathbf{a}_i$ to have a non-zero mean, yet we require the power of its mean to be small compared to the power of the random part of the vector. Intuitively, one would like the measurement vectors to sample diversely from all the directions in the $\mathbb{R}^n$, and not to be biased towards a specific direction. Finally, the last assumption is meant to control the dependencies among the entries of $\mathbf{a}_i$ and is used to prove concentration of $\frac{1}{n}\mathbf{a}_i^T\mathbf{M}\mathbf{a}_i$ around its mean, for a matrix $\mathbf{M}$ with bounded operator norm. For instance, for a Gaussian vector $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, this assumption reduces to finding constants $c_2, \tau_2 > 0$, such that $\frac{\|\boldsymbol{\Sigma}\|_F^2}{\|\boldsymbol{\Sigma}\|_\star^2} \leq c_2\tau^{-n}$.

We are now ready to state our main theorem which shows that the performance of the convex estimator $\mathcal{E}(\mathbf{x}_0, \mathbf{A}, \mathcal{S}, f(\cdot))$ is independent of the distribution of the measurement vectors. Hence, we can replace them with a Gaussian random vectors with the same mean and covariance.

Theorem 3. ***[non-Gaussian=Gaussian]*** *Consider the problem of recovering $\mathbf{x}_0 \in \mathcal{S} \subseteq \mathbb{R}^n$ from the measurements $\mathbf{y} = \mathbf{A}\mathbf{x}_0 \in \mathbb{R}^m$, using a convex penalty function $f(\cdot)$ in the estimator $\mathcal{E}\{\mathbf{x}_0, \mathbf{A}, \mathcal{S}, f(\cdot)\}$ in* (4.7)*. Assume $\mathcal{S}$ is a convex set and m and n are growing to infinity at a fixed rate $m = \theta(n)$. Also assume that*

1. *$f : \mathbb{R}^n \to \mathbb{R}$ is a convex function that satisfies Assumption 2.*
2. *The measurement matrix $\mathbf{A} = \left[\mathbf{a}_1, \ldots, \mathbf{a}_m\right]^T \in \mathbb{R}^{m \times n}$ satisfies Assumption 3, with $\boldsymbol{\mu} := \mathbb{E}[\mathbf{a}_i]$ and $\boldsymbol{\Sigma} := Cov[\mathbf{a}_i]$ for all $i = 1, \ldots, m$.*
3. *$\mathbf{G} = \left[\mathbf{g}_1, \ldots, \mathbf{g}_m\right]^T \in \mathbb{R}^{m \times n}$ is a random Gaussian matrix with independent rows drawn from Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.*

Then the estimator $\mathcal{E}\{\mathbf{x}_0, \mathbf{A}, \mathcal{S}, f(\cdot)\}$ (introduced in Definition 2) succeeds in recovering $\mathbf{x}_0$ with probability approaching one (as m and n grow large), if and only if the estimator $\mathcal{E}\{\mathbf{x}_0, \mathbf{G}, \mathcal{S}, f(\cdot)\}$ succeeds with probability approaching one.

Theorem 3 shows that (when the conditions of Assumption 3 are satisfied) only the mean and covariance of the measurement vectors $\mathbf{a}_i$ affect the required number of measurements for perfect recovery in (4.7).

It is straightforward that if we have sub-Gaussian measurements $\{\mathbf{a}_i\}_{i=1}^m$ that satisfy Assumption 1, then $\text{Vec}\left(\mathbf{a}_i\mathbf{a}_i^T\right)$ would satisfy Assumption 3. Therefore, Proposition 1 is just an immediate consequence of the result of Theorem 3.

Analysis of the Gaussian estimator

The universality result stated above indicates that when the assumptions are satisfied, one can simply replace the measurement vectors with ones with i.i.d. Gaussian distribution while the recovery threshold remains unchanged. Here we state a result on the performance of the Gaussian estimator.

The descent cone of a convex function $f(\cdot)$ at point $\mathbf{x}_0$ is defined as

$$\mathcal{D}_f(\mathbf{x}_0) = \text{Cone}\left(\{\mathbf{z} \ : \ f(\mathbf{x}_0+\mathbf{z}) \leq f(\mathbf{x}_0)\}\right) \ , \tag{4.8}$$

which is a convex cone. Here, $\text{Cone}(\mathcal{S})$ shows the conic-hull of the set $\mathcal{S}$.

The following lemma characterizes the required number of measurements for the equivalent Gaussian estimator.

Lemma 4. *Consider the problem of recovering the vector $\mathbf{x}_0 \in \mathcal{S}$, given the observations $\mathbf{y} = \mathbf{G}\mathbf{x}_0 \in \mathbb{R}^m$, via the estimator $\mathcal{E}\{\mathbf{x}_0, \mathbf{G}, \mathcal{S}, f(\cdot)\}$ introduced earlier. Assume that the rows of $\mathbf{G}$ are independent Gaussian random vectors with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma} = \mathbf{M}\mathbf{M}^T$. Let $\delta := m/n$ and the set $\mathcal{S}$ and the penalty function $f(\cdot)$ are convex. $\mathcal{E}\{\mathbf{x}_0, \mathbf{G}, \mathcal{S}, f(\cdot)\}$ succeed in recovering $\mathbf{x}_0$ with probability approaching one (as m and n grow to infinity), if and only if*

$$\sqrt{\delta} > \sqrt{\delta^\star} = \mathbb{E}\left[\max_{\substack{\mathbf{w}\in(\mathcal{S}-\mathbf{x}_0)\cap D_f(\mathbf{x}_0) \\ \frac{1}{\sqrt{n}}\mathbf{M}^T\mathbf{w}\in\mathbb{S}^{n-1}}} \frac{\mathbf{w}^T\mathbf{g}}{n\sqrt{1+\frac{1}{n}(\mathbf{w}^T\mu)^2}}\right] \tag{4.9}$$

where $\mathbb{S}^{n-1}$ is the n-dimensional unit sphere, and the expected value is over the Gaussian vector $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$.

PRECISE ANALYSIS OF (COMPLEX-VALUED) PHASEMAX: PHASE RETRIEVAL VIA LINEAR PROGRAMMING

In this chapter, we focus on analyzing a recently proposed convex-optimization-formulation for the complex phase retrieval problem known as *PhaseMax*. As explained and analyzed in the previous chapters, conventional convex-relaxation-based methods in phase retrieval resort to the idea of "lifting" which makes them computationally inefficient, since the number of unknowns is effectively squared. In contrast, PhaseMax is a novel convex relaxation that does not increase the number of unknowns. Instead it relies on an initial estimate of the true signal which must be externally provided. Here, we investigate the required number of measurements for exact recovery of the signal in the large system limit and when the linear measurement matrix is random with i.i.d. standard normal entries. If n denotes the dimension of the unknown complex signal and m the number of phaseless measurements, then in the large system limit we show that $\frac{m}{n} > \frac{4}{\cos^2(\theta)}$ measurements are **necessary and sufficient** to recover the signal with high probability, where θ is the angle between the initial estimate and the true signal. Our result indicates a sharp phase transition in the asymptotic regime which matches the empirical result in numerical simulations. Furthermore, from recent works in the literature, we provide some insights on how to find an efficient initialization via a method called spectral initialization.

The organization of this chapter is as follows. In Section 5.1, we provide motivations for this new convex relaxation as well as a discussion on some earlier works that analyzed PhaseMax. We mathematically setup the problem in Section 5.2. Consequently, in Section 5.3, we present our main result followed by discussions and the result of numerical simulations. Finally, Section 5.4 provides some discussion on spectral initialization. The results presented in this chapter are available in the research paper [112][1], and some of the texts appear as it is in the publication.

5.1 Motivations and Background

As explained earlier, the phase retrieval problem has a rich history and occurs in many areas in engineering and applied physics. In most of these cases, measuring the phase is either expensive or even infeasible. For instance, in some optical settings, detection devices like CCD cameras and

[1] F. Salehi, E. Abbasi, and B. Hassibi, "A Precise Analysis of Phasemax in Phase Retrieval." In: 2018 IEEE International Symposium on Information Theory (ISIT), IEEE, 2018, pp. 976–980.

photosensitive films cannot measure the phase of a light wave and instead measure the photon flux.

Reconstructing a signal from magnitude-only measurements is generally very difficult due to loss of important phase information. Therefore, phase retrieval faces fundamental theoretical and algorithmic challenges and a variety of methods were suggested [66]. Convex methods have recently gained significant attention to solve the phase retrieval problem. These methods are mainly based on semidefinite programming by linearizing the resulting quadratic constraints using the idea of *lifting* (e.g. see [27, 67] and references therein). Due to the convex nature of their formulation, these algorithms usually have rigorous theoretical guarantees. However, semidefinite relaxation squares the number of unknowns which makes these algorithms computationally complex, especially in large systems. This caveat makes these approaches intractable in real-world applications.

Introduced in two independent works [58, 9], *PhaseMax* is a recently proposed convex formulation for the phase retrieval problem in the original $n-$dimensional parameter space. This method maximizes a linear functional over a convex feasible set. The constrained set in this optimization is obtained by relaxing the non-convex equality constraints in the original phase retrieval problem to convex inequality constraints. To form the objective function, PhaseMax relies on an initial estimate of the true signal which must be externally provided.

The simple formulation of the PhaseMax method makes it appealing for practical applications. In addition, existing theoretical analysis indicates that this method achieves perfect recovery for a nearly optimal number of random measurements. The analysis in [58, 9, 60] suggests that $m > Cn$, where C is a constant that depends on the quality of initial estimate ($\mathbf{x}_{\text{init}}$), is the sufficient number of measurements for perfect signal reconstruction when the measurement vectors are drawn independently from the Gaussian distribution. The exact phase transition threshold, i.e. the exact value of the constant C, for the *real* PhaseMax has been recently derived in [41, 40]. However, for the practical case of complex signals, previous results could only provide an upper bound on C.

The main contribution of the results presented in this chapter is the characterization of the phase transition regimes for the perfect signal recovery in the (complex-valued) PhaseMax algorithm. Our result is asymptotic and assumes that the measurement vectors are derived independently from Gaussian distribution. To the extent of the author's knowledge, this is the first work that computes the exact phase transition bound of the (complex-valued) PhaseMax in phase retrieval.

In our analysis, we utilize the recently Convex Gaussian Min-max Theorem (CGMT) [130, 129] which uses Gaussian process methods. CGMT has been successfully applied in a number of different problems including the performance analysis of structured signal recovery in M-estimators [129, 3], massive MIMO [131, 1], etc. CGMT has been also used by Dhifallah et. al. [41] to analyze the

real version of the PhaseMax. The complex case, however, does not directly fit into the framework of CGMT. Therefore, in this chapter we introduce a secondary optimization that provably has the same phase transition bounds as PhaseMax and that also can be analyzed by CGMT. A detailed discussion and proof of the equivalence of the phase transition of the secondary optimization with the original PhaseMax optimization is provided in Section 5.5.

5.2 Problem Setup

Let $\mathbf{x}_0 \in \mathbb{C}^n$ denote the underlying signal. We consider the phase retrieval problem with the goal of recovering $\mathbf{x}_0$ from m magnitude-only measurements of the form,

$$b_j = |\mathbf{a}_j^\star \mathbf{x}_0|, \ j = 1, \dots, m. \tag{5.1}$$

Throughout this chapter, we assume that $\{\mathbf{a}_j \in \mathbb{C}^n\}_{j=1}^m$ is the set of known measurement vectors where the $\mathbf{a}_j$'s are independently drawn from the complex Gaussian distribution with mean zero and covariance matrix $\mathbf{I}$.

As mentioned earlier, the PhaseMax method relies on an initial estimate of the true signal. $\mathbf{x}_{\text{init}} \in \mathbb{C}^n$ is used to represent this initial guess. We assume that both $\mathbf{x}_0$ and $\mathbf{x}_{\text{init}}$ are independent of all the measurement vectors. The PhaseMax algorithm provides a convex formulation of the phase retrieval problem by simply relaxing the equality constraints in (5.1) into *convex* inequality constraints. This results in the following convex optimization problem:

$$\begin{aligned} \hat{\mathbf{x}} = \arg\max_{\mathbf{x}\in\mathbb{C}^n} \quad & \Re\{{\mathbf{x}_{\text{init}}}^* \, \mathbf{x}\} \\ \text{subject to:} \quad & |\mathbf{a}_j^* \mathbf{x}| \le b_j \ , \ \ 1 \le j \le m. \end{aligned} \tag{5.2}$$

This optimization searches for a feasible vector that possesses the most real correlation with $\mathbf{x}_{\text{init}}$. Note that because of the global phase ambiguity of the measurements in (5.1), we can estimate $\mathbf{x}_0$ up to a global phase. Therefore, we define the following performance measure for the PhaseMax method,

$$\mathbf{D}(\hat{\mathbf{x}}, \mathbf{x}_0) = \min_{\phi\in[-\pi,\pi]} \frac{\|\hat{\mathbf{x}} e^{j\phi} - \mathbf{x}_0\|}{\|\mathbf{x}_0\|} \ . \tag{5.3}$$

Under this setting, a perfect recovery of $\mathbf{x}_0$ means $\mathbf{D}(\hat{\mathbf{x}}, \mathbf{x}_0) = 0$. In this chapter, we investigate the necessary and sufficient conditions under which the optimization program (5.2) perfectly recovers the true signal.

5.3 Main Result

In this section, we present the main result of this chapter which provides us with the necessary and sufficient number of measurements for the perfect recovery of the PhaseMax method in (5.2) under different scenarios. First, we discuss some of the previous works that analyzed the recovery condition (in terms of the number of measurements) for the (real-valued) PhaseMax. These analyses have shown that in order to successfully recover the underlying signal $\mathbf{x}_0$, $O(n)$ measurements are needed. While these analyses share the same order complexity, they are not providing an exact bound. The exact phase transition for the real-valued PhaseMax has been computed in works by Dhifallah et al. [40, 41]. However, to the extent of the author's knowledge, the exact phase transition has not been computed prior to our work [112] (our work was first released in January 2018).

In what follows, after reviewing some of the prior results for the real setting, we present our main contribution, which is the precise phase transition of the PhaseMax algorithm for recoverying a complex-valued signal.

Prior work: recovery threshold for the real setting

As explained earlier the PhaseMax optimization has been introduced by two independent works [9, 58], where they analyzed the required number of measurements for the optimization program (5.2) to successfully recover the underlying signal. Here, we briefly explain these approaches.

Bahmani and Romberg [9] provided the first analysis for the recovery threshold by using some results from statistical learning theory. They provide an analysis when the measurements are corrupted with a bounded positive noise. In the noiseless setting, their result indicates that $m \overset{\rho_{\text{init}}}{\gtrsim} n$, where ρ_{init} defines the correlation between the underlying signal and the initial guess.

Golstein and Studer [58], who first coined the term PhaseMax, analyzed the problem for the setting where the measurement vectors are drawn from a uniform distribution on the sphere $\mathbb{S}_{\mathbb{C}}^{(n-1)}$. They translated the recovery condition into the problem of intersection of the feasible set and the ascent set (of the objective function), and finding the condition that these two sets have a zero intersection around $\mathbf{x}_0$. They showed that the successful recovery is possible when the number of measurements satisfies the following:

$$m > \frac{4n}{\gamma}\,, \tag{5.4}$$

where $\gamma := 1 - \frac{2}{\pi}\theta$, and $\theta := \operatorname{acos}(\rho_{\text{init}})$ is the angle between $\mathbf{x}_0$ and $\mathbf{x}_{\text{init}}$. We should note that this result is sharper than the earlier bound presented. A similar result has been derived by Hand and

Voroninski [60] using concentration inequalities.

The closest work to our analysis is the results presented by Dhifallah et al. [40, 41] where they presented a precise phase transition for real-valued PhaseMax in the asymptotic regime where $m.n \to \infty$ at a fixed oversampling ratio $\delta := \frac{m}{n} \in [0, \infty)$. They have shown that (provided $\delta > 2$) for isotropic Gaussian measurements, the necessary and sufficient condition for the successful recovery of the (real-valued) signal via PhaseMax is,

$$\frac{\pi}{\delta \tan(\frac{\pi}{\delta})} > 1 - \rho_{\text{init}}^2, \tag{5.5}$$

where ρ_{init} denotes the correlation between the $\mathbf{x}_0$ and $\mathbf{x}_{\text{init}}$. In [41], the authors have shown that by iteratively applying PhaseMax, a method referred to as PhaseLamp, a better recovery threshold would be achieved.

Table 5.1 provides a comparison between different recovery thresholds that have been reported in the previous works in literature.

Authors	**Sample Complexity**
Bahmani-Romberg'16	$\delta > \frac{32}{\sin^4 \gamma} \log(\frac{8e}{\sin^4 \gamma})$
Hand-Voroninski'16	$\delta > C_0(\theta)$
Goldstein-Studer'17	$\delta > \frac{4}{\gamma}$
Dhifallah et al.'17	$\frac{\pi}{\delta \tan(\frac{\pi}{\delta})} > 1 - \rho_{\text{init}}^2$

Table 5.1: Recovery thresholds of PhaseMax reported in prior works in the literature.

Precise phase transition for complex-valued PhaseMax

In this section, we present the main result of the chapter which provides us with the necessary and sufficient number of measurements for the perfect recovery of the PhaseMax method in (5.2) under different scenarios. Our result is asymptotic which assumes a fixed oversampling ratio $\delta := \frac{m}{n} \in [0, \infty)$, while $n \to \infty$. In theorem 4, we introduce δ_{rec} which depends on the problem parameters and prove that the condition $\delta > \delta_{\text{rec}}$ is necessary and sufficient for perfect recovery. Our result reveals significant dependence between δ_{rec} and the quality of the initial guess. We use the following similarity measure to quantify the caliber of the initial estimate:

$$\rho_{\text{init}} := \max_{0 \le \phi < 2\pi} \frac{\Re\{e^{j\phi}\, \mathbf{x}_{\text{init}}^*\, \mathbf{x}_0\}}{||\mathbf{x}_0||\, ||\mathbf{x}_{\text{init}}||} = \frac{|\mathbf{x}_{\text{init}}^*\, \mathbf{x}_0|}{||\mathbf{x}_0||\, ||\mathbf{x}_{\text{init}}||}. \tag{5.6}$$

Note that the multiplication by a unit amplitude scalar in the above definition is due to the global phase ambiguity of the phase retrieval solution (the true phase of $\mathbf{x}_0$ is dissolved in the absolute value in (5.1)). Therefore, for convenience we assume that both $\mathbf{x}_{\text{init}}$ and $\mathbf{x}_0$ are aligned unit norm vectors ($||\mathbf{x}_0|| = ||\mathbf{x}_{\text{init}}|| = 1$), which results in $\rho_{\text{init}} = \mathbf{x}_{\text{init}}^* \mathbf{x}_0$. We also define θ as the angle between $\mathbf{x}_{\text{init}}$ and $\mathbf{x}_0$, and therefore, $\rho_{\text{init}} = \cos\theta$. We now present the main result of the chapter which characterizes the phase transition regimes of PhaseMax for perfect recovery, in terms of δ and ρ_{init}.

Theorem 4. *Consider the PhaseMax problem defined in Section 5.2. For a fixed oversampling ratio $\delta = \frac{m}{n} > 4$, the optimization program* (5.2) *perfectly recovers the true signal (in the sense that $\lim_{n\to\infty} \mathbb{P}(\mathbf{D}(\hat{\mathbf{x}}, \mathbf{x}_0) > \epsilon) = 0$, for any fixed $\epsilon > 0$) if and only if,*

$$\delta > \delta_{rec} := \frac{4}{\cos^2\theta} = \frac{4}{\rho_{init}^2} , \tag{5.7}$$

where ρ_{init} is defined in (5.6).

Theorem 4 establishes a sharp phase transition behavior for the performance of PhaseMax. The inequality (5.7) can also be rewritten in terms of θ (or ρ_{init}) when the oversampling ratio, δ, is fixed,

$$\rho_{\text{init}} = \cos\theta > \sqrt{\frac{4}{\delta}} . \tag{5.8}$$

The proof of Theorem 4 consists of two main steps. First, we introduce a real optimization program with $2n - 1$ variables and prove that it has the same phase transition bounds as PhaseMax in (5.2). The point of this step is that this new real optimization is especially built in a way that its performance can be precisely analyzed using well known tools like CGMT. Therefore, the next step would be to apply the CGMT framework to the new real optimization and to derive its phase transition bounds. We postpone a detailed version of the proof to Section 5.5. The following remarks are in place.

Remark 2. *The condition $\delta > 4$ is proven to be fundamentally necessary for the phase retrieval problem under generic measurements to have a unique solution [36]. This is consistent with Theorem 4 where you can observe that even in the best scenario where $\mathbf{x}_{init}$ is aligned with $\mathbf{x}_0$, we still need $m > 4n$ measurements for PhaseMax to have $\mathbf{x}_0$ as the solution. On the other hand, in the case where $\mathbf{x}_{init}$ carries no information about $\mathbf{x}_0$ ($\mathbf{x}_{init}$ is orthogonal to $\mathbf{x}_0$), recovery of $\mathbf{x}_0$ by PhaseMax is not guaranteed regardless of the number of measurements.*

Remark 3. *It is shown in [58] that $\delta > \frac{4}{1-2\theta/\pi}$ is sufficient for perfect recovery of $\mathbf{x}_0$. This bound is compared to our result in Figure 5.1 which shows phase transition regions of PhaseMax derived*

from empirical results. Although the simulations are run on the signals of size $n = 128$, one can see that the blue line derived from Theorem 4, perfectly predicts the phase transition boundary.

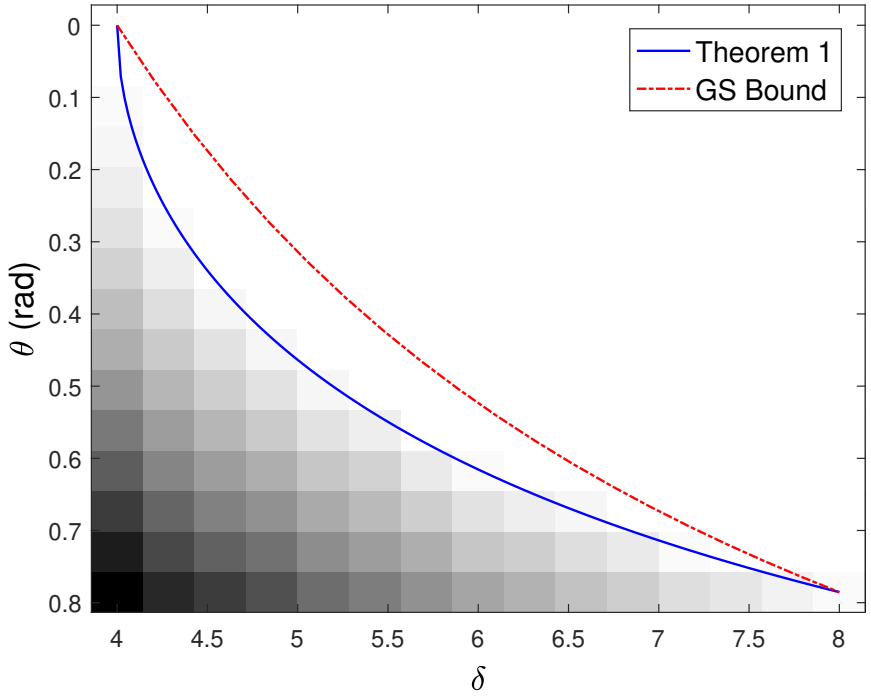


Figure 5.1: Phase transition regimes for the (complex-valued) PhaseMax problem in terms of the oversampling ratio $\delta = \frac{m}{n}$ and θ, the angle between $\mathbf{x}_0$ and $\mathbf{x}_{\text{init}}$. For the empirical results, we generated 10 independent realizations of the measurement vectors with $n = 128$. The blue line indicates the sharp phase transition bounds derived in Theorem 4 and the red line comes from the results of [58], which is referred to as the GS Bound.

5.4 Spectral Initialization

As seen in the previous section, an important ingredient of the PhaseMax optimization is an initialization step to generate the initial guess ($\mathbf{x}_{\text{init}}$) which determines the objective function in (5.2). In addition to PhaseMax, many of the iterative optimization methods which attempt to directly solve the non-convex phase retrieval formulation [97, 34, 145] require an initial estimate of that is well-aligned with the signal. Spectral methods are widely used for generating such initial vectors, and are widely used in generalized linear models (e.g. [86, 70]).

Here, we focus on the *real setting* (i.e., real-valued signal and measurement vectors). Similarly to the previous section, cosine squared similarity is used to measure the alignment of the initial

estimate derived from spectral initialization, say $\mathbf{x}_{\text{init}}$, with the underlying signal, $\mathbf{x}_0$.

$$\rho(\mathbf{x}_{\text{init}}, \mathbf{x}_0)^2 = \frac{|\mathbf{x}_{\text{init}}^T \mathbf{x}_0|^2}{\|\mathbf{x}_0\|^2 \|\mathbf{x}_{\text{init}}\|^2} \tag{5.9}$$

We will assume that our measurement vectors are independently and identically distributed according to the standard normal distribution, $a_i \sim \mathcal{N}(0, I_n)$. We will also assume, without loss of generality, that $||\mathbf{x}_0|| = 1$. The measurement information we have available to us is $y_i = (a_i^T \mathbf{x}_0)^2 = b_i^2$ for measurements $i = 1, ..., m$.

It is worth noting that achieving a cosine similarity that is greater than a positive constant is challenging especially in high dimensions. For instance, if we choose $\mathbf{x}_{\text{init}}$ uniformly at random from $\mathbb{S}^{n-1}$, then with high probability, the correlation would be of $O(\sqrt{\frac{1}{n}})$, which goes to zero as $n \to \infty$.

Consider the matrix $\mathbf{D}_m \equiv \frac{1}{m} \sum_{i=1}^m T(y_i) \mathbf{a}_i \mathbf{a}_i^T$, where $T : \mathbb{R} \to \mathbb{R}$ is some function defined so that the leading eigenvector of $\mathbf{D}_m$ corresponds to $\mathbf{x}_0$ in the limit as the number of measurements m goes to infinity. First, we will show that as $m \to \infty$, the leading eigenvector of $\mathbf{D}_m$ corresponds to $\mathbf{x}_0$ for some suitable function $T(\cdot)$. As $m \to \infty$, due to the law of large numbers, we can see that $\mathbf{D}_m \to \bar{\mathbf{D}}_m = \mathbb{E}[T(y) a a^T]$, where $\mathbf{a} \sim \mathcal{N}(0, \mathbf{I}_n)$ and $y = (\mathbf{a}^T \mathbf{x}_0)^2$. To calculate the converging matrix $\bar{\mathbf{D}}_m$, we simply consider its action on an orthonormal basis for $\mathbb{R}^n$.

Let $\mathcal{B} = \{\mathbf{x}_0, \mathbf{z}_1, ..., \mathbf{z}_{n-1}\}$ be an orthonormal basis of $\mathbb{R}^n$, where $z_1, ..., z_{n-1}$ are all orthogonal to the underlying vector $\mathbf{x}_0$. Presenting $\bar{\mathbf{D}}_m$ in the $\mathcal{B}$ has as its first entry $\mathbf{x}_0^T \bar{\mathbf{D}}_m \mathbf{x}_0$, and $\mathbf{z}_i^T \bar{\mathbf{D}}_m \mathbf{z}_i$ for $i = 1, ..., n-1$ as the rest of its diagonal entries. We have:

$$\mathbf{x}_0^T \bar{\mathbf{D}}_m \mathbf{x}_0 = \mathbf{x}_0^T \mathbb{E}[T(y) \mathbf{a}\mathbf{a}^T] \mathbf{x}_0 = \mathbb{E}[T(y)(\mathbf{a}^T \mathbf{x}_0)^2] = \mathbb{E}[T(y) y],$$

and for $i = 1, 2, \ldots, n-1$,

$$\mathbf{z}_i^T \bar{\mathbf{D}}_m \mathbf{z}_i = \mathbf{z}_i^T \mathbb{E}[T(y) \mathbf{a}\mathbf{a}^T] \mathbf{z}_i = \mathbb{E}[T(y)(\mathbf{a}^T \mathbf{z}_i)^2] = \mathbb{E}[T(y)(\mathbf{a}^T \mathbf{z}_i)^2] \ .$$

Due to the isotropic Gaussian distribution of $\mathbf{a}$, we have that $\mathbf{a}^T \mathbf{z}_i$ has a standard normal distribution and is independent of y, hence,

$$\mathbf{z}_i^T \bar{\mathbf{D}}_m \mathbf{z}_i = \mathbb{E}[T(y)] \mathbb{E}[(\mathbf{a}^T \mathbf{z}_i)^2] = \mathbb{E}[T(y)] \mathbb{E}[y].$$

It can be shown that the off-diagonal entries of $\bar{\mathbf{D}}_m$ are all zero. Therefore, we can write:

$$\bar{\mathbf{D}}_m = \mathbb{E}[T(y)] \mathbb{E}[y] \ \mathbf{I}_n + \text{cov}(y, T(y)) \mathbf{x}_0 \mathbf{x}_0^T. \tag{5.10}$$

Hence, the leading eigenvector of $\bar{\mathbf{D}}_m$ is $\mathbf{x}_0$ if and only if $\text{cov}(y, T(y)) > 0$.

Now that we characterized the condition under which the converging matrix $\bar{\mathbf{D}}_m$ has $\mathbf{x}_0$ as its leading eigenvector, the important question to ask is how many measurements are needed in order for $\mathbf{D}_m$ to be close to its expected value $\bar{\mathbf{D}}_m$. Chen and Candes [34] has shown that $\mathcal{O}(n)$ is sufficient for spectral initialization to give an estimate with absolute positive alignment with the underlying signal. This is an improvement to earlier results in [97, 25] where $m = \mathcal{O}(npolylog(n))$ was reported as a sufficient recovery condition.

Precise characterization of spectral initialization performance

A recent paper by Lu and Li [88] provides a precise characterization of the performance of the spectral initialization method under isotropic Gaussian measurements. Here we briefly review their main result.

In their analysis, they analyze the leading eigenvector of $\mathbf{D}_m$ in the linear asymptotic regime where $m, n \to \infty$ at fixed ratio $\alpha := \frac{m}{n}$. Defining $\kappa := \|\mathbf{x}_0\| > 0$ and $s \sim \mathcal{N}(0, 1)$, $\mathbb{P}(y|\kappa s) = f(y|\kappa s)$. They also assume that $z = T(y)$ has a bounded support $[0, \tau]$, and $\text{cov}(z, s^2) > 0$ is needed to ensure that $\mathbf{x}_0$ is the leading eigenvector of $\mathbf{D}_m$. Their analysis relies on two helper functions $\phi,\ \psi_\alpha : [\tau, +\infty) \to \mathbb{R}$ defined as,

$$\phi(\lambda) := \lambda \cdot \mathbb{E}[\frac{zs^2}{\lambda - z}] \ , \ \psi_\alpha(\lambda) := \lambda \left(\frac{1}{\alpha} + \mathbb{E}\frac{z}{\lambda - z} \right), \tag{5.11}$$

where z and s are defined above.The following theorem indicates the precise phase transition of the spectral initialization:

Theorem 5 (Theorem 1 in [88])**.** *Consider the spectral initialization with the aforementioned assumption and let ζ_α be a function defined as,*

$$\zeta_\alpha(\lambda) := \psi_\alpha \left(\max(\lambda, \bar{\lambda}_\alpha) \right) \ , \ \textit{where} \quad \bar{\lambda}_\alpha := \arg\min_{\lambda > \tau} \psi_\alpha(\lambda).$$

Then, the equation $\zeta_\alpha(\lambda) = \phi(\lambda)$ has a unique solution in $\lambda > \tau$, $\lambda^\star_\alpha$. Let $\mathbf{x}_{init}$ be the leading eigenvalue of the data matrix $\mathbf{D}_m$. As $n \to \infty$,

$$\rho\left(\mathbf{x}_{init}, \mathbf{x}_0\right) \xrightarrow{\mathbb{P}} \begin{cases} 0 & \text{if } \psi'(\lambda^\star_\alpha) < 0, \\ \sqrt{\frac{\psi'(\lambda^\star_\alpha)}{\psi'(\lambda^\star_\alpha) - \phi(\lambda^\star_\alpha)}} & \text{if } \psi'(\lambda^\star_\alpha) > 0. \end{cases} \tag{5.12}$$

The above result indicates an asymptotic phase transition in terms of $\psi'(\lambda^\star_\alpha)$. It turns out that this indeed imposes a phase transition for α, i.e., the cosine similarity converges to a constant bigger than zero iff $\alpha > C\big(\kappa, T, f\big)$.

Remark 4. *As indicated by the above analysis, in order to obtain a reasonable initial guess for the PhaseMax, it is necessary and sufficient to have $O(n)$ measurements. Our result in Theorem 4 indicates that when $\rho_{init} > 0$, the PhaseMax optimization can recover the underlying signal with $\frac{4n}{\rho_{init}^2}$ measurements. Combining these two results indicates that under Gaussian measurement scheme, one can achieve a perfect recovery for (noiseless) phase retrieval with $O(n)$ measurements. However, note that in order to apply our result from Theorem 4, we need the measurement vectors to be independent from $\mathbf{x}_{init}$. Therefore, we need to use a different subset of measurements for the initialization.*

5.5 Proof of Theorem 4

In this section, we introduce the main ideas used in the proof of Theorem 4. As mentioned earlier in section 5.3, we assume $\mathbf{x}_0$ is a unit norm vector aligned with $\mathbf{x}_{\text{init}}$. Due to the rotational invariance of the Gaussian distribution, without loss of generality, we assume $\mathbf{x}_0 = \mathbf{e_1}$, the first vector of the standard basis in $\mathbb{C}^n$. Furthermore, the optimization program (5.2) is scalar invariant. So, we can assume $\|\mathbf{x}_{\text{init}}\| = 1$.

The proof consists of two main steps: In the first step, we analyze the complex optimization problem (5.2) and find the necessary and sufficient condition under which $\hat{\mathbf{x}} = \mathbf{x}_0$. Consequently, we use this condition to build an equivalent real optimization problem. Lemma 8 introduces this equivalent real optimization ERO, in $\mathbb{R}^{2n-1}$, and states that the perfect recovery in the PhaseMax algorithm occurs if and only if the all-zero vector is the unique minimizer of the ERO.

In the second step, we adopt the CGMT framework to analyze the ERO and investigate the conditions on ρ_{init} (or θ) under which the unique answer to the ERO is $\mathbf{0}$. Therefore, as a result of Lemma 8, these conditions will guarantee the perfect recovery in the initial PhaseMax optimization (5.2).

Introducing the equivalent real optimization (ERO)

We define the error vector $\mathbf{w} := \mathbf{x} - \mathbf{e}_1$ and rewrite (5.2) in terms of $\mathbf{w}$,

$$\begin{aligned} \max_{\mathbf{w}\in\mathbb{C}^n} \quad & \Re\{{\mathbf{x}_{\text{init}}}^* \, \mathbf{w}\} \\ \text{subject to:} \quad & |\mathbf{a}_i^*(\mathbf{e}_1 + \mathbf{w})| \le b_i \ , \quad 1 \le i \le m. \end{aligned} \tag{5.13}$$

For $i = 1, 2, \ldots, m$, we use $\phi_i := \angle({\mathbf{a_i}}^*\mathbf{x}_0)$ to define aligned measurement vectors $\tilde{\mathbf{a}}_i := e^{j\phi_i}\mathbf{a}_i$. Therefore, we have,

$$b_i = \tilde{\mathbf{a}}_i^*\mathbf{x}_0 = (\tilde{\mathbf{a}}_i)_1, \quad \text{for } \ i = 1, 2, \ldots, m \ , \tag{5.14}$$

where $(\tilde{\mathbf{a}}_i)_1$ is the first entry of $\tilde{\mathbf{a}}_i$. Let $\mathcal{D}$ be the set of all directions $\mathbf{w}$ with non-negative objective value, i.e.,

$$\mathcal{D} := \{\mathbf{w} \in \mathbb{C}^n : \Re\{\mathbf{x}_{\text{init}}^* \, \mathbf{w}\} \geq 0\}.$$

Also, we define the set $\mathcal{F}$ to represent the feasible set of the optimization problem (5.13).

$$\mathcal{F} := \{\mathbf{w} \in \mathbb{C}^n : |\mathbf{a}_i^*(\mathbf{e}_1 + \mathbf{w})| \leq b_i, \text{ for } i = 1, 2, \ldots, m\}.$$

The following lemmas show necessary and sufficient conditions for perfect recovery in PhaseMax, based on these notations.

Lemma 5. $\mathbf{x}_0$ *is the unique optimal solution of* (5.2) *if and only if* $\mathcal{D} \bigcap \mathcal{F} = \{\mathbf{0}\}$.

Proof. For $\mathbf{w} \in \mathcal{D} \bigcap \mathcal{F}$, $\mathbf{x}_0 + \mathbf{w}$ is a solution of (5.2) with an objective value greater than the value for $\mathbf{x}_0$. Therefore, $\mathcal{D} \bigcap \mathcal{F} = \{\mathbf{0}\}$ is equivalent to $\mathbf{x}_0$ being a local minimizer of (5.2) which is also a global minimum due to the convexity of the problem. □

Lemma 6. $\mathcal{D} \bigcap \mathcal{F} = \{\mathbf{0}\}$ *if and only if* $\mathcal{D} \bigcap cone(\mathcal{F}) = \{\mathbf{0}\}$.

Proof. Note that $\mathcal{D} \subset \mathbb{C}^n$ is a convex cone and $\mathcal{F} \subset \mathbb{C}^n$ is a convex set. The proof is the consequence of the following equality,

$$\mathcal{D} \bigcap \text{cone}(\mathcal{F}) = \text{cone}(\mathcal{D} \bigcap \mathcal{F}). \tag{5.15}$$

□

Lemma 7. $cone(\mathcal{F}) = \bigcap_{i=1}^m \{\mathbf{w} \in \mathbb{C}^n : \Re\{\tilde{\mathbf{a}}_i^* \, \mathbf{w}\} \leq 0\}$.

Proof. Let $\mathbf{d} \in \mathcal{F}$,

$$|b_i + \tilde{\mathbf{a}}_i^* \mathbf{d}| \leq b_i \ , \quad \text{for} \ \ i = 1, 2, \ldots, m. \tag{5.16}$$

Therefore,

$$\begin{aligned} \Re\{\tilde{\mathbf{a}}_i^* \mathbf{d}\} \ &= \Re\{b_i + \tilde{\mathbf{a}}_i^* \mathbf{d}\} - b_i \ , \\ &\leq |b_i + \tilde{\mathbf{a}}_i^* \mathbf{d}| - b_i \ , \\ &\leq 0 \ . \end{aligned} \tag{5.17}$$

This shows that $\text{cone}(\mathcal{F}) \subseteq \bigcap_{i=1}^{m}\{\mathbf{w} \in \mathbb{C}^n : \Re\{\tilde{\mathbf{a}}_i^*\mathbf{w}\} \leq 0\}$. To show the other direction, choose $\mathbf{d} \in \mathbb{C}^n$ such that: $\Re\{\tilde{\mathbf{a}}_i^*\mathbf{d}\} < 0$, for $i = 1, 2, \ldots, m$. One can show that there exists $R > 0$, such that for all $r \leq R$, $r\mathbf{d} \in \mathcal{F}$. Therefore, $\mathbf{d} \in \text{cone}(\mathcal{F})$. This concludes the proof.

□

We have the following corollary as a result of Lemma 5, Lemma 6, and Lemma 7.

Corollary 2. $\mathbf{x}_0$ *is the unique optimal solution of* (5.2) *if and only if,*

$$\{\mathbf{w} : \Re\{\mathbf{x}_{init}^*\mathbf{w}\} \geq 0 \ \textit{and,} \ \Re\{\tilde{\mathbf{a}}_i^*\mathbf{w}\} \leq 0, \ \textit{for} \ 1 \leq i \leq m\} = \{\mathbf{0}\}. \tag{5.18}$$

We are now ready to establish the equivalent real optimization (ERO). We will show that the following optimization has the exact phase transition bounds as PhaseMax in (5.2).

$$\begin{aligned} \max_{\mathbf{w}' \in \mathbb{R}^{2n-1}} \quad & \boldsymbol{\eta}^T \ \mathbf{w}' \\ \text{subject to:} \quad & |\mathbf{a}_i'^T(\mathbf{e}_1 + \mathbf{w}')| \leq b_i \ , \quad 1 \leq i \leq m, \end{aligned} \tag{ERO}$$

where $\mathbf{e}_1$ is the first vector of the standard basis in $\mathbb{R}^{2n-1}$, $\boldsymbol{\eta}$ and $\{\mathbf{a}_i'\}_{i=1}^m$ are $(2n-1)$ dimensional real vectors defined as

$$\eta := \begin{bmatrix} \Re\{\mathbf{x}_{\text{init}}\} \\ -\Im\{\mathbf{x}_{\text{init}}(2:n)\} \end{bmatrix} \text{ and } \mathbf{a}_i' := \begin{bmatrix} \Re\{\tilde{\mathbf{a}}_i\} \\ -\Im\{\tilde{\mathbf{a}}_i(2:n)\} \end{bmatrix}, \ \forall i. \tag{5.19}$$

Here $\Im\{\tilde{\mathbf{a}}_i(2:n)\}$ is the imaginary part of the last $n-1$ entries of $\tilde{\mathbf{a}}_i$.

Lemma 8. $\mathbf{x}_0$ *is the unique optimal solution of the PhaseMax method if and only if* $\mathbf{w}' = \mathbf{0}$ *is the unique optimal solution of* (ERO)*.*

The proof of Lemma 8 is straightforward by defining

$$\mathbf{w}' = \begin{bmatrix} \Re\{\mathbf{w}\} \\ \Im\{\mathbf{w}(2:n)\} \end{bmatrix} \in \mathbb{R}^{2n-1} \ , \tag{5.20}$$

and then showing that the optimality conditions for $\mathbf{w}' = \mathbf{0}$ in (ERO) is equivalent to (5.18).

It is worth mentioning that the result of Lemma 8 is valid for any set of measurement vectors $\{\mathbf{a}_i\}$. In the next part, we use this result to compute the phase transition of PhaseMax when the measurement vectors are drawn independently from the Gaussian distribution.

In the next section, we use the asymptotic CGMT (Lemma 32 in Appendix A.1) to analyze the ERO. To this end, we need to reqrite the ERO in the form of the primaty optimization that is a bilinear form with respect to an i.i.d Gaussian matrix. This enables us to apply Lemma 32 to the ERO and derive an Auxiliary Optimization in the form of (AO). Lemma 32 indicates that if $\|\mathbf{w}'\| \xrightarrow{\mathbb{P}} 0$ for the (AO), then the same also holds for the ERO and we have perfect recovery. We consequently analyze the (AO) using conventional concentration results in high dimensions.

Computing the phase transition for PhaseMax

In this part we adopt the CGMT framework along with the result of Lemma 8 to compute the exact phase transition of the PhaseMax algorithm under the Gaussian measurement scheme.

We start by calculating the distribution of the entries of $\mathbf{a}_i'$ that are defined in (5.19). Recall that $\mathbf{a}_i$'s are independently drawn from the complex Gaussian distribution with mean zero and covariance matrix $\mathbf{I}_n$. Therefore, the distribution of the entries of $\tilde{\mathbf{a}}_i$'s that were are defined above has the following properties:

(i) The first entry of $\tilde{\mathbf{a}}_i$ is the absolute value of the first entry of the $\mathbf{a}_i$. Therefore, it has a Rayleigh distribution, i.e.,

$$(\tilde{\mathbf{a}}_i)_1 \sim \mathcal{R}(1), \tag{5.21}$$

(ii) The remaining entries of $\tilde{\mathbf{a}}_i$ remain standard Gaussian random variables,

$$(\tilde{\mathbf{a}}_i)_k \sim \mathcal{N}_{\mathbb{C}}(0,1), \quad \text{for } 2 \le k \le n\ , \tag{5.22}$$

(iii) The entries of $\tilde{\mathbf{a}}_i$ remain independent.

This implies that all the entries of $\mathbf{a}_i'$ are independent, the first entry of $\mathbf{a}_i'$ has a $\mathcal{R}(1)$ distribution and the rest of the entries have Gaussian distribution $\mathcal{N}(0, \frac{1}{2})$. We form the measurement matrix $\mathbf{A} \in \mathbb{R}^{m\times(2n-1)}$ by row-stacking vectors $\{\mathbf{a_i'}^T,\ 1 \le i \le m\}$. Let $\mathbf{A}_1 \in \mathbb{R}^m$ be the first column of $\mathbf{A}$, and $\bar{\mathbf{A}} \in \mathbb{R}^{m\times(2n-2)}$ be the remaining part (i.e., $\mathbf{A} = \begin{bmatrix}\mathbf{A}_1 & \bar{\mathbf{A}}\end{bmatrix}$). $\mathbf{x}_0 = \mathbf{e}_1$ implies that $\mathbf{A}_1 = \begin{bmatrix}b_1, b_2, \ldots, b_m\end{bmatrix}^T$, where b_i's are defined in (5.1). Using the Lagrange multipliers, we can reformulate (ERO) as the following min-max program,

$$\min_{\substack{w_1\in\mathbb{R}\\ \bar{\mathbf{w}}\in\mathbb{R}^{2n-2}}} \max_{\boldsymbol{\lambda},\boldsymbol{\mu}\in\mathbb{R}_+^m} -\boldsymbol{\eta}^T\mathbf{w} + (\boldsymbol{\lambda}-\boldsymbol{\mu})^T\bar{\mathbf{A}}\bar{\mathbf{w}} - (\boldsymbol{\lambda}+\boldsymbol{\mu})^T\mathbf{A}_1 + (\boldsymbol{\lambda}-\boldsymbol{\mu})^T\mathbf{A}_1(1+w_1), \tag{5.23}$$

where w_1 denotes the first entry of $\mathbf{w}$ and $\bar{\mathbf{w}}$ represents the remaining entries. Define $\mathbf{v} := \boldsymbol{\lambda} - \boldsymbol{\mu}$. It can be shown that optimal values of (5.23) satisfy $\boldsymbol{\lambda} + \boldsymbol{\mu} = |\boldsymbol{\lambda} - \boldsymbol{\mu}|$. Here, $|\cdot|$ denotes the component-wise absolute value. Therefore, (5.23) can be rewritten as an optimization over $\mathbf{v} \in \mathbb{R}^m$ and $\mathbf{w} \in \mathbb{R}^{2n-1}$ in the following form:

$$\min_{\substack{w_1 \in \mathbb{R} \\ \bar{\mathbf{w}} \in \mathbb{R}^{2n-2}}} \max_{\mathbf{v} \in \mathbb{R}^m} -\boldsymbol{\eta}^T \mathbf{w} + \mathbf{v}^T \tilde{\mathbf{A}} \bar{\mathbf{w}} + \mathbf{v}^T \mathbf{A}_1 (1 + w_1) - |\mathbf{v}|^T \mathbf{A}_1. \tag{5.24}$$

Note that $\bar{\mathbf{A}}$ has i.i.d. standard normal entries. One can check that (5.24) satisfies the condition of Lemma 32, i.e., it is convex w.r.t. $\mathbf{w}$ and concave w.r.t $\mathbf{v}$, and all the terms outside the bilinear form are independent of $\bar{\mathbf{A}}$. Hence, we can form the (AO) as follows,

$$\min_{\substack{w_1 \in \mathbb{R} \\ \bar{\mathbf{w}} \in \mathbb{R}^{2n-2}}} \max_{\mathbf{v} \in \mathbb{R}^m} -\boldsymbol{\eta}^T \mathbf{w} + \mathbf{v}^T \mathbf{g} ||\bar{\mathbf{w}}|| + ||\mathbf{v}|| \mathbf{h}^T \tilde{\mathbf{w}} + \mathbf{v}^T \mathbf{A}_1 (1 + w_1) - |\mathbf{v}|^T \mathbf{A}_1, \tag{5.25}$$

where $\mathbf{g} \in \mathbb{R}^m$ and $\mathbf{h} \in \mathbb{R}^{2n-2}$ with entries drawn independently from standard normal distribution. Analysis of (5.25) is skipped here as a similar analysis would be provided in Chapter 6. We conclude this chapter with a theorem that characterizes the performance of the (ERO).

Let $\mathbf{w}^\star$ be the optimizer of (5.25) and define $s^\star := 1 + w_1^\star$ and $t^\star := ||\bar{\mathbf{w}}^\star||$. $s^\star$ simply denotes the first entry of the optimal solution and $t^\star$ indicates the norm of the remaining entries. In order to have a perfect recovery in the PhaseMax optimization (5.2), we should find conditions under which $(t^\star, s^\star) = (0, 1)$ would be achieved by the optimal solution of (AO). The following result characterizes the performance of the (AO) (i.e., optimization (5.25)).

Theorem 6. *In the asymptotic regime where $m, n \to \infty$, and $\delta := \frac{m}{n}$, $s^\star$ and $t^\star$ converge to the solution of the following deterministic optimization,*

$$\begin{aligned} \max_{s \in [-1,1],\ t \geq 0} \ & \rho_{init}\ s + \sqrt{1 - \rho_{init}{}^2} \sqrt{t^2 - \frac{\delta}{2} p(t, s)} \\ s.t.\ & p(t, s) \leq \frac{2t^2}{\delta}. \end{aligned} \tag{5.26}$$

In the above optimization, $p(t, s)$ is defined as,

$$p(t, s) = t^2 + (1 + s)[1 + s - \sqrt{t^2 + (1 + s)^2}] + (1 - s)[1 - s - \sqrt{t^2 + (1 - s)^2}]. \tag{5.27}$$

It can be shown that $\rho_{\text{init}} > \frac{2}{\sqrt{\delta}}$ is the necessary and sufficient condition for $(t^\star, s^\star) = (0, 1)$ to be the unique solution of (5.26) which is equivalent to the perfect recovery in the (ERO).

ACHIEVING OPTIMAL SAMPLE COMPLEXITY VIA REGULARIZED PHASEMAX

[1] F. Salehi et al. "Learning without the Phase: Regularized PhaseMax Achieves Optimal Sample Complexity". In: *Advances in Neural Information Processing Systems* (2018), pp. 8641–8652.

The problem of estimating an unknown signal, $\mathbf{x}_0 \in \mathbb{R}^n$, from a vector $\mathbf{y} \in \mathbb{R}^m$ consisting of m magnitude-only measurements of the form $y_i = |\mathbf{a}_i \mathbf{x}_0|$, where $\mathbf{a}_i$'s are the rows of a known measurement matrix $\mathbf{A}$, is a classical problem known as phase retrieval. This problem arises when measuring the phase is costly or altogether infeasible. In many applications in machine learning, signal processing, statistics, etc., the underlying signal has certain structure (sparse, low-rank, finite alphabet, etc.), opening up the possibility of recovering $\mathbf{x}_0$ from a number of measurements smaller than the ambient dimension, i.e., $m < n$. Ideally, one would like to recover the signal from a number of phaseless measurements that is on the order of the "degrees of freedom" of the structured signal, $\mathbf{x}_0$.

To this end, inspired by the PhaseMax algorithm [58, 9], we formulate a convex optimization problem, where the objective function relies on an initial estimate of the true signal and also includes an additive regularization term to encourage structure. The new formulation is referred to as **regularized PhaseMax**. We analyze the performance of regularized PhaseMax to find the minimum number of phaseless measurements required for perfect signal recovery. The results are asymptotic and are in terms of the geometrical properties (such as the Gaussian width) of certain convex cones. When the measurement matrix has i.i.d. Gaussian entries, we show that our proposed method is indeed order-wise optimal, allowing perfect recovery from a number of phaseless measurements that is only a constant factor away from the optimal number of measurements required when phase information is available. We explicitly compute this constant factor, in terms of the quality of the initial estimate, by deriving the exact phase transition. The theory well matches empirical results in our numerical simulations.

6.1 Motivation and Background

Recovering an unknown signal or model given a limited number of linear measurements is an important problem that appears in many applications. Researchers have developed various methods with rigorous theoretical guarantees for perfect signal reconstruction, e.g. [13, 44, 122, 134]. However, there are many practical scenarios in which the signal should be reconstructed from nonlinear measurements. In particular, in many physical devices, measuring the phase is expensive or even infeasible. For instance, detection devices such as CCD cameras and photosensitive films cannot measure the phase of a light wave and instead measure the photon flux [66].

As explained earlier, the fundamental problem of recovering a signal from magnitude-only measurements is known as *phase retrieval*. This problem has a rich history and occurs in many areas in engineering and applied sciences such as medical imaging [6], X-ray crystallography [93], astronomical imaging [52], and optics [144]. Due to the loss of phase information, signal reconstruction from magnitude-only measurements can be quite challenging. Therefore, despite a variety of proposed methods and analysis frameworks, phase retrieval still faces fundamental theoretical and algorithmic challenges.

Recently, convex methods have gained significant attention to solve the phase retrieval problem. As explained in Chapter 2, the first convex-relaxation-based methods were based on semidefinite programs [27, 25] and resorted to the idea of *lifting* [8, 22, 68, 115] the signal from a vector to a matrix to linearize the quadratic constraints. While the convex nature of this formulation allows theoretical guarantees, the resulting algorithms are computationally inefficient since the number of unknowns is effectively squared. This makes these approaches intractable when the system dimension is large. The PhaseMax, that was introduced in the Chapter 5, is a novel convex relaxation for phase retrieval which works in the original n-dimensional parameter space. Since it does not require lifting and does not square the number of unknowns, it is appealing in practice. It does, however, require an initial estimate of the signal. The exact phase transition for PhaseMax has been explored in details in Chapter 5.

Non-convex methods for phase retrieval have a long history [53]. Recent non-convex methods start with a careful initialization [89, 94] and update the solution iteratively using a gradient-descent-like scheme. Examples of such methods include Wirtinger flow algorithms [24, 34, 119], truncated amplitude flow [145], and alternating minimization [97, 150]. Despite having lower computational cost, precise theoretical analysis of such algorithms seems very technically challenging.

All the aforementioned algorithms essentially demonstrate that a signal of dimension n can be perfectly recovered through $m > Cn$ amplitude-only measurements, where $C > 1$ is a constant that

depends on the algorithm as well as the measurement vectors. However, many interesting signals in practice contain fewer degrees of freedom than the ambient dimension (sparse signals, low-rank matrices, finite alphabet signals, etc.). Such low-dimensional structures open up the possibility of perfect signal recovery with a number of measurements significantly smaller than n.

Summary of contributions

In this chapter, we propose a new approach for recovering *structured* signals. Inspired by the PhaseMax algorithm, we introduce a new convex formulation and investigate necessary and sufficient conditions, in terms of the number of measurements, for perfect recovery. We refer to this new framework as *regularized PhaseMax*. The constrained set in this optimization is obtained by relaxing the non-convex equality constraints in the original phase retrieval problem to convex inequality constraints. The objective function consists of two terms. One is a linear functional that relies on an initial estimate of the true signal which must be externally provided. The second term is an additive regularization term that is formed based on a priori structural information about the signal.

We precisely compute the necessary and sufficient number of measurements for perfect signal recovery when the entries of the measurement matrix are i.i.d. Gaussian. To the extent of our knowledge, this is the first convex optimization formulation for the problem of structured signal recovery given phaseless linear Gaussian measurements that provably requires an order optimal number of measurements. The focus of this chapter is on real signals and real measurements. The complex case is more involved, requires a different analysis, and will be considered as an interesting future direction.

Through our analysis, we make the following main contributions:

- We first provide a sufficient recovery condition, in terms of the number of measurements, for perfect signal recovery. We use this to infer that our proposed method is order-wise optimal.
- We characterize the exact phase transition behavior for the class of absolutely scalable regularization functions.
- We apply our findings to two special examples: unstructured signal recovery and sparse recovery. We observe that the theory well matches the result of numerical simulations for these two examples.

Prior work

Phase retrieval for structured signals has gained significant attention in recent years. We briefly mention some of the most relevant literature for the Gaussian measurement model. Oymak et al. [101] analyzed the performance of the regularized PhaseLift algorithm and observed that the required sample complexity is of a suboptimal order compared to the optimal number of measurements required when phase information is available. For the special case of sparse phase retrieval, similar results have been reported in [85] which indicates that $O(k^2 \log(n))$ measurements are required for recovering of a k-sparse signal, using regularized PhaseLift. Recently, there has been a stream of work on solving phase retrieval using non-convex methods [20, 146]. In particular, Soltanolkotabi [119] has shown that amplitude-based Wirtinger flow can break the $O(k^2 \log(n))$ barrier. We also note that the paper [61] analyzed the PhaseMax algorithm with ℓ_1 regularizer and observed that it achieves perfect recovery with $O(k \log(n/k))$ samples, provided a well-correlated initialization point.

6.2 Preliminaries

Problem setup

Let $\mathbf{x}_0 \in \mathbb{R}^n$ denote the underlying *structured* signal. We consider the *real* phase retrieval problem with the goal of recovering $\mathbf{x}_0$ from m magnitude-only measurements of the form,

$$y_i = |\mathbf{a}_i^T \mathbf{x}_0|, \ i = 1, 2, \ldots, m \ , \tag{6.1}$$

where $\{\mathbf{a}_i \in \mathbb{R}^n\}_{i=1}^m$ is the set of (known) measurement vectors. In practice, this set is identified based on the experimental settings; however, throughout this chapter (for our analysis purposes), we assume that the $\mathbf{a}_i$'s are drawn independently from a Gaussian distribution with mean zero and covariance matrix $\mathbf{I}_n$. In order to exploit the structure of the signal, we assume that $f(\cdot)$ is a *convex* function that measures the "complexity" of the structured solution.

The regularized PhaseMax algorithm also relies on an initial estimate of the true signal. Here, $\mathbf{x}_{\text{init}}$ is used to represent this initial guess. Our analysis is based on the critical assumption that both $\mathbf{x}_{\text{init}}$ and $\mathbf{x}_0$ are **independent** of all the measurement vectors. The constraint set in generalized PhaseMax is derived by simply relaxing the equality constraints in (6.1) into *convex* inequality constraints. We introduce the following convex optimization problem to recover the signal:

$$\begin{aligned} \hat{\mathbf{x}} = \arg\min_{\mathbf{x} \in \mathbb{R}^n} \quad & L_\lambda(\mathbf{x}) = -{\mathbf{x}_{\text{init}}}^T \mathbf{x} + \lambda f(\mathbf{x}) \\ \text{subject to:} \quad & |\mathbf{a}_i^T \mathbf{x}| \leq y_i \ , \ \text{ for } \ 1 \leq i \leq m. \end{aligned} \tag{6.2}$$

The function f is assumed to be sign invariant, i.e., $f(\mathbf{x}) = f(-\mathbf{x})$ for all $\mathbf{x} \in \mathbf{R}^n$ ($-\mathbf{x}$ has the same "complexity" as $\mathbf{x}$.) Note that because of the global phase ambiguity of measurements in (6.1), we can only estimate $\mathbf{x}_0$ up to a sign. Up to this sign ambiguity, we can use the normalized mean squared error (NMSE), defined as $\frac{||\hat{\mathbf{x}}-\mathbf{x}_0||^2}{||\mathbf{x}_0||^2}$, to measure the performance of the solution. Here, we investigate the conditions under which the optimization program (6.2) uniquely identifies the true signal, i.e., $\hat{\mathbf{x}} = \mathbf{x}_0$ (up to the sign). Our results are asymptotic which is valid when $m, n \to \infty$.

Background on convex analysis

Our results give the required number of measurements as a function of certain geometrical properties of the descent cone of the objective function. Here, we recall these definitions from convex analysis.

Definition 3. *(Descent cone) For a function $R : \mathbb{R}^n \to \mathbb{R}$, the descent(tangent) cone at point $\mathbf{x}$ is defined as,*

$$T_R(\mathbf{x}) = cone(\{\mathbf{z} \in \mathbb{R}^n : R(\mathbf{x}+\mathbf{z}) \leq R(\mathbf{x})\}) \ , \tag{6.3}$$

where $cone(\mathcal{S})$ denotes the closed conical hull of the set $\mathcal{S}$.

Definition 4. *Let $\mathcal{S}$ be a closed convex set in $\mathbb{R}^n$. For $\mathbf{x} \in \mathbf{R}^n$, the projection of $\mathbf{x}$ on $\mathcal{S}$, denoted by $\Pi_{\mathcal{S}}(\mathbf{x})$, is defined as follows,*

$$\Pi_{\mathcal{S}}(\mathbf{x}) := \underset{\mathbf{y}\in\mathcal{S}}{\arg\min} ||\mathbf{x}-\mathbf{y}|| \ , \tag{6.4}$$

where $||\cdot||$ is the Euclidean norm. The distance function is defined as: $dist_{\mathcal{S}}(\mathbf{x}) = ||\mathbf{x} - \Pi_{\mathcal{S}}(\mathbf{x})||$.

Definition 5. *(Statistical dimension) [7] The statistical dimension of a closed convex cone C in $\mathbb{R}^n$ is defined as,*

$$d(C) = \mathbb{E}_{\mathbf{g}} \left[||\Pi_C(\mathbf{g})||^2\right] \ , \tag{6.5}$$

where $\mathbf{g} \in \mathbb{R}^n$ is a random vector with independent standard normal entries.

The statistical dimension canonically extends the dimension of linear spaces to convex cones. This quantity has been extensively studied in linear inverse problems. It is well-known that as $n \to \infty$, $m > d(T_{L_\lambda}(\mathbf{x}_0))$ is the necessary and sufficient condition for perfect signal recovery under noiseless linear Gaussian measurements [33, 122]. Our analysis indicates that given *phaseless* linear measurements, the regularized PhaseMax algorithm requires $O(d(T_{L_\lambda}(\mathbf{x}_0)))$ measurements for perfect signal reconstruction. Therefore, it is order-wise optimal in that sense.

6.3 Recovery Thresholds for Regularized PhaseMax

In this section, we present the main results of the chapter which provide us with the required number of measurements for perfect signal recovery in the regularized PhaseMax optimization (6.2). This gives the value $m_0 = m_0(n, \mathbf{x}_0, \mathbf{x}_{\text{init}}, \lambda)$, such that the regularized PhaseMax algorithm uniquely identifies the underlying signal $\mathbf{x}_0$ with high probability whenever $m > m_0$.

First, we investigate sufficient conditions for recovery of the underlying signal. Theorem 7 provides an upper bound on the number of measurements that is equal to a constant factor times the statistical dimension of the descent cone, $d(T_{L_\lambda}(\mathbf{x}_0))$. Therefore, although our analysis is not exact in this section, it leads us to the important observation that our proposed method is order-wise optimal in terms of the required sample complexity for perfect signal reconstruction.

In addition to the sufficient recovery condition, we provide an exact analysis for the phase transition behavior of regularized PhaseMax when the regularizer is an absolutely scalable function. We apply this result to the case of unstructured phaseless recovery as well as sparse phaseless recovery to compute the exact phase transitions. We then compare the result of our theory with the empirical results from numerical simulations.

Sufficient recovery condition

Let $\mathbf{P} := \frac{1}{||\mathbf{x}_0||^2}\mathbf{x}_0\mathbf{x}_0^T$ and $\mathbf{P}^\perp := \mathbf{I} - \mathbf{P}$ denote the projections onto the span of $\mathbf{x}_0$ and its orthogonal complement, respectively, where $||\cdot||$ denotes the ℓ_2-norm of the vectors. We also define $d^{(n)} := d(T_{L_\lambda}(\mathbf{x}_0))$ as the statistical dimension of the descent cone of the objective function at point $\mathbf{x}_0$. Our analysis rigorously characterizes the phase transition behavior of the regularized PhaseMax in the large system limit, i.e., when $n \to \infty$, while m and $d^{(n)}$ grow at a proportional ratio $\delta = \frac{m}{d^{(n)}}$. δ is often called the oversampling ratio. Here, the superscript (n) is used to denote the elements of a sequence. To streamline the notations, we often drop this when understood from the context.

Theorem 7 provides sufficient conditions for the successful recovery of $\mathbf{x}_0$. The recovery threshold depends on λ and the initialization vector, $\mathbf{x}_{\text{init}}$. We define $\rho_{\text{init}} := \mathbf{x}_{\text{init}}^T\mathbf{x}_0$ to quantify the caliber of the initial estimate. Due to the sign invariance property of the solution, we can assume without loss of generality that $\rho_{\text{init}} \geq 0$. Before stating the theorem, we shall introduce the function $R : (2, +\infty) \to \mathbb{R}_+$.

Definition 6. *For $x > 2$, $R(x)$ is the unique nonzero solution of the following equation:*

$$t^2 = \frac{x}{\pi}((1 + t^2)atan(t) - t) \ . \tag{6.6}$$

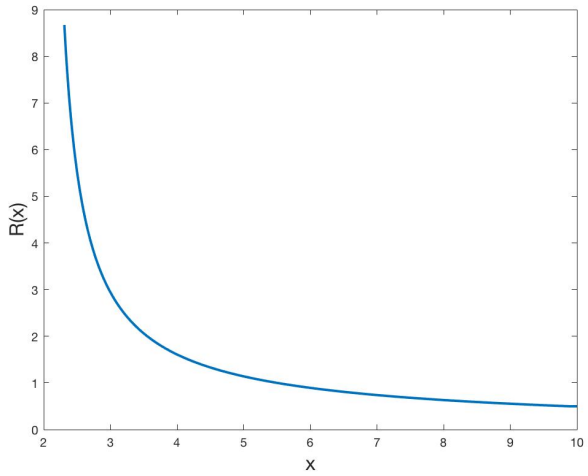


Figure 6.1: The function $R(x)$, which is defined in Definition 6, for different values of x. R is a monotonically decreasing function that approaches 0 in the limit.

Figure 6.1 depicts the evaluation of the function $R(x)$ for different input values x. As observed, $R(x)$ is a decreasing function with respect to x, and it approaches zero as x grows to infinity. It can be shown that for large values of the input x, $R(x)$ decays with the rate $\frac{1}{x}$.

Theorem 7 (Sufficient recovery condition)**.** *For a fixed oversampling ratio $\delta > 2$, the regularized PhaseMax optimization* (6.2) *perfectly recovers the target signal (in the sense that $\lim_{n\to\infty} \mathbb{P}\{||\hat{\mathbf{x}} - \mathbf{x}_0||^2 > \epsilon||\mathbf{x}_0||^2\} = 0$, for any fixed $\epsilon > 0$) if,*

$$R(\delta) < \sup_{\mathbf{v}\in\partial L_\lambda(\mathbf{x}_0)} \frac{||\mathbf{P}\mathbf{v}||}{||\mathbf{P}^\perp\mathbf{v}||}\,, \tag{6.7}$$

where $\partial L_\lambda(\mathbf{x}_0)$ denotes the sub-differential set of the objective function $L_\lambda(\cdot)$ at point $\mathbf{x}_0$.

It is worth noting that $\partial L_\lambda(\mathbf{x}_0)$ is a convex and compact set, and it can be expressed in terms of the sub-differential of the regularization function $\partial f(\mathbf{x}_0)$ as follows,

$$\partial L_\lambda(\mathbf{x}_0) = \{\lambda\mathbf{u} - \mathbf{x}_{\text{init}} : \mathbf{u} \in \partial f(\mathbf{x}_0)\}\,. \tag{6.8}$$

Observe that since $R(\cdot)$ is a monotonically decreasing function, the inequality (6.7) gives a lower bound for the oversampling ratio δ. In fact, we can restate the result in terms of this lower bound as the following corollary:

Corollary 3. *If there exists a fixed constant $\tau > 0$ such that,*

$$\sup_{\mathbf{v}\in\partial L_\lambda(\mathbf{x}_0)} \frac{||\mathbf{P}\mathbf{v}||}{||\mathbf{P}^\perp\mathbf{v}||} > \tau, \tag{6.9}$$

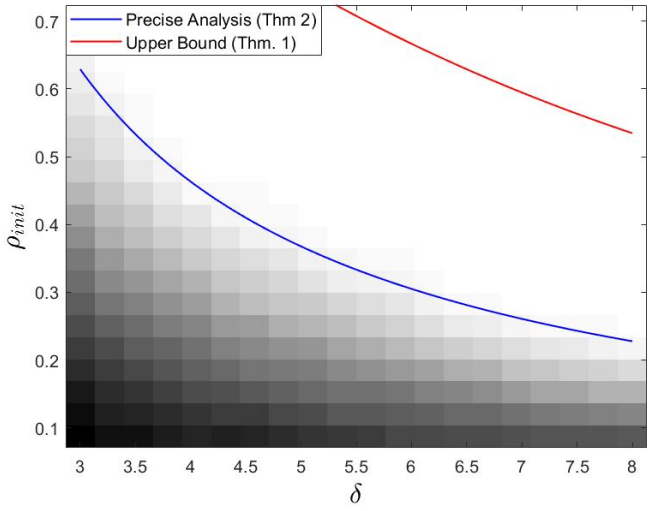


Figure 6.2: Phase transition regimes for the regularized PhaseMax problem in terms of the oversampling ratio δ and $\rho_{\text{init}} = \mathbf{x}_{\text{init}}^T \mathbf{x}_0$, for the cases of $\mathbf{x}_0$ with **no structure**. The blue line indicates the theoretical estimate for the phase transition derived from Theorem 8. The red line corresponds to the upper bound calculated by Theorem 7. In the simulations, we used signals of size $n = 128$. The result is averaged over 10 independent realizations of the measurement vectors.

then the regularized PhaseMax optimization (6.2) *has perfect recovery for $\delta > C$, where C is a constant that only depends on τ.*

Proof. This corollary is an immediate consequence of Theorem 7 by choosing $C = R^{-1}(\tau)$ and noting that $R(\cdot)$ is monotonically decreasing. □

This result indicates that if $\mathbf{x}_{\text{init}}$ and λ are chosen in such a way that the inequality (6.9) is satisfied for some positive constant τ, then one needs $m > Cd^{(n)}$ measurement samples for perfect recovery, where C is a constant and $d^{(n)}(= d)$ is the statistical dimension of the descent cone of the objective function at point $\mathbf{x}_0$. As motivating examples, we use Theorem 7 to find upper bounds on the phase transition when $\mathbf{x}_0$ has no structure or it is a sparse signal.

Example 1: Assume the target signal $\mathbf{x}_0$ has no a priori structure. The objective function in this case would be $L(\mathbf{x}) = -\mathbf{x}_{\text{init}}^T \mathbf{x}$, and $\partial L(\mathbf{x}_0) = \{-\mathbf{x}_{\text{init}}\}$. It can be shown that the statistical dimension is $d^{(n)} = n - 1/2$. Due to the absence of the regularization term in this case, without loss of generality, we can assume that $\|\mathbf{x}_0\| = \|\mathbf{x}_{\text{init}}\| = 1$. Theorem 7 provides the following sufficient condition for

perfect recovery:

$$\frac{||\mathbf{P}\mathbf{x}_{\text{init}}||}{||\mathbf{P}^{\perp}\mathbf{x}_{\text{init}}||} = \frac{\rho_{\text{init}}}{\sqrt{1-\rho_{\text{init}}^2}} > R(\delta)\ . \tag{6.10}$$

This indicates that $O(n)$ measurements are sufficient for perfect recovery as long as $\rho_{\text{init}} \geq \rho_0$, where $\rho_0 > 0$ is a constant that does not approach zero as $n \to \infty$. The exact phase transition for the unstructured case (PhaseMax), which was presented in Section 5.3, is compatible with this result. Figure 6.2 shows the result of numerical simulation for different values of δ and ρ_{init}, when $n = 128$. As depicted in the figure, the sufficient recovery condition from Theorem 7 is approximately a factor of 2 away from the actual phase transition.

Example 2: Let $\mathbf{x}_0$ be a k-sparse signal. In this case, we use $\|\cdot\|_1$ as the regularization function. We show in Section 6.8 that if $\lambda > \frac{c}{\sqrt{k}}$, then $d^{(n)} \leq Ck\log(n/k)$ for some constants $c, C > 0$. This matches the well-known order for the statistical dimension derived in the compressed sensing literature [122].

Moreover, in order to satisfy the condition in Corollary 3, we need to have $\frac{\rho_{\text{init}}}{||\mathbf{x}_0||_1} > (1+\epsilon)\lambda$, for some $\epsilon > 0$. Therefore, $\mathbf{x}_0$ can be perfectly recovered having $O(k\log(n/k))$ samples when the hyper-parameter λ is tuned properly, i.e., $\frac{c}{\sqrt{k}} < \lambda < \frac{\rho_{\text{init}}}{||\mathbf{x}_0||_1}$. Figure 6.3 compares this upper bound with the precise analysis that we will show in the next section. As depicted in this figure, the sufficient recovery condition is a valid upper bound on the phase transition, but it is not sharp.

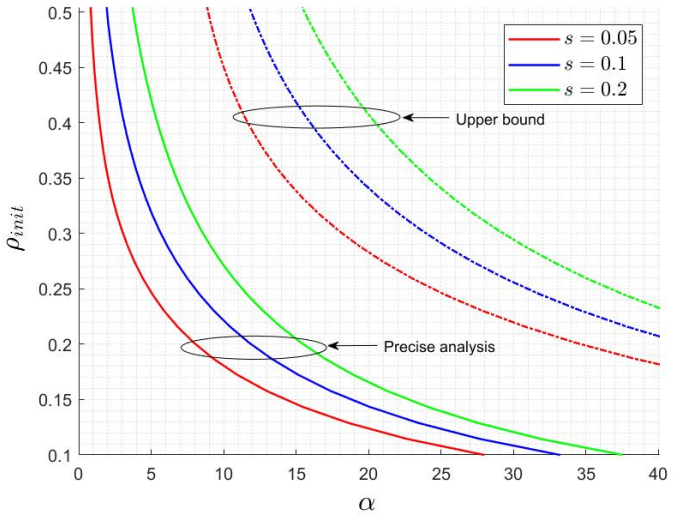


Figure 6.3: Comparing the upper bounds on the phase transition, derived by Theorem 7 (dashed lines) and the precise phase transition by Theorem 8 (solid lines), for three values of the sparsity factor $s = k/n$.

Precise phase transition

So far, we have provided a sufficient condition for perfect signal recovery in the regularized PhaseMax. In this section, we give the exact phase transition, i.e., the minimum number of measurements m_0 required for perfect recovery of the unknown vector $\mathbf{x}_0$. For our analysis, we assume that the function $f(\cdot)$ is absolutely homogeneous (scalable), i.e., $f(\tau \cdot \mathbf{x}) = |\tau| \cdot f(\mathbf{x})$, for any scalar τ, and every $\mathbf{x} \in \mathbb{R}^n$. This covers a large range of regularization functions such as norms and semi-norms. Let $\partial L_\lambda{}^\perp(\mathbf{x}_0) \subset \mathbb{R}^n$ denote the projection of the sub-differential set into the orthogonal complement of $\mathbf{x}_0$, i.e.,

$$\partial L_\lambda{}^\perp(\mathbf{x}_0) = \{\mathbf{P}^\perp \mathbf{u} : \mathbf{u} \in \partial L_\lambda(\mathbf{x}_0)\} \ , \tag{6.11}$$

which is a convex and compact set. To state the result in a general framework, we require one further assumption on functions $L_\lambda^{(n)}(\cdot)$.

Assumption 4 (Asymptotic functionals)**.** *The following uniform convergences exist, as $n \to \infty$,*

$$\begin{aligned} \beta \ - \ \mathbb{E}\big[\frac{1}{\sqrt{n}}\, \mathbf{h}^T \, \Pi_{\partial L_\lambda{}^\perp(\mathbf{x}_0)}(\frac{\beta}{\sqrt{n}}\mathbf{h})\big] &\xrightarrow{\textit{Unif.}} F_\lambda(\beta), \ \textit{and,} \\ \mathbb{E}\big[dist_{\partial L_\lambda{}^\perp(\mathbf{x}_0)}(\frac{\beta}{\sqrt{n}}\mathbf{h})\big] &\xrightarrow{\textit{Unif.}} G_\lambda(\beta) \ , \end{aligned} \tag{6.12}$$

where $\mathbf{h} \in \mathbb{R}^n$ has i.i.d. standard normal entries and $F_\lambda, \ G_\lambda : \mathbb{R}_+ \to \mathbb{R}$ denote the functions that the sequences uniformly converge to.

One can show that, under some mild conditions on the regularization function $f(\cdot)$, Assumption 4 holds and also $F_\lambda(\beta) = G_\lambda(\beta) G'_\lambda(\beta)$, where $G'_\lambda(\cdot)$ denotes the derivative of the function $G_\lambda(\cdot)$. This assumption especially holds for the class of separable regularizers, where $f(\mathbf{v}) = \sum_i \tilde{f}(v_i)$ (e.g. ℓ_1 norm for the case of sparse phase-retrieval). Later in this section, we will see validity of this assumption for the two examples discussed earlier.

Our precise phase transition results indicate the required number of measurements as the solution of a set of two nonlinear equations with two unknowns. We define a new parameter $\alpha := \frac{m}{n}$, where $\alpha_{\text{opt}} = \frac{m_0}{n}$ indicates the exact phase transition of the regularized PhaseMax optimization. The following theorem gives an implicit formula to derive α_{opt}.

Theorem 8 (Precise phase transition)**.** *Let $\hat{\mathbf{x}}$ be the solution to the regularized PhaseMax optimization* (6.2) *with the objective function $L_\lambda(\mathbf{x}) = -\mathbf{x}_{init}^T \mathbf{x} + \lambda f(\mathbf{x})$, where the convex function $f(\cdot)$ is absolutely homogeneous and Assumption 4 holds. The regularized PhaseMax optimization would perfectly recover the target signal $\mathbf{x}_0$ if and only if:*

1. $\alpha > \alpha_{opt}$, *where* α_{opt} *is the solution of the following system of nonlinear equations with two unknowns,* α *and* β*,*

$$\begin{cases} -G_\lambda(\beta)\ L_\lambda(\mathbf{x}_0) = \tan(\frac{\pi}{\alpha\beta}F_\lambda(\beta))\ (G_\lambda^2(\beta) - \beta F_\lambda(\beta))\ , \\ \tan(\frac{\pi}{\alpha\beta}F_\lambda(\beta))\ (G_\lambda(\beta) + \frac{\pi}{\alpha\beta}F_\lambda(\beta)\ L_\lambda(\mathbf{x}_0)) = \frac{\pi}{\alpha\beta}F_\lambda(\beta)\ G_\lambda(\beta) \end{cases} \tag{6.13}$$

2. *and,* $L_\lambda(\mathbf{x}_0) < L_\lambda(0) = 0$

where the functions $F_\lambda(\cdot)$ *and* $G_\lambda(\cdot)$ *are defined in* (6.12)*.*

A few remarks are in place for this theorem:

Remark 5 (Solving equations (6.13))**.** *The system of nonlinear equations* (6.13) *only involves two scalars* β *and* α*, and the functions* $F_\lambda(\beta)$ *and* $G_\lambda(\beta)$ *are determined by the objective function* $L_\lambda(\mathbf{x})$*. For our numerical simulations, we used a fixed-point iterative method that can quickly find the solution given a proper initialization.*

Remark 6 (Tuning λ)**.** *Theorem 8 requires the objective function to satisfy* $L_\lambda(\mathbf{x}_0) = \lambda f(\mathbf{x}_0) - \rho_{init} < 0$*. Therefore, it is necessary to choose* λ *in such a way that* $\lambda < \frac{\rho_{init}}{f(\mathbf{x}_0)}$*. Some additional assumptions on the unknown vector* $\mathbf{x}_0$ *enables us to calculate the proper range for* λ*. For instance, if we consider the case where the entries of* $\mathbf{x}_0$ *are drawn form a specific distribution, where the non-zero entries of* $\mathbf{x}_0$ *are Gaussian (or other random variables),* $\mathbb{E}[f(\mathbf{x}_0)]$ *gives a reasonable estimation on* $f(\mathbf{x}_0)$ *that can help us in choosing* λ *appropriately. We will see an example of such case in the next section. Figure 6.4 shows an example of how the phase transition of the regularized PhaseMax, or equivalently the required sample complexity, behaves as a function of the hyper-parameter* λ*.*

In the next section, we use the result of Theorem 8 to compute the exact phase transition for the case of unstructured signal as well as the sparse signal recovery. Since the regularizer $f(\mathbf{x})$ is absolutely scalable, for both examples, we assume that $\|\mathbf{x}_0\| = 1$.

6.4 Applications of Theorem 8

Unstructured signal recovery

When there is no a priori information about the structure of the target signal, we use the following optimization (PhaseMax) for signal recovery:

$$\begin{aligned} \hat{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathbf{R}^n} \quad & L(\mathbf{x}) = -{\mathbf{x}_{\text{init}}}^T\mathbf{x} \\ \text{subject to:} \quad & |\mathbf{a}_i^T\mathbf{x}| \le y_i\ , \quad \text{for}\ \ 1 \le i \le m\ . \end{aligned} \tag{6.14}$$

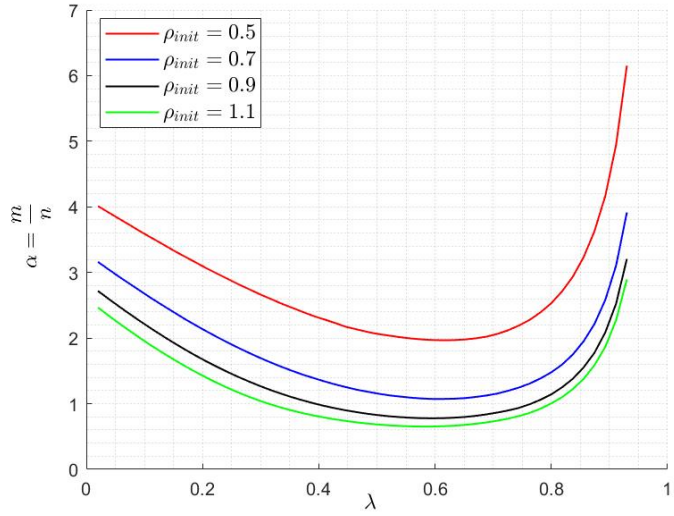


Figure 6.4: The phase transition behavior as a function of the regularization parameter λ, derived from the result of Theorem 8. As depicted in the figure, there is a suitable region for tuning λ which gives a lower recovery threshold for the regularized PhaseMax.

Due to the absence of the regularization term, without loss of generality we can assume that $\|\mathbf{x}_{\text{init}}\| = 1$. Moreover, $L(\mathbf{x}_0) = -\rho_{\text{init}}$ which indicates that the second condition in Theorem 8 is satisfied. To apply the result of our theorem, we first compute explicit formulas for the functions $F_\lambda(\beta)$ and $G_\lambda(\beta)$ as follows,

$$F_\lambda(\beta) = \beta\,, \quad G_\lambda(\beta) = \sqrt{\beta^2 + 1 - \rho_{\text{init}}^2}\,. \tag{6.15}$$

We can now form the system of nonlinear equations (6.13) as follows,

$$\begin{cases} \sqrt{\beta^2 + 1 - \rho_{\text{init}}^2}\ \frac{\rho_{\text{init}}}{1-\rho_{\text{init}}^2} = \tan(\frac{\pi}{\alpha})\,, \\ \tan(\frac{\pi}{\alpha})\ (\sqrt{\beta^2 + 1 - \rho_{\text{init}}^2} - \frac{\pi\rho_{\text{init}}}{\alpha}) = \frac{\pi}{\alpha}\sqrt{\beta^2 + 1 - \rho_{\text{init}}^2}\,. \end{cases} \tag{6.16}$$

Finally, solving equations (6.16) yields the following necessary and sufficient condition for perfect recovery,

$$\frac{\pi}{\alpha \tan(\pi/\alpha)} > 1 - \rho_{\text{init}}^2\,, \tag{6.17}$$

which also verifies the result presented in Section 5.3.

Figure 6.2 shows the result of numerical simulations of running the PhaseMax algorithm for different values of ρ_{init} and δ. The intensity level of the color of each square in Figure 6.2 represents the

error of PhaseMax in recovering $\mathbf{x}_0$. As seen in the figure, although our theoretical results have been established for the asymptotic setting (when the problem dimensions approach infinity), the blue line, which is derived from (6.17), reasonably predicts the phase transition for $n = 128$. The sufficient condition that is derived from Theorem 7 is also depicted by the red line in the same figure.

Sparse recovery

We consider the case where the target signal $\mathbf{x}_0$ is sparse with k non-zero entries. The convex function $f(\mathbf{x}) = \frac{1}{\sqrt{n}}\|\mathbf{x}\|_1$, which is known to be a proper regularizer that enforces sparsity [132], is used in the regularized PhaseMax optimization to recover $\mathbf{x}_0$,

$$\begin{aligned} \hat{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \quad & L_\lambda(\mathbf{x}) = -{\mathbf{x}_{\text{init}}}^T\mathbf{x} + \frac{\lambda}{\sqrt{n}}\|\mathbf{x}\|_1 \\ \text{subject to:} \quad & |\mathbf{a}_i^T\mathbf{x}| \le y_i \ , \quad \text{for } \ 1 \le i \le m \ . \end{aligned} \tag{6.18}$$

To streamline notations, here we assume that the non-zero entries of $\mathbf{x}_0$ are the first k entries and decompose vector $\mathbf{v} \in \mathbb{R}^n$ as $\mathbf{v} = \begin{bmatrix} \mathbf{v}^\Delta \\ \mathbf{v}^{\Delta^c} \end{bmatrix}$, where $\mathbf{v}^\Delta \in \mathbb{R}^k$ denotes the first k entries of $\mathbf{v}$, and $\mathbf{v}^{\Delta^c} \in \mathbb{R}^{n-k}$ is the remaining $n-k$ entries. As $m, n \to \infty$, we would like to apply the result of Theorem 8 to compute the exact phase transition. Due to the rotational invariance property of the Gaussian distribution, it can be shown that multiplying the last $(n-k)$ entries of $\mathbf{x}_{\text{init}}$ by a unitary matrix $\mathbf{U} \in \mathbb{R}^{(n-k)\times(n-k)}$ does not change the phase transition behavior in (6.2). Hence, we can assume that the entries of $\mathbf{x}_{\text{init}}^{\Delta^c}$ have a Gaussian distribution, i.e.,

$$\mathbf{x}_{\text{init}} = \begin{bmatrix} \mathbf{x}_{\text{init}}^\Delta \\ \mathbf{x}_{\text{init}}^{\Delta^c} \end{bmatrix}, \quad \text{and} \quad \mathbf{x}_{\text{init}}^{\Delta^c} = \frac{1}{\sqrt{n-k}} \ \|\mathbf{x}_{\text{init}}^{\Delta^c}\| \ \mathbf{g} \ , \tag{6.19}$$

where $\mathbf{g} \in \mathbb{R}^{n-k}$ has standard normal entries. This observation enables us to establish the following lemma:

Lemma 9. *Consider the optimization problem* (6.18) *to recover the k-sparse signal $\mathbf{x}_0$. We assume that the entries of $\mathbf{x}_{init}$ are distributed as in* (6.19) *and define $\tilde{\rho} := \frac{1}{\sqrt{k}} sign(\mathbf{x}_0^\Delta)^T \mathbf{x}_{init}^\Delta$, where $sign(\cdot)$*

denotes the component-wise sign function. Then, Assumption 4 holds with:

$$F_\lambda(\beta) = \beta(s + 2(1-s)\cdot Q(\frac{\lambda}{\sqrt{\beta^2 + \frac{\|\mathbf{x}_{init}^{\Delta^c}\|^2}{1-s}}})\ ),$$
$$G_\lambda^2(\beta) = s\cdot(\beta^2+\lambda^2) + \|\mathbf{x}_{init}^{\Delta}\|^2 - 2\lambda\sqrt{s}\tilde{\rho} - L^2(\mathbf{x}_0)$$
$$+ (1-s)(\beta^2 + \frac{\|\mathbf{x}_{init}^{\Delta^c}\|^2}{1-s})\cdot \mathbb{E}_H[\ shrink^2(H, \frac{\lambda}{\sqrt{\beta^2 + \frac{\|\mathbf{x}_{init}^{\Delta^c}\|^2}{1-s}}})\ ] \tag{6.20}$$

where $Q(\cdot)$ is the tail distribution of the standard normal distribution, H has standard normal distribution, and $s := k/n$ is the sparsity factor. The shrinkage function $shrink(\cdot,\cdot) : \mathbb{R}\times\mathbb{R}_+ \to \mathbb{R}_+$ is defined as:

$$shrink(x,\tau) = (|x|-\tau)\mathbf{1}\{|x|\geq\tau\}\ . \tag{6.21}$$

It is worth noting that the function $\text{shrink}(\cdot,\cdot)$ also appeared in computing the statistical dimension for ℓ_1 regularization (see Section 6.8) which indicates some implicit relation to α_{opt}.

We have numerically computed the solution of the nonlinear system (6.20). Figure 6.5 shows the error of regularized PhaseMax over a range of ρ_{init} and δ. The comparison between our upper bound derived from Theorem 7 and precise analysis of Theorem 7 is depicted in Figure 6.3 for three values of the sparsity factor $s = 0.05, 0.1, 0.2$. Observe that the upper bound is only a constant factor away from the precise phase transition, while its derivation involves simpler formulas. Finally, Figure 6.4 illustrates impact of the regularization parameter λ on the phase transition of the regularized PhaseMax optimization for four values of ρ_{init}. The values of λ in this figure are normalized by $\frac{\rho_{\text{init}}\sqrt{n}}{\|\mathbf{x}_0\|}$, which is the maximum acceptable value of λ in the regularized PhaseMax.

6.5 Conclusion and Future Directions

In this chapter, we introduced a new convex optimization framework, *regularized PhaseMax*, to solve the structured phase retrieval problem. We have shown that, given a proper initialization, the regularized PhaseMax optimization perfectly recovers the underlying signal from a number of phaseless measurements that is only a constant factor away from the number of measurements required when the phase information is available. We explicitly computed this constant factor.

An important (yet still open) research problem is to investigate the required sample complexity to construct a proper initialization vector, $\mathbf{x}_{\text{init}}$. As an example, for the case of sparse phase retrieval, although our analysis indicates that $O(k\log\frac{n}{k})$ is the required sample complexity of the regularized PhaseMax optimization, the best known initialization technique [20] needs $O(k^2\log n)$ samples to

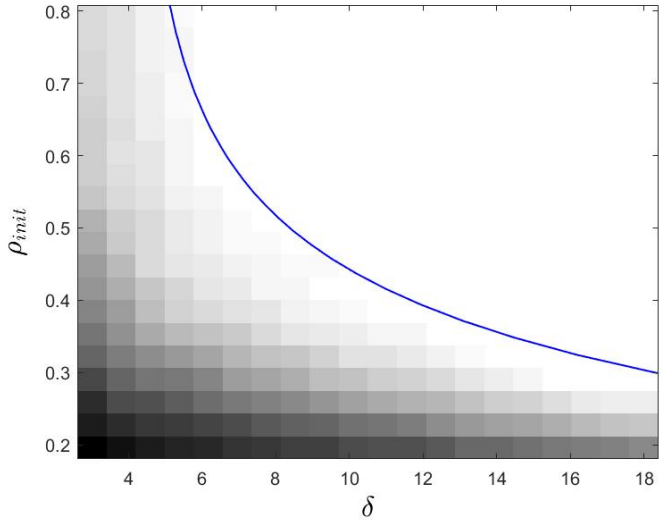


Figure 6.5: Phase transition regimes for the regularized PhaseMax problem in terms of the oversampling ratio δ and $\rho_{\text{init}} = \mathbf{x}_{\text{init}}^T \mathbf{x}_0$, for the cases of $\mathbf{x}_0$ with **sparse structure**. The blue line indicates the theoretical estimate for the phase transition derived from Theorem 8. In the simulations, we used signals of size $n = 128$. The result is averaged over 10 independent realizations of the measurements.

generate a meaningful initialization, which is suboptimal. An important future direction is to study initialization techniques that break this sample complexity barrier, or to exploit information theoretic arguments (as in [94]) to show that the sample complexity for the initialization cannot be improved.

To form the objective function in the regularized PhaseMax, we exploited some a priori knowledge about the structure of the underlying signal. In many practical settings, such prior information is not available. There has been some interesting recent publications (e.g. [10, 148, 51]) which introduce efficient algorithms to learn the structure of the underlying signal. An interesting research direction is to investigate new optimization frameworks that do not rely on the prior information about the structure of the underlying signal.

6.6 Proofs and Technical Derivations

In order to establish the results, we use the following lemma which provides an equivalent optimization that has the same error performance as PhaseMax, and is the key ingredient in deriving the main results of the paper.

Lemma 10 (Equivalent Optimization)**.** *Consider the regularized PhaseMax problem introduced in Section 6.2. As $n \to \infty$, the error performance converges in probability as follows:*

$$\left(\frac{||\hat{\mathbf{x}}-\mathbf{x}_0||^2}{||\mathbf{x}_0||^2}\right)-\left(||\mathbf{w}^\star||^2+(1-s^\star)^2\right)\overset{n\to\infty}{\longrightarrow}0\ . \tag{6.22}$$

Here $s^\star\in\mathbb{R}$ and $\mathbf{w}^\star\in\mathbb{R}^n$ are the unique optimizers of the following optimization program,

$$\begin{aligned}\min_{s\in\mathbb{R}}\ \min_{\mathbf{w}\in\mathbb{R}^n,\mathbf{w}\perp\mathbf{x}_0}\quad & -\mathbf{x}_{init}^T(s\mathbf{x}_0+\mathbf{w})+\lambda f(s\mathbf{x}_0+\mathbf{w})\\ \textit{subject to:}\quad & \mathbf{h}^T\mathbf{w}\geq\sqrt{m\ \mathbf{c_d}(s,||\mathbf{w}||)}\quad ,\end{aligned} \tag{6.23}$$

where $\mathbf{h}\in\mathbb{R}^n$ has i.i.d. standard normal entries and the function $\mathbf{c_d}:\mathbb{R}\times\mathbb{R}_+\to\mathbb{R}$ is defined as,

$$\mathbf{c_d}(s,r)=\frac{1}{\pi}[((1+s)^2+r^2)\ atan(\frac{r}{1+s})\ +((1-s)^2+r^2)\ atan(\frac{r}{1-s})-2r]. \tag{6.24}$$

The full technical details of obtaining this result is explained in Section 6.7. In short, to show the equivalence, we start from (6.2) and define new variables $s:=\mathbf{x}_0^T\mathbf{x}$ and $\mathbf{w}=\mathbf{P}^\perp\mathbf{x}$. Then reformulate it as an unconstrained optimization using Lagrange multipliers. The result is a consequence of applying CGMT (Lemma 32, see Appendix A.1) with some simplifications.

Before explaining the technical details of the proofs of our main results, we state one more lemma which will be used in the proof of Theorem 8. In the path of analyzing the auxiliary optimization, we replace several functions with their limits in probability. This can be done through the same tricks used in section A.4 of [129] and Lemma B.1 in the same paper. Here, we state the following lemma without proof.

Lemma 11 (Min-convergence – Open Sets)**.** *Consider a sequence of proper, convex stochastic functions $M_n:(0,\infty)\to\mathbb{R}$ and a deterministic function $M:(0,\infty)\to\mathbb{R}$ such that:*

1. *$M_n(x)\xrightarrow{P}M(x)$, for all $x>0$,*
2. *there exists $z>0$ such that $M(x)>\inf_{x>0}M(x)$ for all $x\geq z$.*

Then, $\inf_{x>0}M_n(x)\xrightarrow{P}\inf_{x>0}M(x)$.

The objective function in our optimization problems satisfies the assumptions of this lemma at the points where we replace them with their limits.

Proof of Theorem 7

Consider the following optimization:

$$\begin{aligned} \min_{s\in\mathbb{R}} \min_{\mathbf{w}\in\mathbb{R}^n, \mathbf{w}\perp\mathbf{x}_0} \quad & -\mathbf{x}_{\text{init}}^T(s\mathbf{x}_0+\mathbf{w}) + \lambda f(s\mathbf{x}_0+\mathbf{w}) \\ \text{subject to:} \quad & \mathbf{h}^T\mathbf{w} \geq \sqrt{m\ \mathbf{c_d}(s, \|\mathbf{w}\|)}\ . \end{aligned} \tag{6.25}$$

The result of Lemma 10 established that as $n \to \infty$, the error performance of the regularized PhaseMax converges to the error performance in (6.25). The following corollary indicates the necessary and sufficient condition for perfect recovery:

Corollary 4. *As $n \to \infty$, $\mathbf{x}_0$ is the unique solution of the regularized PhaseMax optimization, if and only if $(s^\star, \mathbf{w}^\star) = (1, \mathbf{0})$ be the unique optimizer of the equivalent optimization* (6.25).

Proof. This is an immediate consequence of Lemma 10, noticing that $\|\hat{\mathbf{x}} - \mathbf{x}_0\| = 0$ is the condition for perfect recovery. □

We proceed onwards with analyzing (6.25). For simplicity, we assume $\|\mathbf{x}_0\| = 1$. Define a new function $\hat{f}(\cdot) : \mathbb{R}^n \to \mathbb{R}$ as follows,

$$\hat{f}(\mathbf{x}) = f(\mathbf{x}_0) + \max_{\mathbf{v}\in\partial f(\mathbf{x}_0)} \mathbf{v}^T(\mathbf{x}-\mathbf{x}_0)\ . \tag{6.26}$$

$\partial f(\mathbf{x}_0)$ is the sub-differential set of function $f(\cdot)$ at point $\mathbf{x}_0$ which is a convex and compact set. $\hat{f}(\cdot)$ is basically the first-order approximation of the regularization function $f(\cdot)$ at point $\mathbf{x}_0$. This replacement cannot be done in general, but since we are only investigating the phase transition regime where the norm of the error, $\|\hat{\mathbf{x}} - \mathbf{x}_0\|$, approaches to zero, we may perform this exchange. To investigate the phase transition behavior in (6.25), we bound $1 - s$ and $||\mathbf{w}||$ to a small neighborhood of 0. Therefore, it is valid to replace f with $\hat{f}$ in that small neighborhood around $\mathbf{x}_0$. Reformulating the optimization using this replacement would give us the following,

$$\begin{aligned} \min_{s\in\mathbb{R}, \mathbf{w}\perp\mathbf{x}_0} \max_{\mathbf{v}\in\partial f(\mathbf{x}_0)} \quad & -\mathbf{x}_{\text{init}}^T(s\mathbf{x}_0+\mathbf{w}) + \lambda f(\mathbf{x}_0) + \lambda\mathbf{v}^T((s-1)\mathbf{x}_0+\mathbf{w}) \\ \text{subject to:} \quad & \mathbf{h}^T\mathbf{w} \geq \sqrt{m\ \mathbf{c_d}(s, ||\mathbf{w}||)}\ . \end{aligned} \tag{6.27}$$

We add the constant term, $\mathbf{x}_{\text{init}}^T\mathbf{x}_0 - \lambda f(\mathbf{x}_0)$, to the objective function and reformulate the maximization in terms of $\partial L(\mathbf{x}_0)$ as follows,

$$\begin{aligned} \min_{s\in\mathbb{R},\mathbf{w}\perp\mathbf{x}_0} \max_{\mathbf{v}\in\partial L(\mathbf{x}_0)} \quad & (s-1)\mathbf{x}_0^T\mathbf{v} + \mathbf{w}^T\mathbf{v} \\ \text{subject to:} \quad & \mathbf{h}^T\mathbf{w} \geq \sqrt{m\ \mathbf{c_d}(s, \|\mathbf{w}\|)}\ . \end{aligned} \tag{6.28}$$

If $|s| > 1$ in (6.28), we have the following inequalities:

$$||\mathbf{w}||^2\ d(T_L(\mathbf{x}_0)) \geq (\mathbf{h}^T\mathbf{w})^2 \geq m\ \mathbf{c_d}(s, ||\mathbf{w}||) > \frac{m}{2}||\mathbf{w}||^2\ . \tag{6.29}$$

The first inequality is due to the fact that $\mathbf{x} - \mathbf{x}_0 = (s-1)\mathbf{x}_0 + \mathbf{w}$ is in $T_L(\mathbf{x}_0)$ (the descent cone of the objective at point $\mathbf{x}_0$). The second inequality appeared as a constraint in the optimization problem (6.28). The last inequality holds since $\mathbf{c_d}(s, r) > r^2/2$, when $|s| \geq 1$. Therefore, using the assumption $\delta = \frac{m}{d} > 2$, it can be shown that the feasible set of (6.28) is nonempty if and only if $|s| \leq 1$.

Since the regularized PhaseMax optimization is convex, in order to show that $s^\star = 1$ and $\mathbf{w}^\star = \mathbf{0}$ are the unique optimizers of (6.28), it is sufficient to check the optimality condition in a small neighborhood of $(s^\star = 1, \mathbf{w}^\star = \mathbf{0})$. We also use the following approximation of the function $\mathbf{c_d}(s, r)$ which is valid in a small neighborhood around the point $(s, r) = (1, 0)$:

$$\mathbf{c_d}(s, r) = \frac{1}{\pi}\Big[\big((1-s)^2 + r^2\big)\text{atan}(\frac{r}{1-s}) - r(1-s)\Big]\ . \tag{6.30}$$

Next, for fixed $|s| < 1$, we will find an upper bound for $r := ||\mathbf{w}||$ such that s and $\mathbf{w}$ satisfy the constraint in (6.28). To this goal, we use the following inequalities:

$$r^2\ d(T_L(\mathbf{x}_0)) \geq (\mathbf{h}^T\mathbf{w})^2 \geq m\ \mathbf{c_d}(s, r) \Rightarrow r^2 \geq \delta\ \mathbf{c_d}(s, r)\ . \tag{6.31}$$

Replacing the approximation (6.30) for $\mathbf{c_d}(s, r)$ when $s \uparrow 1$, we have,

$$r \leq R(\delta)\big(1 - s\big)\ , \tag{6.32}$$

where $R(\delta)$ is the unique nonzero solution of the following nonlinear equation:

$$t^2 = \frac{\delta}{\pi}\Big[(1 + t^2)\text{atan}(t) - t\Big]\ . \tag{6.33}$$

We are now at the stage to establish the result of Theorem 7. Assume that $\tilde{\mathbf{v}} \in \partial L(\mathbf{x}_0)$ achieves the supremum in (6.7) (note that $\tilde{\mathbf{v}}$ always exists because the set $\partial L(\mathbf{x}_0)$ is compact). $\tilde{\mathbf{v}}$ then satisfies the following conditions:

1. $\mathbf{x}_0^T \tilde{\mathbf{v}} < 0$,
2. $||\mathbf{P}\tilde{\mathbf{v}}|| > R(\delta)\ ||\mathbf{P}^{\perp}\tilde{\mathbf{v}}||$.

We have the following inequalities:

$$\begin{aligned} \min_{|s|\leq 1, \mathbf{w}\perp\mathbf{x}_0} \max_{\mathbf{v}\in\partial L(\mathbf{x}_0)} \ (s-1)\mathbf{x}_0^T\mathbf{v} + \mathbf{w}^T\mathbf{v} \quad &\geq \min_{|s|\leq 1, \mathbf{w}\perp\mathbf{x}_0} (s-1)\mathbf{x}_0^T\tilde{\mathbf{v}} + \mathbf{w}^T\tilde{\mathbf{v}} \\ &\geq \min_{|s|\leq 1, \mathbf{w}\perp\mathbf{x}_0} (1-s)||\mathbf{P}\tilde{\mathbf{v}}|| - ||\mathbf{w}||\ ||\mathbf{P}^{\perp}\tilde{\mathbf{v}}|| \ , \end{aligned} \tag{6.34}$$

where for the first inequality, we used the fact that maximization over $\mathbf{v}$ gives a larger value compared to choosing the specific vector $\tilde{\mathbf{v}}$. For the second inequality we used Cauchy-Schwarz to bound $\mathbf{w}^T\tilde{\mathbf{v}}$ from below. When $s \uparrow 1$, we use the approximation (6.32) which bounds $\|\mathbf{w}\|$ from above. Therefore, we have:

$$(1-s)||\mathbf{P}\tilde{\mathbf{v}}|| - ||\mathbf{w}|| \cdot ||\mathbf{P}^{\perp}\tilde{\mathbf{v}}|| > (1-s)(||\mathbf{P}\tilde{\mathbf{v}}|| - R(\delta)||\mathbf{P}^{\perp}\tilde{\mathbf{v}}||) > 0. \tag{6.35}$$

This gives the final result that $s^{\star} = 1$, $\mathbf{w}^{\star} = \mathbf{0}$ is the unique solution of (6.28). The perfect recovery in the generalized PhaseMax follows from the result of Corollary 4.

Proof of Theorem 8

We start from the equivalent optimization derived as the result of Lemma 10, defined as,

$$\begin{aligned} \min_{s\in\mathbb{R}} \min_{\mathbf{w}\in\mathbb{R}^n, \mathbf{w}\perp\mathbf{x}_0} \quad & -\mathbf{x}_{\text{init}}^T(s\mathbf{x}_0 + \mathbf{w}) + \lambda f(s\mathbf{x}_0 + \mathbf{w}) \\ \text{subject to:} \quad & \mathbf{h}^T\mathbf{w} \geq \sqrt{m\ \mathbf{c}_{\mathbf{d}}(s, \|\mathbf{w}\|)} \ . \end{aligned} \tag{6.36}$$

One key idea to analyze this optimization is to replace $f(s\mathbf{x}_0 + \mathbf{w})$ with its first-order linear approximation around the point $\mathbf{x}_0$. Let $\hat{f}$ denote the approximation function,

$$\hat{f}(\mathbf{x}) = f(\mathbf{x}_0) + \max_{\mathbf{v}\in\lambda\partial f(\mathbf{x}_0)} \mathbf{v}^T(\mathbf{x} - \mathbf{x}_0) \ . \tag{6.37}$$

Here, $\partial f(\mathbf{x}_0)$ denotes the sub-differential of $f(\cdot)$ at point $\mathbf{x}_0$ which is well-defined for convex functions and is a compact and convex set. Replacing $f(\cdot)$ with its approximation enables us to precisely analyze the conditions for perfect signal recovery in the equivalent optimization (6.36) which determines the precise phase transition in the regularized PhaseMax optimization. This approximation is tight when the norm of the error approaches zero (which occurs in perfect recovery). We refer the interested reader to [99] for more details.

Therefore, for the rest of this section, we will analyze the following optimization,

$$\begin{aligned} \min_{\substack{s\in\mathbb{R}\\ \mathbf{w}\in\mathbb{R}^n,\mathbf{w}\perp\mathbf{x}_0}} \max_{\mathbf{v}\in\lambda\partial f(\mathbf{x}_0)} \quad & f(\mathbf{x}_0)+\lambda\mathbf{v}^T(s\mathbf{x}_0+\mathbf{w}-\mathbf{x}_0)-s\rho_{\text{init}}-\mathbf{x}_{\text{init}}^T\mathbf{w} \\ \text{subject to:} \quad & \mathbf{h}^T\mathbf{w}\geq\sqrt{m\ \mathbf{c_d}(s,\|\mathbf{w}\|)}\ . \end{aligned} \tag{6.38}$$

Next, we use the dual variable β to rewrite (6.38) as,

$$\min_{\substack{s\in\mathbb{R}\\ \mathbf{w}\in\mathbb{R}^n,\mathbf{w}\perp\mathbf{x}_0}} \max_{\substack{\mathbf{v}\in\lambda\partial f(\mathbf{x}_0)\\ \beta\geq 0}} f(\mathbf{x}_0)+\lambda\mathbf{v}^T(s\mathbf{x}_0+\mathbf{w}-\mathbf{x}_0)-s\rho_{\text{init}}-\mathbf{x}_{\text{init}}^T\mathbf{w}-\frac{\beta}{\sqrt{n}}\mathbf{h}^T\mathbf{w}+\frac{\beta}{\sqrt{n}}\sqrt{m\ \mathbf{c_d}(s,||\mathbf{w}||)}\ . \tag{6.39}$$

In the next step, we would like to switch the minimization over $\mathbf{w}$ with the maximization over β. But since the objective function is not convex with respect to $\mathbf{w}$, such an exchange would not be a direct result of the Sion's min-max theorem. However, note that the initial optimization satisfies the conditions of the Sion's min-max theorem. In the asymptotic settings, using the same techniques as in [129] (see section A.2.4 in the appendix of the paper), one can show that changing the order of min and max does not change the solution of the optimization problem.

Hence, we are now able to first do the minimization over $\mathbf{w}$. To do this, we define $r := \|\mathbf{w}\|$ and by fixing r, we are computing the minimization with respect to the direction of $\mathbf{w}$. The following optimization is the result of minimization over the direction of $\mathbf{w}$:

$$\min_{\substack{s\in\mathbb{R}\\ r\geq 0}} \max_{\substack{\mathbf{v}\in\lambda\partial f(\mathbf{x}_0)\\ \beta\geq 0}} (1-s)(\rho_{\text{init}}-\mathbf{v}^T\mathbf{x}_0)-r\cdot\|\mathbf{P}^{\perp}(\lambda\mathbf{v}-\mathbf{x}_{\text{init}}-\frac{\beta}{\sqrt{n}}\mathbf{h})\|+\beta\sqrt{\alpha\ \mathbf{c_d}(s,r)}\ , \tag{6.40}$$

where, as defined in Section 6.3, $\alpha=\frac{m}{n}$ is the oversampling ratio and $\mathbf{P}^{\perp}=\mathbf{I}-\mathbf{x}_0\mathbf{x}_0^T$ is the projection to the orthogonal subspace of $\mathbf{x}_0$.

Up to this point, the result is valid for every convex function $f(\cdot)$. But in order to continue our analysis, we need the following lemma which restricts us to a specific class of functions, i.e., the class of absolutely scalable functions.

Lemma 12. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function such that for all $\mathbf{x} \in \mathbb{R}^n$ and $\alpha \geq 0$, $f(\alpha\, \mathbf{x}) = \alpha\, f(\mathbf{x})$. Then, for all $\mathbf{v} \in \partial f(\mathbf{x})$,*

$$\mathbf{v}^T \mathbf{x} = f(\mathbf{x}) \ , \tag{6.41}$$

where $\partial f(\mathbf{x})$ is the set of sub-differentials of function $f(\cdot)$ at point $\mathbf{x}$.

Proof. Since $f(\cdot)$ is convex, for all $\mathbf{v} \in \partial f(\mathbf{x})$ and any $\epsilon < 1$, we have

$$(1-\epsilon)\ f(\mathbf{x}) = f((1-\epsilon)\mathbf{x}) \geq f(\mathbf{x}) - \epsilon \mathbf{v}^T \mathbf{x} \ . \tag{6.42}$$

Thus, $\epsilon\ f(\mathbf{x}) \leq \epsilon \mathbf{v}^T \mathbf{x}$. Choosing $\epsilon_1 = 1/2$ and $\epsilon_2 = -1/2$ yields $\mathbf{v}^T \mathbf{x} = f(\mathbf{x})$ which concludes the proof. □

If we apply Lemma 12 to the objective function in (6.40), we can replace $\mathbf{v}^T \mathbf{x}_0$ with $f(\mathbf{x}_0)$ for all $\mathbf{v} \in \partial f(\mathbf{x}_0)$, which gives the following optimization,

$$\min_{\substack{s\in\mathbb{R} \\ r\geq 0}} \max_{\beta \geq 0} \ -(1-s)L(\mathbf{x}_0) - r \cdot \min_{\mathbf{v} \in \lambda \mathbf{P}^\perp \partial L(\mathbf{x}_0)} \|\mathbf{v} - \frac{\beta}{\sqrt{n}}\mathbf{h}\| + \beta\sqrt{\alpha\ \mathbf{c_d}(s,r)} \ . \tag{6.43}$$

Recall that $(1-s)$ and $r \geq 0$ respectively represent the norm of the error in the direction of $\mathbf{x}_0$ and its orthogonal complement. Therefore, the perfect recovery in our optimization corresponds to the case where the optimizers are $r^\star = 0$ and $s^\star = 1$, and we are interested in the phase transition ratio $\alpha^\star$ for which this happens.

We use the following approximation of the objective function near the point $(r, s) = (0, 1)$, which was introduced earlier in (6.30),

$$\mathbf{c_d}(s,r) = \frac{1}{\pi}\Big[((1-s)^2 + r^2)\text{atan}(\frac{r}{1-s}) - r(1-s)\Big] \ . \tag{6.44}$$

Next, we define the new variable $t := \frac{r}{1-s}$ and rewrite the optimization in terms of t and s. One can show from (6.44) that as $r \downarrow 0$ and $s \uparrow 1$, the value of $\mathbf{c_d}(s,r)$ will only depends on the ratio t.

$$\min_{\substack{s\in\mathbf{R} \\ t\geq 0}} \max_{\beta \geq 0} \ \Psi(s,t,\beta) = -(1-s)L(\mathbf{x}_0) - t(1-s)\cdot \text{dist}_{\lambda \partial L^\perp(\mathbf{x}_0)}(\frac{\beta}{\sqrt{n}}\mathbf{h}) + \beta(1-s)\sqrt{\alpha\big((1+t^2)\text{atan}(t) - t\big)} \ , \tag{6.45}$$

where $\partial L^{\perp}(\mathbf{x}_0) = \mathbf{P}^{\perp}\partial L(\mathbf{x}_0)$ and $\text{dist}_{\mathcal{S}}(\mathbf{x})$ is the distance function defined in Definition 4. Since we have a convex-concave objective function over three scalars, we can write the first order optimality conditions for the solutions to (6.45) as follows,

$$\begin{cases} \frac{\partial}{\partial \beta}\Psi(s,t,\beta)\big|_{(s^\star=1,t^\star,\beta^\star)} = 0 \\ \frac{\partial}{\partial s}\Psi(s,t,\beta)\big|_{(s^\star=1,t^\star,\beta^\star)} = 0 \\ \frac{\partial}{\partial t}\Psi(s,t,\beta)\big|_{(s^\star=1,t^\star,\beta^\star)} = 0 \end{cases} \tag{6.46}$$

Next, we want to find the conditions (on α) under which the solution to (6.46) happens at $s^\star = 1$. Therefore, we aim to solve the system of nonlinear equations (6.46), for three unknowns t, β and δ.

These equations can be written in the following form,

$$\begin{aligned} &- t \cdot \frac{\frac{\beta}{n}\|\mathbf{h}\|^2 - \frac{\mathbf{h}^T}{\sqrt{n}}\Pi_{\partial L^{\perp}(\mathbf{x}_0)}(\frac{\beta}{\sqrt{n}}\mathbf{h})}{\text{dist}_{\lambda\partial L^{\perp}(\mathbf{x}_0)}(\frac{\beta}{\sqrt{n}}\mathbf{h})} + \sqrt{\alpha((1+t^2)\text{atan}(t) - t)} = 0 \\ &L(\mathbf{x}_0) + t \cdot \text{dist}_{\lambda\partial L^{\perp}(\mathbf{x}_0)}(\frac{\beta}{\sqrt{n}}\mathbf{h}) - \beta\sqrt{\alpha((1+t^2)\text{atan}(t) - t)} = 0 \\ &- \text{dist}_{\lambda\partial L^{\perp}(\mathbf{x}_0)}(\frac{\beta}{\sqrt{n}}\mathbf{h}) + \frac{\beta\, t\, \text{atan}(t)\sqrt{\alpha}}{\sqrt{\alpha((1+t^2)\text{atan}(t) - t)}} = 0 \end{aligned} \tag{6.47}$$

Next, we exploit the conditions of Assumption 4. Using theorem 5.2.2. in [141], both the functions $\text{dist}_{\lambda\partial L^{\perp}(\mathbf{x}_0)}(\frac{\beta}{\sqrt{n}}\mathbf{h})$ and $\frac{\mathbf{h}^T}{\sqrt{n}}\Pi_{\partial L^{\perp}(\mathbf{x}_0)}(\frac{\beta}{\sqrt{n}}\mathbf{h})$ converge point-wise to their expected value. Moreover, from Assumption 4, we know that both $\mathbb{E}[\text{dist}_{\lambda\partial L^{\perp}}(\mathbf{x}_0)(\frac{\beta}{\sqrt{n}}\mathbf{h})]$ and $\mathbb{E}[\frac{\mathbf{h}^T}{\sqrt{n}}\Pi_{\partial L^{\perp}(\mathbf{x}_0)}(\frac{\beta}{\sqrt{n}}\mathbf{h})]$ converge uniformly to $G_\lambda(\beta)$ and $\beta - F_\lambda(\beta)$, respectively. Therefore, using the same arguments as in [129], we can replace $\text{dist}_{\lambda\partial L^{\perp}(\mathbf{x}_0)}(\frac{\beta}{\sqrt{n}}\mathbf{h})$ with $F_\lambda(\beta)$ in the optimization (6.45), and then apply the result of Theorem 7.17 in [111], we can show that $F'_\lambda(\beta) = G_\lambda(\beta)\; G'_\lambda(\beta)$.

Therefore, we are able to use the functions F_λ, and G_λ to rewrite the system of non-linear equations (6.47):

$$\begin{aligned} &- t \cdot \frac{F_\lambda(\beta)}{G_\lambda(\beta)} + \sqrt{\alpha((1+t^2)\text{atan}(t) - t)} = 0 \\ &L(\mathbf{x}_0) + t \cdot G_\lambda(\beta) - \beta\sqrt{\alpha((1+t^2)\text{atan}(t) - t)} = 0 \\ &- G_\lambda(\beta) + \frac{\beta\, t\, \text{atan}(t)\sqrt{\alpha}}{\sqrt{\alpha((1+t^2)\text{atan}(t) - t)}} = 0 \end{aligned} \tag{6.48}$$

By combining the first and third equations, we will get

$$t = \tan(\frac{\pi}{\alpha\beta} F_\lambda(\beta)) \tag{6.49}$$

Finally, using (6.49) in (6.48) reduces the number of equations to 2, and yields the following system of non-linear equations.

$$\begin{cases} G_\lambda(\beta)\ l = \tan(\frac{\pi}{\alpha\beta} F_\lambda(\beta))\ (G_\lambda^2(\beta) - \beta F_\lambda(\beta))\ , \\ \tan(\frac{\pi}{\alpha\beta} F_\lambda(\beta))\ (G_\lambda(\beta) - \frac{\pi l}{\alpha\beta} F_\lambda(\beta)) = \frac{\pi}{\alpha\beta} F_\lambda(\beta)\ G_\lambda(\beta)\ , \end{cases} \tag{6.50}$$

This concludes the proof.

6.7 Proof of Lemma 10

Define matrix $\mathbf{A} \in \mathbf{R}^{m\times n}$ with i^{th} row equal to the measurement vector $\mathbf{a}_i$, for $i = 1, 2, \ldots, m$. Let $\mathbf{y} := |\mathbf{A}\mathbf{x}_0| \in \mathbb{R}^m$ denote the measurement values. To streamline our analysis, we assume $\|\mathbf{x}_0\| = 1$. One can rewrite the constraint set of the optimization problem (6.2) as following,

$$|\mathbf{A}\mathbf{x}| \le \mathbf{y} \ \Leftrightarrow\ -\mathbf{A}\mathbf{x} + \mathbf{y} \ge \mathbf{0}\ , \text{ and } \ \mathbf{A}\mathbf{x} + \mathbf{y} \ge \mathbf{0}, \tag{6.51}$$

where all the inequalities are component-wise. Exploiting the Lagrange multipliers, we can reformulate the generalized PhaseMax optimization as,

$$\min_{\mathbf{x}\in\mathbb{R}^n} \max_{\boldsymbol{\mu},\boldsymbol{\eta}\in\mathbb{R}^m_+} \quad -{\mathbf{x}_{\text{init}}}^T\mathbf{x} + \lambda f(\mathbf{x}) + (\boldsymbol{\mu}-\boldsymbol{\eta})^T\mathbf{A}\mathbf{x} - (\boldsymbol{\mu}+\boldsymbol{\eta})^T\mathbf{y}\ , \tag{6.52}$$

where μ_i and η_i are Lagrange multipliers for the inequalities $\mathbf{a}_i^T\mathbf{x} \le y_i$ and $\mathbf{a}_i^T\mathbf{x} \ge -y_i$, respectively. Assume $y_i > 0$ (which happens with probability 1), these two inequalities cannot be active at the same time. Therefore, at least one of μ_i and η_i must be equal to 0, for every $i = 1, 2, \ldots, m$. Hence, we have $\boldsymbol{\mu} + \boldsymbol{\eta} = |\boldsymbol{\mu} - \boldsymbol{\eta}|$. Here $|\cdot|$ denotes the component-wise absolute value function. Define $\mathbf{v} := \boldsymbol{\mu} - \boldsymbol{\eta}\ \in \mathbb{R}^m$ and rewrite the optimization in terms of $\mathbf{v}$ gives the following,

$$\min_{\mathbf{x}\in\mathbb{R}^n} \max_{\mathbf{v}\in\mathbb{R}^m} \quad -{\mathbf{x}_{\text{init}}}^T\mathbf{x} + \lambda f(\mathbf{x}) + \mathbf{v}^T\mathbf{A}\mathbf{x} - |\mathbf{v}|^T|\mathbf{A}\mathbf{x}_0|\ . \tag{6.53}$$

Since the term $|\mathbf{v}|^T|\mathbf{A}\mathbf{x}_0|$ depends on the matrix $\mathbf{A}$, it is not possible to apply the CGMT to the bilinear form $\mathbf{v}^T\mathbf{A}\mathbf{x}$. In order to apply CGMT, we use the following key decomposition for $\mathbf{x}$:

$$\mathbf{x} = s\mathbf{x}_0 + \mathbf{w}, \tag{6.54}$$

where $s = \mathbf{x}_0^T\mathbf{x} \in \mathbb{R}$ is a scalar and the vector $\mathbf{w} = \mathbf{P}^{\perp}\mathbf{x} \in \mathbb{R}^n$ is orthogonal to $\mathbf{x}_0$. We can rewrite the optimization problem (6.53) in terms of s and $\mathbf{w}$ as follows,

$$\min_{s\in\mathbb{R},\ \mathbf{w}\perp\mathbf{x}_0}\ \max_{\mathbf{v}\in\mathbb{R}^m} \quad -s\rho_{\text{init}} - \mathbf{x}_{\text{init}}^T\mathbf{w} + \lambda f(s\mathbf{x}_0 + \mathbf{w}) + \mathbf{v}^T\mathbf{A}\mathbf{w} + s\mathbf{v}^T\mathbf{A}\mathbf{x}_0 - |\mathbf{v}|^T|\mathbf{A}\mathbf{x}_0|, \tag{6.55}$$

where $\rho_{\text{init}} = \mathbf{x}_{\text{init}}^T\mathbf{x}_0$. Next, we use the following property of Gaussian matrices.

Lemma 13. *Let $\mathbf{G} \in \mathbb{R}^{m\times n}$ be a random matrix with i.i.d. standard normal entries, and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ are such that $\mathbf{u} \perp \mathbf{v}$. The random vectors $\mathbf{G}\mathbf{u}$ and $\mathbf{G}\mathbf{v}$ are independent.*

Proof. Let $\mathbf{G} = [g_{i,j}]_{m\times n}$ and define $\mathbf{a} = \mathbf{G}\mathbf{u}$, and $\mathbf{b} = \mathbf{G}\mathbf{v}$. Since $\mathbf{G}$ has Gaussian entries $\mathbf{a}, \mathbf{b}$ are Gaussian vectors in $\mathbb{R}^m$. Therefore, to show their independence it is sufficient to show that $\mathbb{E}[\mathbf{a}, \mathbf{b}^T] = \mathbf{0}_{m\times m}$.

$$\mathbb{E}[a_i b_j] = \sum_{k=1}^{n}\sum_{l=1}^{n} u_k v_l\ \mathbb{E}[g_{i,k}\ g_{j,l}] = \begin{cases} \sum_{k=1}^{n} u_k v_k = 0, & \text{if } i = j \\ 0\,, & \text{if } i \neq j \end{cases}, \tag{6.56}$$

where we used the fact that $\mathbf{u}^T\mathbf{v} = \sum_{k=1}^{n} u_k v_k = 0$.

□

Using the result of Lemma 13 , the random vectors $\mathbf{A}\mathbf{x}_0$ and $\mathbf{A}\mathbf{w}$ are independent. So, we are allowed to change the matrix $\mathbf{A}$ in the bilinear form $\mathbf{v}^T\mathbf{A}\mathbf{w}$ with its independent copy $\mathbf{H} \in \mathbb{R}^{m\times n}$ which also has i.i.d. standard normal entries. We also define $\mathbf{q} = \mathbf{A}\mathbf{x}_0 \in \mathbb{R}^m$, which is independent from $\mathbf{H}$. Note that since $\mathbf{A}$ has i.i.d. normal entries and $\|\mathbf{x}_0\| = 1$, the entries of $\mathbf{q}$ also has i.i.d. standard normal distribution. We can rewrite the optimization (6.55) as follows:

$$\min_{s\in\mathbb{R},\ \mathbf{w}\perp\mathbf{x}_0}\ \max_{\mathbf{v}\in\mathbb{R}^m} \quad -s\rho_{\text{init}} - \mathbf{x}_{\text{init}}^T\mathbf{w} + \lambda f(s\mathbf{x}_0 + \mathbf{w}) + \mathbf{v}^T\mathbf{H}\mathbf{w} + s\mathbf{v}^T\mathbf{q} - |\mathbf{v}|^T|\mathbf{q}|\ . \tag{6.57}$$

Next, we apply the CGMT framework in Lemma 32 to equation (6.57), in order to replace the bilinear form $\mathbf{v}^T\mathbf{H}\mathbf{w}$ with two linear forms $\|\mathbf{v}\|\mathbf{h}^T\mathbf{w} + \mathbf{v}^T\mathbf{g}\|\mathbf{w}\|$. But this lemma requires the set that we optimize $\mathbf{w}$ over to be compact. In order to be able to apply CGMT, we enforce an "artificial" bound on the norm of $\mathbf{w}$. Note that our goal is to eventually prove that, $\hat{\mathbf{w}}$ converges to a finite number

$\alpha^\star$. We define $K_\alpha = \alpha^\star + \Delta$ for some $\Delta > 0$ and also the compact set $\mathcal{S}_{\mathbf{w}} = \{\mathbf{w} | \mathbf{w} \perp \mathbf{x}_0 \,,\ \|\mathbf{w}\| \leq K_\alpha\}$. Let $\hat{\mathbf{w}}^{\text{temp}}$ to be the optimizer to the version of (6.57) where we optimize $\mathbf{w}$ over $\mathcal{S}_{\mathbf{w}}$. It is simple to verify that if $\|\hat{\mathbf{w}}^{\text{temp}}\| \xrightarrow{\mathbb{P}} \alpha^\star$, then $\|\hat{\mathbf{w}}\| \xrightarrow{\mathbb{P}} \alpha^\star$. This means that if in the final equation, we get a unique finite solution for the asymptotic behavior of $\|\hat{\mathbf{w}}\|$ (which is what we do) , the proof goes though and we can apply the CGMT.

Now that this concern is taken care of, the following corollary will be the result of applying CGMT to the equation (6.57).

Corollary 5. *Let $\hat{\mathbf{x}}$ be the unique optimizer of the generalized PhaseMax algorithm* (6.2)*. As $n \to \infty$ the error performance converges in probability as follows:*

$$\left(\frac{\|\hat{\mathbf{x}} - \mathbf{x}_0\|^2}{\|\mathbf{x}_0\|^2}\right) - \left(\|\mathbf{w}^\star\|^2 + (1 - s^\star)^2\right) \overset{n\to\infty}{\longrightarrow} 0 \,, \tag{6.58}$$

where $s^\star$, $\mathbf{w}^\star$ are the unique optimizers of the following (auxiliary) optimization:

$$\min_{s\in\mathbb{R},\, \mathbf{w}\perp\mathbf{x}_0} \max_{\mathbf{v}\in\mathbb{R}^m} \quad -s\rho_{init} - \mathbf{x}_{init}^T\mathbf{w} + \lambda f(s\mathbf{x}_0 + \mathbf{w}) - \|\mathbf{v}\|\mathbf{h}^T\mathbf{w} + \mathbf{v}^T\mathbf{g}\|\mathbf{w}\| + s\mathbf{v}^T\mathbf{q} - |\mathbf{v}|^T|\mathbf{q}| \,. \tag{6.59}$$

$\mathbf{h} \in \mathbb{R}^n$ and $\mathbf{g} \in \mathbb{R}^m$ are random vectors with i.i.d. standard normal entries.

We proceed onward with analyzing (6.59). Observe that if we fix $|\mathbf{v}|$, then the optimal $\mathbf{v}$ satisfies $\text{sign}(\mathbf{v}) = \text{sign}(\|\mathbf{w}\|\mathbf{g} + s\mathbf{q})$ which simplifies the optimization to the following,

$$\min_{s\in\mathbb{R},\, \mathbf{w}\perp\mathbf{x}_0} \max_{\mathbf{v}\in\mathbb{R}^m} \quad -s\rho_{\text{init}} - \mathbf{x}_{\text{init}}^T\mathbf{w} + \lambda f(s\mathbf{x}_0 + \mathbf{w}) - ||\mathbf{v}||\mathbf{h}^T\mathbf{w} + |\mathbf{v}|^T(|\ s\mathbf{q} + \|\mathbf{w}\|\mathbf{g}\ | - |\mathbf{q}|), \tag{6.60}$$

By fixing the norm of $\mathbf{v}$ and optimizing over its direction, the optimization problem (6.60) can be reduced to the following:

$$\begin{aligned} \min_{s\in\mathbb{R},\, \mathbf{w}\perp\mathbf{x}_0} \quad & -s\rho_{\text{init}} - \mathbf{x}_{\text{init}}^T\mathbf{w} + \lambda f(s\mathbf{x}_0 + \mathbf{w}) \\ \text{subject to:} \quad & \mathbf{h}^T\mathbf{w} \geq ||\{|\ s\mathbf{q} + ||\mathbf{w}||\mathbf{g}\ | - |\mathbf{q}|\}_+|| \,, \end{aligned} \tag{6.61}$$

where, for a vector $\mathbf{c}$, we let $\{\mathbf{c}\}_+$ denote the component-wise positive part function, with i^{th} entry equal to $\max(0, \mathbf{c}_i)$. Next, note that $\mathbf{q}$ and $\mathbf{g}$ are independent vectors in $\mathbb{R}^m$ with i.i.d. standard normal entries. We introduce the function $\mathbf{c_d}(s, r)$ as follows:

Definition 7. *The function* $\mathbf{c_d} : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ *is defined as,*

$$\mathbf{c_d}(s, r) = \mathbb{E}_{X_1, X_2}[\{|sX_1 + rX_2| - |X_1|\}_+^2] \ , \tag{6.62}$$

where $X_1, X_2 \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$.

Lemma 14. $\frac{1}{m}||\{|\ s\mathbf{q} + ||\mathbf{w}||\mathbf{g}\ | - |\mathbf{q}|\}_+||^2 \xrightarrow{\mathbb{P}} \mathbf{c_d}(s, \|\mathbf{w}\|)$, *as* $m \to \infty$.

Proof. Define the vector $\mathbf{u} \in \mathbb{R}_+^m$ as,

$$\mathbf{u} := \{|\ s\mathbf{q} + \|\mathbf{w}\|\mathbf{g}\ | - |\mathbf{q}|\}_+ \ . \tag{6.63}$$

The entries of $\mathbf{u}$ are i.i.d and $\mathbb{E}[u_i^2] = \mathbf{c_d}(s, \|\mathbf{w}\|)$, for $1 \leq i \leq m$. Therefore, the weak law of large number gives the following:

$$\frac{1}{m}||\mathbf{u}||^2 = \frac{1}{m}\sum_{i=1}^{m} u_i^2 \ \xrightarrow{\mathbb{P}} \ \mathbb{E}[u_i^2] = \mathbf{c_d}(s, \|\mathbf{w}\|) \ . \tag{6.64}$$

□

To conclude the proof of Lemma 10, we exploit the result of Lemma 14 to replace $\|\{|s\mathbf{q} + ||\mathbf{w}||\mathbf{g}| - |\mathbf{q}|\}_+||$ in (6.61), which gives us the following optimization:

$$\begin{aligned} \min_{s\in\mathbb{R}} \min_{\mathbf{w}\in\mathbb{R}^n, \mathbf{w}\perp\mathbf{x}_0} \quad & -\mathbf{x}_{\text{init}}^T(s\mathbf{x}_0 + \mathbf{w}) + \lambda f(s\mathbf{x}_0 + \mathbf{w}) \\ \text{subject to:} \quad & \mathbf{h}^T\mathbf{w} \geq \sqrt{m\ \mathbf{c_d}(s, \|\mathbf{w}\|)} \ . \end{aligned} \tag{6.65}$$

We are not going through the technical details of obtaining the convergence result in (6.65). The point-wise convergence, for fixed values of s and $\|\mathbf{w}\|$, follows from Lemma 14. To show the uniform convergence, we appeal to the convexity of the objective function. The corresponding convergence of the optimal cost follows from the uniform convergence.

The following lemma gives an explicit formula for the function $\mathbf{c_d}(\cdot, \cdot)$ in terms of its two input arguments.

Lemma 15.

$$\mathbf{c_d}(s, t) = \frac{1}{\pi}\Big[\big((1+s)^2 + t^2\big)\ atan\big(\frac{t}{1+s}\big)\ + \big((1-s)^2 + t^2\big)\ atan\big(\frac{t}{1-s}\big) - 2t\Big]. \tag{6.66}$$

Proof.

$$\mathbf{c_d}(s,t) = \mathbb{E}\big[\{|sX_1 + tX_2| - |X_1|\}_+^2\big] \tag{6.67}$$

$$= \frac{1}{\pi}\int_0^\infty e^{-x_1^2/2}\int_{\frac{1-s}{t}x_1}^\infty e^{-x_2^2/2}(tx_2 - (1-s)x_1)^2\ dx_2\ dx_1 \tag{6.68}$$

$$+ \frac{1}{\pi}\int_0^\infty e^{-x_1^2/2}\int_{-\infty}^{-\frac{1+s}{t}x_1} e^{-x_2^2/2}\,(tx_2 + (1+s)x_1)^2\ dx_2\ dx_1\ ,$$

□

where, due to the symmetry, we have computed the expectation only for $X_1 > 0$ and multiplied the result by two. Next, we rewrite the integral in the polar coordinates (r,θ) where $x_1 = r\cos(\theta)$, and $x_2 = r\sin(\theta)$.

$$\mathbf{c_d}(s,t) = \frac{1}{\pi}\int_{\text{atan}\left(\frac{1-s}{t}\right)}^{\pi/2}\int_0^\infty r^3 e^{-r^2/2}\,(t\sin(\theta) - (1-s)\cos(\theta))^2\ dr\ d\theta \tag{6.69}$$

$$+ \frac{1}{\pi}\int_{\text{atan}(\frac{1+s}{t})}^{\pi/2}\int_0^\infty r^3 e^{-r^2/2}\,(t\sin(\theta) - (1+s)\cos(\theta))^2\ dr\ d\theta$$

$$= \frac{2}{\pi}\int_{\text{atan}\left(\frac{1-s}{t}\right)}^{\pi/2}(t\sin(\theta) - (1-s)\cos(\theta))^2\ d\theta + \frac{2}{\pi}\int_{\text{atan}(\frac{1+s}{t})}^{\pi/2}(t\sin(\theta) - (1+s)\cos(\theta))^2\ d\theta \tag{6.70}$$

$$= \frac{2}{\pi}\left[\frac{t^2+(1-s)^2}{2}\text{atan}\Big(\frac{t}{1-s}\Big) - \frac{t(1-s)}{2}\right] + \frac{2}{\pi}\left[\frac{t^2+(1+s)^2}{2}\text{atan}\Big(\frac{t}{1+s}\Big) - \frac{t(1+s)}{2}\right] \tag{6.71}$$

$$= \frac{1}{\pi}\left[((1+s)^2+t^2)\ \text{atan}\Big(\frac{t}{1+s}\Big)\ + \big((1-s)^2+t^2\big)\ \text{atan}\Big(\frac{t}{1-s}\Big) - 2t\right]\ . \tag{6.72}$$

We derived (6.70) using the fact that $\int_0^\infty r^3 e^{-r^2/2}dr = 2$. Computing each of the integrals with respect to θ would result in (6.71), and the final result in (6.72) is derived by some simplifications.

6.8 Computing the Statistical Dimension for the ℓ_1 Regularization

In this section we bound the statistical dimension of the descent cone of the objective function in (6.2), where $f(\cdot) = \|\cdot\|_1$ is used for regularization. We assume that the underlying signal, $\mathbf{x}_0$, is k-sparse and define function $L_\lambda : \mathbb{R}^n \to \mathbb{R}$ as follows,

$$L_\lambda(\mathbf{x}) := -\mathbf{x}_{\text{init}}^T\mathbf{x} + \lambda\|\mathbf{x}\|_1, \tag{6.73}$$

In order to derive an upper bound for the statistical dimension $d\left(T_{L_\lambda}(\mathbf{x}_0)\right)$, we first introduce another summary parameter for convex sets called the Gaussian width.

Definition 8 (Gaussian width [141])**.** *The Gaussian width of a subset $\mathcal{T} \subset \mathbb{R}^n$ is defined as,*

$$\omega(\mathcal{T}) = \mathbb{E} \sup_{\mathbf{x} \in \mathcal{T}} \langle \mathbf{x}, \mathbf{g} \rangle, \quad \textit{where } \mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \ . \tag{6.74}$$

The following proposition indicates the relationship between the Gaussian width and statistical dimension of a convex cone.

Proposition 2 (Proposition 10.2 in [7])**.** *Let $C \subset \mathbb{R}^n$ be a convex cone. Then*

$$\omega^2(C \cap \mathbb{S}^{n-1}) \leq d(C) \leq \omega^2(C \cap \mathbb{S}^{n-1}) + 1 \ , \tag{6.75}$$

where $\mathbb{S}^{n-1} \subset \mathbb{R}^n$ is the unit sphere.

The Proposition 2 shows that in order to bound the statistical dimension of a convex cone, we need to bound the squared Gaussian width of that cone. Hence, in the remaining of this section we will bound the squared Gaussian width. To this goal, we briefly review some known properties of Gaussian width of convex cones.

Some properties of Gaussian width

The Gaussian width is one of the intrinsic volumes of a body studied in combinatorial geometry. It is invariant under translation and unitary transformation and has deep connections to convex geometry. While discussing all the properties of Gaussian width is beyond the scope of this paper, we refer the interested reader to [110, 133, 141] and references therein.

Inspired by [122, 33], here we bound the Gaussian width of a cone via the distance to its polar cone. Before stating the proposition, we review the definition of the polar cone.

Definition 9 (Polar cone)**.** *Let $C \subset \mathbb{R}^n$ be a non-empty convex cone. The polar cone of C, denoted by $C^\star$, is defined as follows,*

$$C^\star = \{\mathbf{z} \in \mathbb{R}^n : \langle \mathbf{z}, \mathbf{x} \rangle \leq 0 \ \textit{ for all } \mathbf{x} \in C\} \ . \tag{6.76}$$

The following proposition establishes a connection between the Gaussian width of the cone C and its polar cone $C^\star$:

Proposition 3 (Proposition 3.6 in [33]). *Let C be any non-empty convex cone in $\mathbb{R}^n$, and let $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I})$ be a random Gaussian vector. Then we have the following bound:*

$$\omega(C \cap \mathbb{S}^{n-1}) \leq \mathbb{E}_{\mathbf{g}}[dist(\mathbf{g}, C^{\star})] \ , \tag{6.77}$$

where the $dist(\cdot, \cdot)$ function here denotes the Euclidean distance between a point and a set.

Applying Jensen's inequality will result in the following,

$$\omega^2(C \cap \mathbb{S}^{n-1}) \leq \mathbb{E}_{\mathbf{g}}[\text{dist}^2(\mathbf{g}, C^{\star})] \ . \tag{6.78}$$

This is very useful in bounding the Gaussian width of the descent cone of a convex function due to We next appeal the following lemma to bound the Gaussian width:

Lemma 16 ([108]). *For a convex function $f : \mathbb{R}^n \to \mathbb{R}$,*

$$(T_f(\mathbf{x}))^{\star} = cone(\partial f(\mathbf{x})) \ , \tag{6.79}$$

where $\partial f(\mathbf{x})$ is the sub-differential set of function f at point $\mathbf{x}$.

The polar cone of $T_f(\mathbf{x})$ is also called the *normal cone*, $N_f(\mathbf{x})$, at point $\mathbf{x}$. Exploiting the above results, we can bound $\omega^2\left(T_L(\mathbf{x}_0) \cap \mathbb{S}^{n-1}\right)$ in terms of the squared distance to the normal cone at point $\mathbf{x}_0$, i.e., we have the following,

$$d(T_{L_\lambda}(\mathbf{x}_0)) \leq \ \omega^2\left(T_{L_\lambda}(\mathbf{x}_0) \cap \mathbb{S}^{n-1}\right) + 1 \ \leq \mathbb{E}_{\mathbf{g}}\left[\text{dist}^2(\mathbf{g}, \text{cone}(\partial L_\lambda(\mathbf{x}_0))\right] + 1 \ . \tag{6.80}$$

For simplicity in the remaining formulations we omit the sub-script λ and denote the objective function by L. Let Δ denote the set of coordinates where $\mathbf{x}_0$ is non-zero. The sub-differential set of the function L (defined in (6.73)) can be formally characterized as,

$$\partial L(\mathbf{x}_0) = \{-\mathbf{x}_{\text{init}} + \frac{\lambda}{\sqrt{k}}\mathbf{v} : \mathbf{v} \in \mathbb{R}^n \text{ s.t. } \mathbf{v}[i] = \text{sign}(\mathbf{x}_0[i]) \text{ for } i \in \Delta, \ |\mathbf{v}[i]| \leq 1 \text{ for } i \in \Delta^c\}. \tag{6.81}$$

Here $\Delta^c := [n] \backslash \Delta$ represents the zero entries of $\mathbf{x}_0$. Without the loss of generality, we are going to assume that the first k entries of $\mathbf{x}_0$ are non-zero, while the rest are zero. Then, cone of the sub-differential can be rewritten as

$$\text{cone}(\partial f(\mathbf{x})) = \{\beta \cdot (-\mathbf{x}_{\text{init}} + \frac{\lambda}{\sqrt{k}}\mathbf{v}) : \mathbf{v}[1:k] = \mathbf{1},\ \|\mathbf{v}[k+1:n]\|_\infty \leq 1,\ \beta \geq 0\}. \tag{6.82}$$

The squared distance to the normal cone can be formulated as the following optimization:

$$\begin{aligned} \text{dist}^2(\mathbf{g}, N_L(\mathbf{x}_0)) = \min_{t \geq 0} & \sum_{i \in \Delta} (\mathbf{g}[i] + t\mathbf{x}_{\text{init}}[i] - t\lambda \text{sign}(\mathbf{x}_0[i]))^2 \\ & + \sum_{j \in \Delta^c} \min_{|u_j| < t} (\mathbf{g}[j] + t\mathbf{x}_{\text{init}}[j] - \lambda u_j)^2 \end{aligned} \tag{6.83}$$

Define $\mathbf{z} := \mathbf{z}(t) = \mathbf{g} + t\mathbf{x}_{\text{init}}$. We can rewrite the equation (6.83) as,

$$\begin{aligned} \text{dist}^2(\mathbf{g}, N_L(\mathbf{x}_0)) = \min_{t \geq 0} & \sum_{i \in \Delta} (\mathbf{z}[i] - t\lambda \text{sign}(\mathbf{x}_0[i]))^2 \\ & + \sum_{j \in \Delta^c} \text{shrink}(\mathbf{z}[j], t\lambda)^2 \,, \end{aligned} \tag{6.84}$$

where the function $\text{shrink}(\cdot, \cdot)$ is defined as, $\text{shrink}(x, T) = \begin{cases} x + T \,, & x < -T \\ 0 \,, & -T \leq x \leq T \\ x - T \,, & x > T \end{cases}$. This function is known as ℓ_1-shrinkage function and is used in sparse denoising. Taking the expectation with respect to $\mathbf{g}$ will provide us with the quantity we would like to bound. We are bounding the expectation of the squared distance by bounding each of the two terms of the sum. For the first term, we have:

$$\mathbb{E}\big[\sum_{i \in \Delta} (\mathbf{z}[i] - t\lambda \text{sign}(\mathbf{x}_0[i]))^2\big] = k + t^2(\lambda^2 k - ||\mathbf{x}_{\text{init}}^\Delta||^2) - 2t\lambda \text{sign}(\mathbf{x}_0)^T \mathbf{x}_{\text{init}}^\Delta \tag{6.85}$$

Bounding the expectation of the second term in (6.84) would result in,

$$\mathbb{E}[\sum_{j \in \Delta^c} \text{shrink}^2(\mathbf{z}[j], t\lambda)] \leq \frac{2(n-k)\sigma^3}{\sqrt{2\pi} t\lambda} exp(\frac{-t^2\lambda^2}{2\sigma^2}) \,, \tag{6.86}$$

where $\sigma^2 := 1 + \frac{||\mathbf{x}_{\text{init}}^{\Delta}||^2}{n-k}\, t^2$.

Using the result of equations (6.85) and (6.86), we can see that when $\lambda > \frac{c}{\sqrt{k}}$, then by choosing $t = \frac{\sqrt{2 \log \frac{n}{k}}}{\lambda}$, both of the terms in the sum are bounded by $Ck \log \frac{n}{k}$, where c and C are constants that are independent of the problem's parameters.

Therefore, when $\lambda > \frac{c}{\sqrt{k}}$ (where k is the number of non-zero entries), the statistical dimension of $T_{L_\lambda}(\mathbf{x}_0)$ is bounded by $Ck \log \frac{n}{k}$. This has been used in Example 2 in Section 6.3. Using the result of Theorem 7, we can conclude that for the sparse phase retrieval problem the required sample complexity of the regularized PhaseMax is $O(k \log \frac{n}{k})$. This indicates that regularized PhaseMax is order-wise optimal (given a proper initialization).

Part II:

Linear Classification with Structured Parameters

LINEAR MODELS FOR BINARY CLASSIFICATION

In this chapter, we introduce the problem of classification with linear decision boundaries. As explained earlier in Chapter 1, this problem falls under the category of generalized linear models. In binary classification, a linear decision boundary can be seen as a hyperplane that separates the two classes. Given a training data set, the goal is to find the parameters of the classifier (equivalently the normal vector of the separating hyperplane).

Mathematically speaking, we assume that we have access to the training data $\mathcal{D} = \{(\mathbf{x}_i, y_i) | 1 \leq i \leq n\}$, where, for $i = 1, 2, \ldots, n$, $\mathbf{x}_i \in \mathbb{R}^p$ denotes the feature vector and $y_i \in \{+1, -1\}$ indicates the class label, which is a symmetric Bernoulli random variable with,

$$\mathbb{P}[y_i = 1 | \mathbf{x}_i] = \Phi(\mathbf{x}_i^T \mathbf{w}^\star). \tag{7.1}$$

Here, $\mathbf{w}^\star \in \mathbb{R}^p$ is the parameter vector of the underlying model, and $\Phi : \mathbb{R} \to \mathbb{R}$ is a known real-valued (non-linear) function, often referred to as *the link function*. A well-known example of such model is logistic regression in which the function Φ is the logistic function.

$$\text{logistic model:} \quad \mathbb{P}[y_i = 1 | \mathbf{x}_i] = \frac{e^{\mathbf{x}_i^T \mathbf{w}^\star}}{e^{\mathbf{x}_i^T \mathbf{w}^\star} + e^{-\mathbf{x}_i^T \mathbf{w}^\star}}. \tag{7.2}$$

Another important example is the noise-free model, in which the hyperplane with the normal vector $\mathbf{w}^\star$ perfectly separates the two classes.

$$\text{noise-free model:} \quad \mathbb{P}[y_i = 1 | \mathbf{x}_i] = \text{SIGN}(\mathbf{x}_i^T \mathbf{w}^\star). \tag{7.3}$$

It is worth noting that thew logistic model converges to a noise-free model as $||\mathbf{w}^\star|| \to +\infty$.

Here, we assume that the link function, $\Phi(\cdot)$ is known, and the goal is to find $\mathbf{w}^\star$ given the data set $\mathcal{D}$. We assume that the data points $\mathbf{x}_i$'s are generated *independently* from a distribution $\mathbb{P}_x$. Inspired by the maximum-likelihood estimator in logistic regression, one can attempt to find $\mathbf{w}^\star$ by solving an optimization problem of the form:

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{R}^p} \sum_{i=1}^{n} \ell(\mathbf{x}_i^T \mathbf{w}, y_i), \tag{7.4}$$

where the objective function is additive with respect to each of the data points, and the function $\ell : \mathbb{R}^2 \to \mathbb{R}$ is defined based on the link function $\Phi(\cdot)$. For instance, we will show in Chapter 8 that the maximum-likelihood estimator can be derived by choosing $\ell(\cdot, \cdot)$ as,

$$\ell(u, v) = \log(1 + \exp(-uv)), \tag{7.5}$$

which gives the following optimization program:

$$\hat{\mathbf{w}}_{LR} = \arg \min_{\mathbf{w} \in \mathbb{R}^p} \sum_{i=1}^{n} \log \left[1 + \exp(y_i \mathbf{x}_i^T \mathbf{w})\right]. \tag{7.6}$$

It turns out that the optimization program (7.6) exhibits different behaviors depending on the *separability* of the data set.

Definition 10. *The data set $\mathcal{D} = \{(\mathbf{x}_i, y_i) | 1 \leq i \leq n\}$ is said to be separable if there exists $\mathbf{w}_0 \in \mathbb{R}^p$ such that $y_i = SIGN(\mathbf{w}_0^T \mathbf{x}_i)$ for $i = 1, 2, \ldots, n$.*

7.1 Non-separable Data sets

When n/p is sufficiently large, i.e., when we have access to a sufficiently large number of samples, the maximum-likelihood estimator is well-defined. In this case, the training data set $\mathcal{D}$ is non-separable, and the optimization program (7.6) has a unique solution. Since this is a convex optimization, a descent method (such as gradient descent) can be used to achieve the optimal solution. There has been multiple studies in classical statistics on the performance of the estimate derived from the logistic regression when the data is non-separable. The main underlying assumptions in these studies is that the number of data samples, n, far exceeds the number of parameters, p. More specifically, it has been shown that when $n/p \to \infty$, the solution of the logistic regression $\hat{\mathbf{w}}_{LR}$ becomes an efficient estimator of the underlying parameters $\mathbf{w}^\star$, i.e.,

- $\hat{\mathbf{w}}_{LR}$ is an unbiased estimator of $\mathbf{w}^\star$.
- $\hat{\mathbf{w}}_{LR}$ has the minimum variance among all unbiased estimators. The covariance matrix of this estimator is equal to the inverse of the Fisher Information matrix.

We will observe in Chapter 8 that the above results, derived in the classical regime, do not hold when the number of samples is proportional to the number of parameters. In particular, we will see that when $\delta = n/p < +\infty$, the estimator derived from the logistic model is not even an unbiased estimator!

In practice, we often assume that the underlying parameters possess certain structure(s). To enforce the structure, we often add a regularization term to the objective function. The regularized logistic regression can be formulated as following,

$$\hat{\mathbf{w}}_{RLR} = \arg\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \log\left[1+\exp(y_i\mathbf{x}_i^T\mathbf{w})\right] + \frac{\lambda}{p}\sum_{k=1}^{p} R(w_k). \tag{7.7}$$

where $R(\cdot)$ is the regularization function that enforces the desired/expected structure.

In Chapter 8, we provide a precise asymptotic analysis of the solution of the regularized logistic regression. In particular, we characterize the performance via the solution to a nonlinear system of equations with 6 unknowns. We will show that any performance measures of the resulting estimator can be computed form the solution of this nonlinear system. We then investigate the performance for certain choices of the regularization functions: (1) ridge regularization, which is often used in ML applications to improve the performance of the estimator and to avoid overfitting, (2) ℓ_1 regularization, which is used to enforce sparse structures, and (3) ℓ_∞ regularization, which is used when the underlying parameter has a discrete (binary) structure.

7.2 Separable Data sets (Interpolating Regime)

In the setting where the data set $\mathcal{D}$ is separable, the optimization problem (7.6) does not have a unique solution. Let $\mathcal{W}$ denote the set of parameters that perfectly separates the data, i.e.,

$$\mathcal{W} = \{\mathbf{w}\in\mathbb{R}^p |\ y = \text{SIGN}(\mathbf{w}^T\mathbf{x})\ ,\ \text{for } (\mathbf{x},y)\in\mathcal{D}\}. \tag{7.8}$$

The regime of parameters when $\mathcal{D}$ is a separable data set (equivalently, when the set $\mathcal{W}$ is non-empty) is known as the *interpolating regime*.

To find an optimal solution of (7.6), we often use an iterative solver, which starts from an initialization point and follows an update rule till convergence. It has been observed that in many machine learning tasks, iterative solvers converge to one of the points in the set $\mathcal{W}$ (i.e., the training error converges to zero). Therefore, one can qualitatively pose the following important (yet still mysterious) question:

Which point(s) in $\mathcal{W}$ is (are) "better" estimator(s) of the actual parameter, $\mathbf{w}^\star$?

Studies on the performance of different classifiers for binary classification dates back to the seminal work of Vapnik in the 1980's [140]. In an effort to find the "optimal" hyperplane that separates the data, he presented an upper bound on the test error which is inversely proportional to the margin (minimum distance of the data points to the separating hyperplane), and concluded that the

max-margin classifier is indeed the desired classifier. It has also been observed that to construct such optimal hyperplanes one only needs to only take into consideration a small amount of the training data, the so-called support vectors [37].

A more recent line of research studies the convergence of iterative optimization algorithms (such as gradient descent) on the logistic loss in the interpolating regime. Soudry et al. [120] studied the behavior of gradient descent updates when applied on the logistic loss, i.e.,

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \times \nabla L(\mathcal{D}, \mathbf{w}) \text{ for } t \geq 0, \tag{7.9}$$

where η is the step-size and $L(\mathcal{D}, \mathbf{w}) = \sum_{i=1}^{n} \log\left[1 + \exp(y_i \mathbf{x}_i^T \mathbf{w})\right]$ is the objective function in logistic regression. The following Theorem characterizes the convergence behavior of GD iterates:

Theorem 9 (Theorem 3 in [120])**.** *Consider the optimization problem* (7.6)*, and the gradient descent iterates initialized at* $\mathbf{w}_0$ *and updated as in* (7.9) *with* $\eta < C_{\mathcal{D}}/n$ [1]*. We have,*

(i) $||\mathbf{w}_t|| \to +\infty$*, as* $t \to \infty$*, and,*

(ii) $\lim_{t\to\infty} \frac{\mathbf{w}_t}{||\mathbf{w}_t||} = \frac{\hat{\mathbf{w}}}{||\hat{\mathbf{w}}||}$*, where* $\hat{\mathbf{w}}$ *is the parameters of the max-margin classifier defined as,*

$$\begin{aligned} \hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^p} \quad & ||\mathbf{w}||_2 \\ s.t. \ \ & y_i(\mathbf{x}_i^T \mathbf{w}) \geq 1, \ \ for\ 1 \leq i \leq n. \end{aligned} \tag{7.10}$$

This result is very interesting as it sheds light on the convergence behavior of the GD iterates. While, there are multiple hyperplanes that separates the data, the GD converges to a special one that is the max-margin classifier. More importantly, in order to assess the performance of the result achieve by running gradient descent method, we should analyze the performance of the max-margin classifier.

In Chapter 9, we attempt to extend this result by considering the case where the underlying parameter, $\mathbf{w}^\star$ possesses certain structure (sparse, low-rank, block-sparse, etc.), and consider a convex function $\psi : \mathbb{R}^p \to \mathbb{R}$ which encourages this structure. We introduce the *Extended Margin Maximizer (EMM)* as the solution of the following optimization,

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^p} \quad & \psi(\mathbf{w}) \\ \text{s.t. } & y_i(\mathbf{x}_i^T \mathbf{w}) \geq 1, \ \ \text{for } 1 \leq i \leq n. \end{aligned} \tag{7.11}$$

[1] $C_{\mathcal{D}}$ is a constant that depends on the data set $\mathcal{D}$. It has been shown in [120] that for logistic loss $C_{\mathcal{D}} = \frac{2}{\sigma^2_{\max}(X)}$.

We will show that EMM can be viewed as an optimization problem that simultaneously considers enforcing the structure and the maximiztion of the margin. After introducing extended margin-maximizers as a suitable extension of the max-margin classifier in the structured setting, in Chapter 9 we also analyze the asymptotic behavior of the EMM optimizer when the data points are derived independently from a Gaussian distribution. It will be shown in our theoretical results as well as numerical simulations that an appropriate choice of the potential function $\psi(\cdot)$ can lead to a classifier which outperforms the max-margin classifier.

Before proceeding into the analysis of the performance of the optimal classifiers in Chapters 8 and 9, we need to answer an important question. **What is the condition for the separability of the training data?** In the remaining of this chapter, we focus on answering this question.

7.3 Separability Condition

In this section, we study the necessary and sufficient conditions for the separability of the data set, $\mathcal{D}$. Here, we assume that the data points, $\{\mathbf{x}_i\}_{i=1}^n$, are drawn independently from the standard normal distribution $\mathcal{N}(\mathbf{0}, \frac{1}{p}\mathbf{I}_p)$. In addition, the labels are generated from a logistic-type model, with the underlying parameter, $\mathbf{w}^\star \in \mathbb{R}^p$, as follows,

$$y_i = \mathrm{RAD}\big(\Phi(\mathbf{x}_i^T \mathbf{w}^\star)\big) \ , \quad i = 1, 2, \ldots, n, \tag{7.12}$$

where the link function $\Phi : \mathbb{R} \to [0, 1]$ is non-decreasing.

We analyze this problem in the linear asymptotic regime in which the problem dimensions, p and n, grow to infinity at a proportional rate $\delta := \frac{p}{n}$, known as the *overparameterization ratio*. Our analysis on the separability of the data set also relies on another parameter, the *signal strength* defined as $\kappa := \frac{||\mathbf{w}^\star||}{\sqrt{p}}$. Under these assumptions, we provide the conditions on the separability of the data set. Our results rely on a summary function $c_t(\cdot, \cdot)$, which incorporates the information about the underlying model.

Definition 11. *For the parameter $t > 0$, the function $c_t : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}_+$ is defined as,*

$$c_t(s, r) = \mathbb{E}\big[(1 - tsZ_1Y - rZ_2)_+^2\big], \tag{7.13}$$

where $Z_1, Z_2 \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$, and $Y \sim RAD(\Phi(tZ_1))$.

Asymptotic phase transition

Theorem 10 provides the necessary and sufficient conditions for the separability of the data.

Theorem 10 (Phase transition)**.** *Consider the data set* $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n \subset \mathbb{R}^p \times \{+1, -1\}$*, where, for* $i = 1, 2, \ldots, n$*,* $\mathbf{x}_i$ *is independently drawn from* $\mathcal{N}(\mathbf{0}, \mathbf{I}/p)$ *and* y_i *is derived from* (7.12)*. As* $n, p \to \infty$ *at a fixed overparameterization ratio* $\delta := \frac{p}{n} \in (0, \infty)$*,* $\mathcal{D}$ *is (almost surely) separable (or equivalently, the set* $\mathcal{W}$ *is nonempty) if and only if,*

$$\delta > \delta^* = \delta^*(\kappa) := \inf_{s,r \geq 0} \frac{c_\kappa(s, r)}{r^2} \,. \tag{7.14}$$

Theorem 10 indicates the necessary and sufficient conditions for the separability of the data. From earlier discussion in Section 7.2, the separability of the data set is the condition required for the existence of the EMM classifier (7.11). Therefore, (7.14) also indicates the necessary and sufficient condition for the existence of EMM. The proof of Theorem 10 is provided in Section 9.8 of Chapter 9. A few remarks are in place:

Remark 7. *The condition for the existence of the EMM classifier* (7.11) *is independent of the choice of the potential function,* $\psi(\cdot)$*.*

Remark 8. *The phase transition* (7.14) *is valid for any choice of the link function* $\Phi(\cdot)$*. This generalizes the former result by Candes and Sur [31]. It is worth noting that the summary functional* c_κ *depends on the choice of the link function. This function is often computed numerically (computing the integral corresponds to the expectation.)*

The following lemma, describes the changes in δ^* with respect to the signal strength, κ.

Lemma 17. *Assume* $\Phi(\cdot)$ *is an increasing function with* $\Phi(0) = 1/2$ *and* $\lim_{u \to -\infty} \Phi(u) = 1 - \lim_{u \to +\infty} \Phi(u) = 0$*.* δ^* *is a decreasing function of* κ*, with* $\delta^*(0) = \frac{1}{2}$ *and* $\lim_{\kappa \to \infty} \delta^*(\kappa) = 0$*.*

Note that the assumptions on the link function in Lemma 17 are satisfied by typical choices of link function, as we expect the output (the probability of the label being equal to +1) to be close to 1 when the feature vector is well-aligned with the underlying parameter vector, $\mathbf{w}^\star$, and close to 0 when the feature vector is aligned with $-\mathbf{w}^\star$.

The result of Lemma 17 can be intuitively verified. Recall that $\kappa = \frac{\|\mathbf{w}^\star\|}{\sqrt{p}}$ and $y_i \sim \text{RAD}(\Phi(\mathbf{x}_i^T \mathbf{w}^\star))$. Therefore, $\kappa \to \infty$ translates to having $y_i = \text{SIGN}(\mathbf{x}_i^T \mathbf{w}^\star)$. In this case, our training data is always separable for any number of observations n. Besides, the case of $\kappa = 0$ corresponds to having random labels assigned to feature vectors $\mathbf{x}_i$. [38] showed that in this case, as $p \to \infty$, $\delta > 0.5$ is the necessary and sufficient condition for the separability of the data set.
Figure 7.1 provides a comparison between the theoretical result in Theorem 10 and the empirical

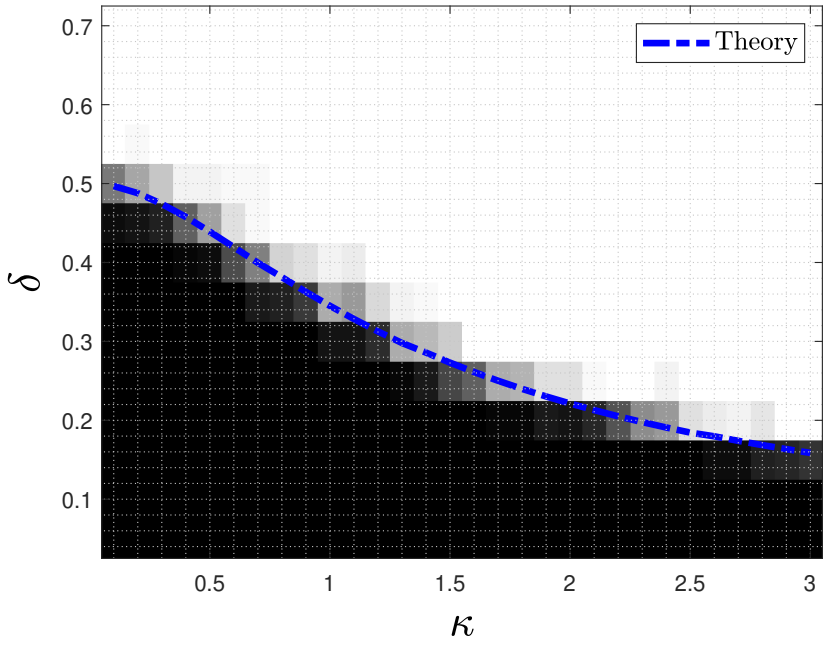


Figure 7.1: The phase transition, δ^*, for the separability of the data set, where the feature vector, $\mathbf{x}_i$ is drawn from the Gaussian distribution, $\mathcal{N}(\mathbf{0}, \frac{1}{p}\mathbf{I}_p)$, and the labels are $y_i \sim \text{RAD}\big(\Phi(\mathbf{x}_i^T \mathbf{w}^\star)\big)$, for $\Phi(z) = \frac{e^t}{e^t+e^{-t}}$. The empirical result is the average over 20 trials with $p = 150$, and the theoretical results are from Theorem 10.

results derived from numerical simulations for $p = 150$ and 20 trials. As seen in this plot, the theory matches well with the empirical simulations.

Conclusion

In this chapter, we introduced the problem of binary classification as a special example of generalized linear models which is commonly used in machine learning applications. With the goal of finding the optimal parameter of the logistic model, we formed an optimization consisting of a loss function and a regularization term. Separability of the training data set, $\mathcal{D}$, is the distinguishing factor for the behavior of the optimal solution of this optimization. Our analysis on the separability of $\mathcal{D}$ shows a sharp asymptotic phase transition with respect to the overparameterization ratio $\delta = p/n$. For data sets generated from a Gaussian distribution, we precisely characterize the phase transition which shows the necessary and sufficient condition on the separability of $\mathcal{D}$.

PRECISE PERFORMANCE ANALYSIS OF REGULARIZED LOGISTIC REGRESSION IN HIGH DIMENSIONS

[1] F. Salehi et al. "The Impact of Regularization on High-dimensional Logistic Regression". In: *Advances in Neural Information Processing Systems* (2019), pp. 11982–11992.

Logistic regression is the most commonly used statistical model for predicting dichotomous outcomes [63]. It has been extensively employed in many areas of engineering and applied sciences, such as in the medical [18, 135] and social sciences [78]. As an example, in medical studies, logistic regression can be used to predict the risk of developing a certain disease (e.g. diabetes) based on a set of observed characteristics from the patient (age, gender, weight, etc.)

Linear regression is a very useful tool for predicting a quantitative response. However, in many situations the response variable is qualitative (or categorical) and linear regression is no longer appropriate [71]. This is mainly due to the fact that least-squares regression often succeeds under the assumption that the error components are independent with normal distribution. In categorical predictions, however, the error components are neither independent nor normally distributed [96].

In logistic regression, we model the probability that the label, Y, belongs to a certain category. When no prior knowledge is available regarding the structure of the parameters, maximum likelihood is often used for fitting the model. Maximum likelihood estimation (MLE) is a special case of maximum a posteriori estimation (MAP) that assumes a uniform prior distribution on the parameters.

In many applications in statistics, machine learning, signal processing, etc., the underlying parameter obeys some sort of *structure* (sparse, group-sparse, low-rank, finite-alphabet, etc.). For instance, in modern applications where the number of features far exceeds the number of observations, one typically enforces the solution to contain only a few non-zero entries. To exploit such structural information, inspired by the Lasso [132] algorithm for linear models, researchers have studied regularization methods for generalized linear models [118, 54]. From a statistical viewpoint, adding a regularization term provides a MAP estimate with a non-uniform prior distribution that has higher densities in the set of structured solutions.

8.1 Prior Work

Classical results in logistic regression mainly concern the regime where the sample size, n, is overwhelmingly larger than the feature dimension p. It can be shown that in the limit of large samples when p is fixed and $n \to \infty$, the maximum likelihood estimator provides an efficient estimate of the underlying parameter, i.e., an unbiased estimate with the covariance matrix approaching the inverse of the Fisher information [139, 84]. However, in most modern applications in data science, the data sets often have a huge number of features, and therefore, the assumption $\frac{n}{p} \gg 1$ is not valid. Sur and Candes [31, 123, 124] have recently studied the performance of the maximum likelihood estimator for logistic regression in the regime where n is proportional to p. Their findings challenge the conventional wisdom, as they have shown that in the linear asymptotic regime, the maximum likelihood estimate is not even unbiased. Their analysis provides the precise performance of the maximum likelihood estimator.

There have been many studies in the literature on the performance of regularized (penalized) logistic regression, where a regularizer is added to the negative log-likelihood function (a partial list includes [19, 76, 138]). These studies often require the underlying parameter to be heavily structured. For example, if the parameters are sparse, the sparsity is taken to be $o(p)$. Furthermore, they provide order-wise bounds on the performance, but do not give a precise characterization of the quality of the resulting estimate. A major advantage of adding a regularization term is that it allows for recovery of the parameter vector even in regimes where the maximum likelihood estimate does not exist (due to an insufficient number of observations).

Summary of results and contributions

Here, we study regularized logistic regression (RLR) for parameter estimation in high-dimensional logistic models. Inspired by recent advances in the performance analysis of M-estimators for linear models [42, 48, 129], we precisely characterize the asymptotic performance of the RLR estimate. Our characterization is through a system of six nonlinear equations in six unknowns, through whose solution all locally-Lipschitz performance measures such as the mean, mean-squared error, probability of support recovery, etc., can be determined. In the special case when the regularization term is absent, our 6 nonlinear equations reduce to the 3 nonlinear equations reported in [123]. When the regularizer is quadratic in parameters, the 6 equations also simplify to 3. When the regularizer is the ℓ_1 norm, which corresponds to the popular sparse logistic regression [80, 81], our equations can be expressed in terms of Q-functions, and quantities such as the probability of correct support recovery can be explicitly computed. Numerous numerical simulations validate the theoretical findings across a range of problem settings. To the extent of the author's knowledge, the

result presented here is the first work that precisely characterizes the performance of the regularized logistic regression in high dimensions[1]. The result presented in this chapter is adopted from our paper [113].

8.2 Mathematical Setup

Assume we have n samples from a logistic model with parameter $\mathbf{w}^\star \in \mathbb{R}^p$. Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ denote the set of samples, where for $i = 1, 2, \ldots, n$, $\mathbf{x}_i \in \mathbb{R}^p$ is the feature vector and the label $y_i \in \{-1, +1\}$ is a symmetric Bernoulli random variable with,

$$\mathbb{P}[y_i = 1|\mathbf{x}_i] = \rho'(\mathbf{x}_i^T \mathbf{w}^\star) \ , \quad \text{for } \ i = 1, 2, \ldots, n \ , \tag{8.1}$$

where $\rho'(t) := \frac{e^{\frac{t}{2}}}{e^{\frac{t}{2}} + e^{\frac{-t}{2}}}$ is the standard logistic function. The goal is to compute an estimate for $\mathbf{w}^\star$ from the training data $\mathcal{D}$. The maximum likelihood estimator, $\hat{\mathbf{w}}_{ML}$, is defined as,

$$\begin{aligned} \hat{\mathbf{w}}_{ML} = \arg\max_{\mathbf{w}\in\mathbb{R}^p} \prod_{i=1}^n \mathbb{P}_{\mathbf{w}}(y_i|\mathbf{x}_i) &= \arg\max_{\mathbf{w}\in\mathbb{R}^p} \prod_{i=1}^n \frac{e^{\frac{y_i \mathbf{x}_i^T \mathbf{w}}{2}}}{e^{\frac{\mathbf{x}_i^T \mathbf{w}}{2}} + e^{-\frac{\mathbf{x}_i^T \mathbf{w}}{2}}} \\ &= \arg\min_{\mathbf{w}\in\mathbb{R}^p} \sum_{i=1}^n \rho(\mathbf{x}_i^T \mathbf{w}) - (\frac{1+y_i}{2})(\mathbf{x}_i^T \mathbf{w}) \ , \end{aligned} \tag{8.2}$$

where $\rho(t) := \log(1 + e^t)$ is the *link function* which has the standard logistic function as its derivative. The last optimization is simply minimization over the negative log-likelihood. This is a convex optimization program as the log-likelihood is concave with respect to $\mathbf{w}$.

As explained earlier, in many interesting settings the underlying parameter possesses certain structure(s) (sparse, low-rank, finite-alphabet, etc.). In order to exploit this structure, we assume $f : \mathbb{R}^p \to \mathbb{R}$ is a *convex* function that measures the (so-called) "complexity" of the structured solution. We fit this model by the regularized maximum (binomial) likelihood defined as follows,

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^p} \ \frac{1}{n} \cdot \big[\sum_{i=1}^n \rho(\mathbf{x}_i^T \mathbf{w}) - (\frac{1+y_i}{2})(\mathbf{x}_i^T \mathbf{w})\big] + \frac{\lambda}{p} \ f(\mathbf{w}) \ . \tag{8.3}$$

Here, $\lambda \in \mathbb{R}_+$ is the regularization parameter that must be tuned properly. In this chapter, we study the linear asymptotic regime in which the problem dimensions p, n grow to infinity at a proportional rate, $\delta := \frac{p}{n} > 0$. Our main result characterizes the performance of $\hat{\mathbf{w}}$ in terms of the ratio, δ, and the signal strength, $\kappa = \frac{||\mathbf{w}^\star||}{\sqrt{p}}$. For our analysis, we assume that the regularizer $f(\cdot)$ is separable, $f(\mathbf{w}) = \sum_i \tilde{f}(w_i)$, and the data points are drawn independently from the Gaussian distribution,

[1]The statement refers to the time of the first submission of these results in May, 2019.

$\{\mathbf{x}_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \frac{1}{p}\mathbf{I}_p)$. We further assume that the entries of $\mathbf{w}^\star$ are drawn from a distribution Π. Our main result characterizes the performance of the resulting estimator through the solution of a system of six nonlinear equations with six unknowns. In particular, we use the solution to compute some common descriptive statistics of the estimate, such as the mean and the variance.

8.3 Characterization of the Performance of Regularized Logistic Regression

In this section, we present the main results of this chapter, that is the characterization of the asymptotic performance of regularized logistic regression (RLR). When the estimation performance is measured via a locally-Lipschitz function (e.g. mean-squared error), Theorem 11 precisely predicts the asymptotic behavior of the error. The derived expression captures the role of the regularizer, $f(\cdot)$, and the particular distribution of $\mathbf{w}^\star$, through a set of scalars derived by solving a system of nonlinear equations. We first present this system of nonlinear equations along with some insights on how to numerically compute its solution. Afterwards, we formally state our result in Theorem 11. The result of this theorem will then be used to predict the general behavior of $\hat{\mathbf{w}}$. In particular, we compute its correlation with the true signal as well as its mean-squared error.

A nonlinear system of equations

As we will see in Theorem 11, given the signal strength κ and the ratio δ, the asymptotic performance of RLR is characterized by the solution to the following system of nonlinear equations with six unknowns $(\alpha, \sigma, \gamma, \theta, \tau, r)$.

$$
\begin{cases}
\kappa^2\alpha = \mathbb{E}\big[W \operatorname{Prox}_{\lambda\sigma\tau\tilde{f}(\cdot)}\big(\sigma\tau(\theta W + r\sqrt{\delta}Z)\big)\big] , \\
\gamma = \dfrac{\sqrt{\delta}}{r}\,\mathbb{E}\big[Z \operatorname{Prox}_{\lambda\sigma\tau\tilde{f}(\cdot)}\big(\sigma\tau(\theta W + r\sqrt{\delta}Z)\big)\big] , \\
\kappa^2\alpha^2 + \sigma^2 = \mathbb{E}\left[\operatorname{Prox}_{\lambda\sigma\tau\tilde{f}(\cdot)}\big(\sigma\tau(\theta W + r\sqrt{\delta}Z)\big)^2\right] , \\
\gamma^2 = \dfrac{2}{r^2}\,\mathbb{E}\big[\rho'(-\kappa Z_1)\big(\kappa\alpha Z_1 + \sigma Z_2 - \operatorname{Prox}_{\gamma\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2)\big)^2\big] , \\
\theta\gamma = -2\,\mathbb{E}\big[\rho''(-\kappa Z_1)\operatorname{Prox}_{\gamma\rho(\cdot)}\big(\kappa\alpha Z_1 + \sigma Z_2\big)\big] , \\
1 - \dfrac{\gamma}{\sigma\tau} = \mathbb{E}\Big[\dfrac{2\rho'(-\kappa Z_1)}{1 + \gamma\rho''\big(\operatorname{Prox}_{\gamma\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2)\big)}\Big] .
\end{cases}
\tag{NLS}
$$

Here Z, Z_1, Z_2 are standard normal variables, and $W \sim \Pi$, where Π denotes the distribution on the entries of $\mathbf{w}^\star$. The following remarks provide some insights on solving this nonlinear system.

Remark 9 (Proximal Operators)**.** *It is worth noting that the equations in* (NLS) *include the expectation of functionals of two proximal operators. The first three equations are in terms*

of $Prox_{\tilde{f}(\cdot)}$, which can be computed explicitly for most widely used regularizers. For instance, in ℓ_1-regularization, the proximal operator is the well-known shrinkage function defined as $\eta(x,t) := \frac{x}{|x|}(|x|-t)_+$. The remaining equations depend on computing the proximal operator of the link function $\rho(\cdot)$. For $x \in \mathbb{R}$, $Prox_{t\rho(\cdot)}(x)$ is the unique solution of $z + t\rho'(z) = x$.

Remark 10 (Numerical Evaluation)**.** *Define $\mathbf{v} := [\alpha, \sigma, \gamma, \theta, \tau, r]^T$ as the vector of unknowns. The nonlinear system (NLS) can be reformulated as $\mathbf{v} = S(\mathbf{v})$ for a properly defined $S : \mathbb{R}^6 \to \mathbb{R}^6$. We have empirically observed in our numerical simulations that a fixed-point iterative method, $\mathbf{v}_{t+1} = S(\mathbf{v}_t)$, converges to $\mathbf{v}^*$, such that $\mathbf{v}^* = S(\mathbf{v}^*)$.*

Asymptotic performance of regularized logistic regression

We are now able to present our main result. Theorem 11 below describes the average behavior of the entries of $\hat{\mathbf{w}}$, the solution of the RLR. The derived expression is in terms of the solution of the nonlinear system (NLS), denoted by $(\bar{\alpha}, \bar{\sigma}, \bar{\gamma}, \bar{\theta}, \bar{\tau}, \bar{r})$. An informal statement of our result is that as $n \to \infty$, the entries of $\hat{\mathbf{w}}$ converge as follows,

$$\hat{\mathbf{w}}_j \xrightarrow{d} \Gamma(\mathbf{w}_j^\star, Z) \ , \quad \text{for } \ j = 1, 2, \ldots, p \ , \tag{8.4}$$

where Z is a standard normal random variable, and $\Gamma : \mathbb{R}^2 \to \mathbb{R}$ is defined as,

$$\Gamma(c, d) := \text{Prox}_{\lambda\bar{\sigma}\bar{\tau}\tilde{f}(\cdot)}\big(\bar{\sigma}\bar{\tau}(\bar{\theta}c + \bar{r}\sqrt{\delta}d)\big) \ . \tag{8.5}$$

In other words, the RLR solution has the same behavior as applying the proximal operator on the "perturbed signal", i.e., the true signal added with a Gaussian noise.

Theorem 11. *Consider the optimization program* (8.3)*, where for $i = 1, 2, \ldots, n$, $\mathbf{x}_i$ has the multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}, \frac{1}{p}\mathbf{I}_p)$, and $y_i = RAD\big(\rho(\mathbf{x}_i^T\mathbf{w}^\star)\big)$, and the entries of $\mathbf{w}^\star$ are drawn independently from a distribution Π. Assume that the parameters δ, κ, and λ are such that the nonlinear system (NLS) has a unique solution $(\bar{\alpha}, \bar{\sigma}, \bar{\gamma}, \bar{\theta}, \bar{\tau}, \bar{r})$. Then, as $p \to \infty$, for any locally-Lipschitz[2] function $\Psi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, we have,*

$$\frac{1}{p}\sum_{j=1}^{p} \Psi(\hat{\mathbf{w}}_j, \mathbf{w}_j^\star) \xrightarrow{P} \mathbb{E}\big[\Psi\big(\Gamma(W, Z), W\big)\big] \ , \tag{8.6}$$

where $Z \sim \mathcal{N}(0,1)$, $W \sim \Pi$ is independent of Z, and the function $\Gamma(\cdot,\cdot)$ is defined in (8.5)*.*

[2] A function $\Psi : \mathbb{R}^d \to \mathbb{R}$ is said to be *locally-Lipschitz* if,

$$\forall M > 0, \ \exists L_M \geq 0, \text{ such that } \ \forall \mathbf{x}, \mathbf{y} \in \big[-M, +M\big]^d \ : \quad |\Psi(\mathbf{x}) - \Psi(\mathbf{y})| \leq L_M ||\mathbf{x} - \mathbf{y}|| \ .$$

We defer the detailed proof to Section 8.7. In short, to show this result we first represent the optimization as a bilinear form, $\mathbf{u}^T\mathbf{X}\mathbf{v}$, where $\mathbf{X}$ is the measurement matrix. Applying the CGMT to derive an equivalent optimization, we then simplify this optimization to obtain an unconstrained optimization with six scalar variables. The nonlinear system (NLS) represents the first-order optimality condition of the resulting scalar optimization.
Before stating the consequences of this result, a few remarks are in order.

Remark 11 (Assumptions)**.** *The assumptions in Theorem 11 are chosen in a conservative manner. In particular, we could relax the separability condition on $f(\cdot)$ to some milder condition in terms of asymptotic convergence of its proximal operator. Furthermore, one can relax the assumption on the entries of $\mathbf{w}^\star$ being i.i.d. to a weaker assumption on the empirical distribution of its entries. However, for the applications discussed in this chapter, the theorem in its current form is adequate. In Chapter 9, we will perform a more general analysis with less restrictive assumptions.*

Remark 12 (Choosing Ψ)**.** *The performance measure in Theorem 11 is computed in terms of evaluation of a locally-Lipschitz function, $\Psi(\cdot,\cdot)$. As an example, $\Psi(u,v) = (u-v)^2$ can be used to compute the mean-squared error. In the next section, we will appeal to this theorem with various choices of Ψ to evaluate different performance measures on $\hat{\mathbf{w}}$.*

Correlation and variance of the RLR estimate

As the first application of Theorem 11, we compute common descriptive statistics of the estimate $\hat{\mathbf{w}}$. In the following corollaries, we establish that the parameters $\bar{\alpha}$, and $\bar{\sigma}$ in (NLS), respectively, correspond to the correlation and the mean-squared error of the resulting estimate.

Corollary 6. *As $p \to \infty$, $\frac{1}{||\mathbf{w}^\star||^2}\,\hat{\mathbf{w}}^T\mathbf{w}^\star \xrightarrow{P} \bar{\alpha}$.*

Proof. Recall that $||\mathbf{w}^\star||^2 = p\kappa^2$. Applying Theorem 11 with $\Psi(u,v) = uv$ gives,

$$\frac{1}{||\mathbf{w}^\star||^2}\hat{\mathbf{w}}^T\mathbf{w}^\star = \frac{1}{\kappa^2 p}\sum_{j=1}^{p}\hat{\mathbf{w}}_j\mathbf{w}_j^\star \xrightarrow{P} \frac{1}{\kappa^2}\mathbb{E}\big[W\ \mathrm{Prox}_{\lambda\bar{\sigma}\bar{\tau}\tilde{f}(\cdot)}\big(\bar{\sigma}\bar{\tau}(\bar{\theta}W + \bar{r}\sqrt{\delta}Z)\big)\big] = \bar{\alpha}\ , \tag{8.7}$$

where the last equality is derived from the first equation in the nonlinear system (NLS), along with the fact that $(\bar{\alpha},\bar{\sigma},\bar{\gamma},\bar{\theta},\bar{\tau},\bar{r})$ is the solution to this nonlinear system. □

Corollary 6 states that upon centering $\hat{\mathbf{w}}$ around $\bar{\alpha}\mathbf{w}^\star$, it becomes uncorrelated from $\mathbf{w}^\star$. Therefore, we define $\tilde{\mathbf{w}} := \frac{\hat{\mathbf{w}}}{\bar{\alpha}}$ to be an unbiased estimator of $\mathbf{w}^\star$. The following corollary computes the mean-squared error of $\tilde{\mathbf{w}}$.

Corollary 7. *As* $p \to \infty$, $\frac{1}{p}||\tilde{\mathbf{w}} - \mathbf{w}^\star||^2 \xrightarrow{P} \frac{\bar{\sigma}^2}{\bar{\alpha}^2}$.

Proof. We appeal to Theorem 11 with $\Psi(u, v) = (u - \bar{\alpha}v)^2$,

$$\frac{1}{p}||\tilde{\mathbf{w}} - \mathbf{w}^\star||^2 = \frac{1}{\bar{\alpha}^2}\big(\frac{1}{p}||\hat{\mathbf{w}} - \bar{\alpha}\mathbf{w}^\star||^2\big) \xrightarrow{P} \frac{1}{\bar{\alpha}^2}\mathbb{E}\big[\big(\mathrm{Prox}_{\lambda\bar{\sigma}\bar{\tau}\tilde{f}(\cdot)}\big(\bar{\sigma}\bar{\tau}(\bar{\theta}W + \bar{r}\sqrt{\delta}Z)\big) - \bar{\alpha}W\big)^2\big] = \frac{\bar{\sigma}^2}{\bar{\alpha}^2}\ , \tag{8.8}$$

where the last equality is derived from the third equation in the nonlinear system (NLS) together with the result of Corollary 6. □

In the next two sections, we investigate other properties of the estimate $\hat{\mathbf{w}}$ under ℓ_1 and ℓ_2^2 regularization.

8.4 Impact of ℓ_2 regularization on Logistic Regression

The ℓ_2 norm regularization is commonly used in machine learning applications to stabilize the model. Adding this regularization would simply shrink all the parameters toward the origin and hence decrease the variance of the resulting model. Here, we provide a precise performance analysis of the RLR with ℓ_2^2-regularization, i.e.,

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{n} \cdot \big[\sum_{i=1}^{n} \rho(\mathbf{x}_i^T\mathbf{w}) - (\frac{1+y_i}{2})(\mathbf{x}_i^T\mathbf{w})\big] + \frac{\lambda}{2p}\sum_{i=1}^{p}\mathbf{w}_i^2\ . \tag{8.9}$$

To analyze (8.9), we use the result of Theorem 11. It can be shown that in the nonlinear system (NLS), $\bar{\theta}$, $\bar{\tau}$, $\bar{r}$ can be derived explicitly from solving the first three equations. This is due to the fact that the proximal operator of $\tilde{f}(\cdot) = \frac{1}{2}(\cdot)^2$ can be expressed in the following closed-form,

$$\mathrm{Prox}_{t\tilde{f}(\cdot)}(x) = \arg\min_{y\in\mathbb{R}} \frac{1}{2t}(y-x)^2 + \frac{1}{2}y^2 = \frac{x}{1+t}\ . \tag{8.10}$$

This indicates that the proximal operator in this case is just a simple scaling. Substituting (8.10) in the nonlinear system (NLS), we can rewrite the first three equations as follows,

$$\begin{cases} \theta = \dfrac{\alpha\delta}{\gamma}\ , \\ \tau = \dfrac{\gamma}{\sigma\big(\delta - \lambda\gamma\big)}\ , \\ r = \dfrac{\sigma\sqrt{\delta}}{\gamma}\ . \end{cases} \tag{8.11}$$

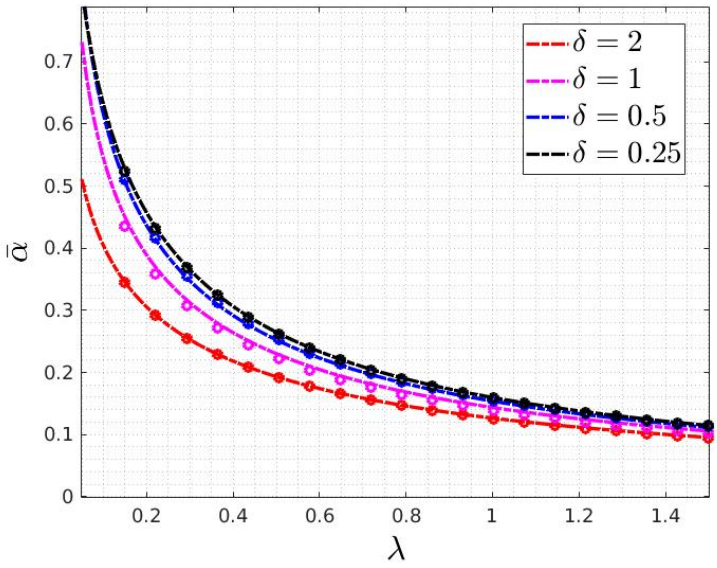


Figure 8.1: The correlation factor ($\bar{\alpha}$) of the solution of logistic regression with ℓ_2^2 penalty.

Using this simplification, we can state the following Theorem which gives the performance of logistic regression under ℓ_2^2-regularization:

Theorem 12. *Consider the optimization* (8.9) *with parameters κ, δ, and γ, and the same assumptions as in Theorem 11. As $p \to \infty$, for any locally-Lipschitz function $\Psi(\cdot,\cdot)$, the following convergence holds,*

$$\frac{1}{p}\sum_{j=1}^{p} \Psi(\hat{\mathbf{w}}_j - \bar{\alpha}\mathbf{w}_j^\star, \mathbf{w}_j^\star) \xrightarrow{P} \mathbb{E}\big[\Psi(\bar{\sigma}Z, W)\big] \,, \tag{8.12}$$

where Z is standard normal, $W \sim \Pi$ is independent of Z, and $\bar{\alpha}$, $\bar{\sigma}$ are the unique solution to the following nonlinear system of equations,

$$\begin{cases} \dfrac{\delta\sigma^2}{2} = \mathbb{E}\big[\rho'(-\kappa Z_1)\big(\kappa\alpha Z_1 + \sigma Z_2 - Prox_{\gamma\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2)\big)^2\big] \,, \\ -\dfrac{\delta\alpha}{2} = \mathbb{E}\big[\rho''(-\kappa Z_1) Prox_{\gamma\rho(\cdot)}\big(\kappa\alpha Z_1 + \sigma Z_2\big)\big] \,, \\ 1 - \delta + \lambda\gamma = \mathbb{E}\big[\dfrac{2\rho'(-\kappa Z_1)}{1 + \gamma\rho''\big(Prox_{\gamma\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2)\big)}\big] \,. \end{cases} \tag{NLS - L_2}$$

The proof is provided in Section 8.8. Theorem 12 states that upon centering the estimate $\hat{\mathbf{w}}$, it becomes uncorrelated from $\mathbf{w}^\star$ and the distribution of the entries approach a zero-mean Gaussian distribution with variance $\bar{\sigma}^2$.

Figures 8.1, 8.2, and 8.3 depict the performance of the regularized estimate for different values of λ.

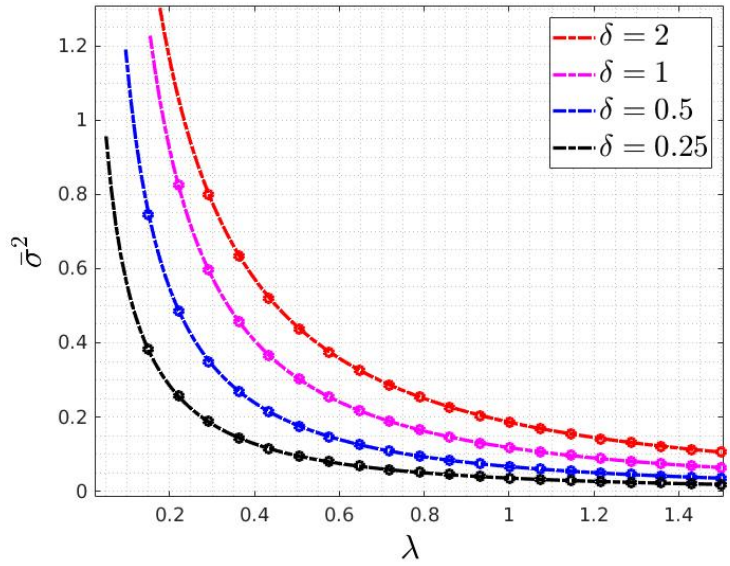


Figure 8.2: The variance $\bar{\sigma}^2$ of the solution of logistic regression with ℓ_2^2 penalty.

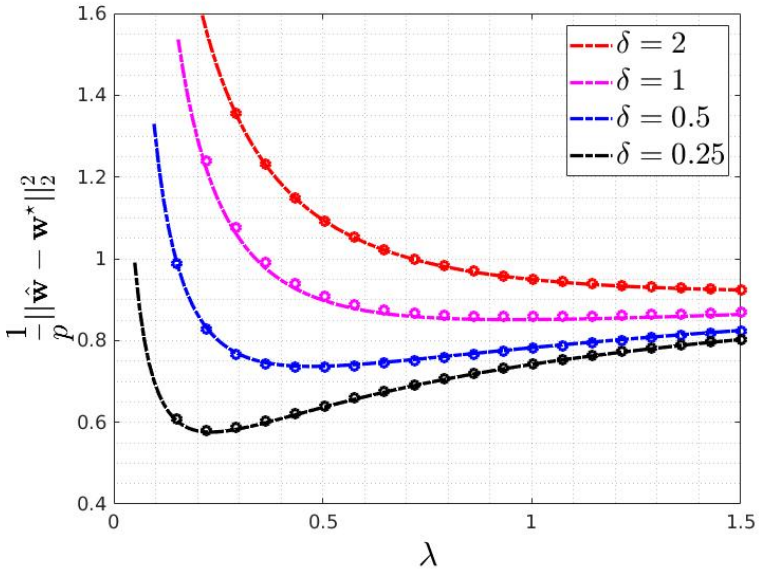


Figure 8.3: The mean-squared error $\frac{1}{p}\|\hat{\mathbf{w}} - \mathbf{w}^\star\|^2$ of the solution of logistic regression with ℓ_2^2 penalty.

The dashed lines depict the theoretical results derived from Theorem 12, and the dots are the results of empirical simulations. The empirical results are the average over 100 independent trials with $p = 250$ and $\kappa = 1$. Although our theoretical results are asymptotic, we observe that the theory well matches the empirical results in our numerical simulations.

As observed in these figures, increasing the value of λ reduces the correlation factor $\bar{\alpha}$ (Figure 8.1) and the variance $\bar{\sigma}^2$ (Figure 8.2). Figure 8.3 shows the mean-squared-error of the estimate as a function of λ . It indicates that for different values of δ, there exists an optimal value λ_{opt} that achieves the minimum mean-squared error.

Unstructured case

By setting $\lambda = 0$ in (8.9), we obtain the optimization with no regularization, i.e., the maximum likelihood estimate. When we set λ to zero in $(NLS - L_2)$, Theorem 12 gives the same result as Sur and Candes reported in [123]. In their analysis, they have also provided an interesting interpretation of $\bar{\gamma}$ in terms of the likelihood ratio statistics.

8.5 Sparse Logistic Regression

In this section, we study the performance of our estimate when the regularizer is the ℓ_1 norm. In modern machine learning applications, the number of features, p, is often overwhelmingly large. Therefore, to avoid overfitting, one typically needs to perform feature selection, that is to exclude irrelevant variables from the regression model [71]. Adding an ℓ_1 penalty to the loss function is the most popular approach for feature selection.

As a natural consequence of the result of Theorem 11, we study the performance of RLR with ℓ_1 regularizer (referred to as "sparse LR") and evaluate its success in recovery of the sparse signals. Here, we extend our general analysis to the case of sparse LR. In other words, we will precisely analyze the performance of the solution of the following optimization,

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^p} \ \frac{1}{n} \cdot \Big[\sum_{i=1}^{n} \rho(\mathbf{x}_i^T\mathbf{w}) - (\frac{1+y_i}{2})(\mathbf{x}_i^T\mathbf{w})\Big] + \frac{\lambda}{p}||\mathbf{w}||_1 \ . \tag{8.13}$$

In what follows, we first explicitly describe the expectations in the nonlinear system (NLS) using two Q-functions[3]. Consequently, we analyze the support recovery in the resulting estimate and show that the two Q-functions represent the probability of on and off support recovery.

[3]The Q-function is the tail distribution of the standard normal random variable defined as $Q(t) := \int_t^\infty \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx$.

Convergence behavior of sparse LR

For our analysis in this section, we assume that each entry $\mathbf{w}_i^\star$, for $i = 1, \ldots, p$, is sampled i.i.d. from a distribution,

$$\Pi(W) = (1-s)\cdot\delta_0(W) + s\cdot\Big(\frac{\phi(\frac{W}{\frac{\kappa}{\sqrt{s}}})}{\frac{\kappa}{\sqrt{s}}}\Big), \tag{8.14}$$

where $s \in (0,1)$ is the *sparsity factor*, $\phi(t) := \frac{e^{-t^2/2}}{\sqrt{2\pi}}$ is the density of the standard normal distribution, and $\delta_0(\cdot)$ is the Dirac delta function. In other words, entries of $\mathbf{w}^\star$ are zero with probability $1-s$, and the non-zero entries have a Gaussian distribution with appropriately defined variance. Although our analysis can be extended further, here we only present the result for a Gaussian distribution on the non-zero entries. The proximal operator of $\tilde{f}(\cdot) = |\cdot|$ is the soft-thresholding operator defined as $\eta(x,t) = \frac{x}{|x|}(x-t)_+$. Therefore, we are able to explicitly compute the expectations with respect to $\tilde{f}(\cdot)$ in the nonlinear system (NLS). To streamline the representation, we introduce the following two proxies,

$$t_1 = \frac{\lambda}{\sqrt{r^2\delta + \frac{\theta^2\kappa^2}{s}}}\ ,\quad t_2 = \frac{\lambda\sqrt{\delta}}{r}\ . \tag{8.15}$$

In the next section, we provide an interpretation for t_1 and t_2. In particular, we will show that $Q(\bar{t_1})$, and $Q(\bar{t_2})$ are related to the probabilities of on and off support recovery, which would allow us to compute the type I and type II errors in support recovery. Considering the distribution Π (in 8.14) for the entries of $\mathbf{w}^\star$, we can rewrite the first three equations (which are in terms of the proximal operator of $|\cdot|$) in (NLS) as follows,

$$\begin{cases} \dfrac{\alpha}{2\sigma\tau} = \theta\cdot Q(t_1)\ , \\ \dfrac{\gamma}{2\delta\sigma\tau} = s\cdot Q\big(t_1\big) + (1-s)\cdot Q\big(t_2\big)\ , \\ \dfrac{\kappa^2\alpha^2+\sigma^2}{2\sigma^2\tau^2} = \dfrac{\gamma\lambda^2}{2\delta\sigma\tau} + \dfrac{\gamma r^2}{2\sigma\tau} + \kappa^2\theta^2\cdot Q\big(t_1\big) - \lambda^2(s\cdot\dfrac{\phi(t_1)}{t_1} + (1-s)\cdot\dfrac{\phi(t_2)}{t_2})\ . \end{cases} \tag{8.16}$$

Appending the three equations in (8.16) to the last three equations in (NLS) gives the nonlinear system for sparse logistic regression. Upon solving these system of nonlinear equations, we can use the result of Theorem 11 to compute various performance measures on the estimate $\hat{\mathbf{w}}$.

Figures 8.4, 8.5, and 8.6 show the performance of our estimate as a function of λ. The dashed lines depict the theoretical results derived from Theorem 11, and the dots are the results of empirical simulations. The empirical results are the average over 100 independent trials with $p = 250$ and

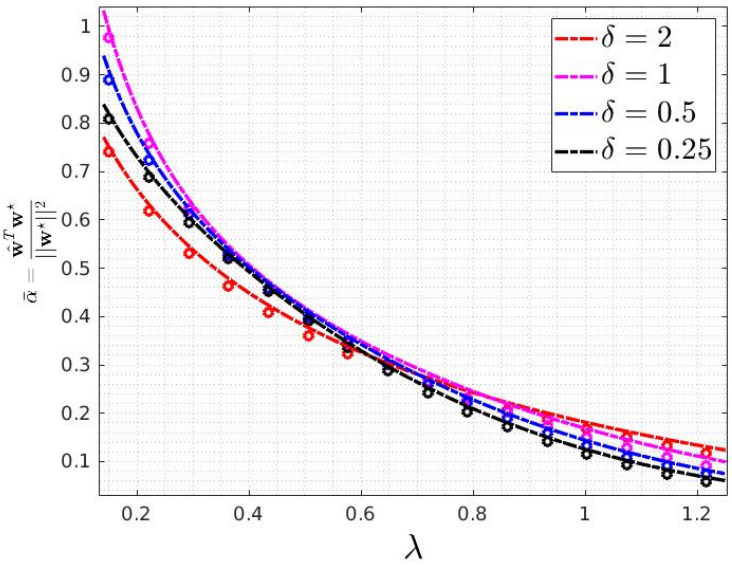


Figure 8.4: The correlation factor ($\bar{\alpha}$) of the solution of logistic regression with ℓ_1 penalty.

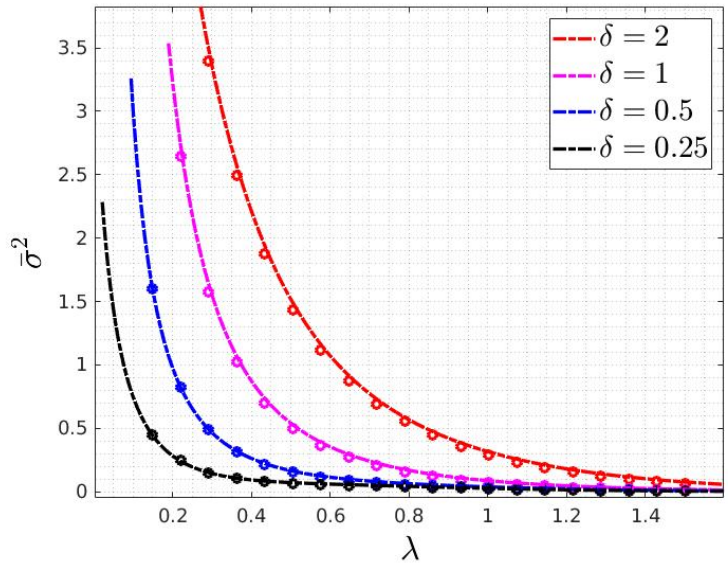


Figure 8.5: The variance $\bar{\sigma}^2$ of the solution of logistic regression with ℓ_1 penalty.

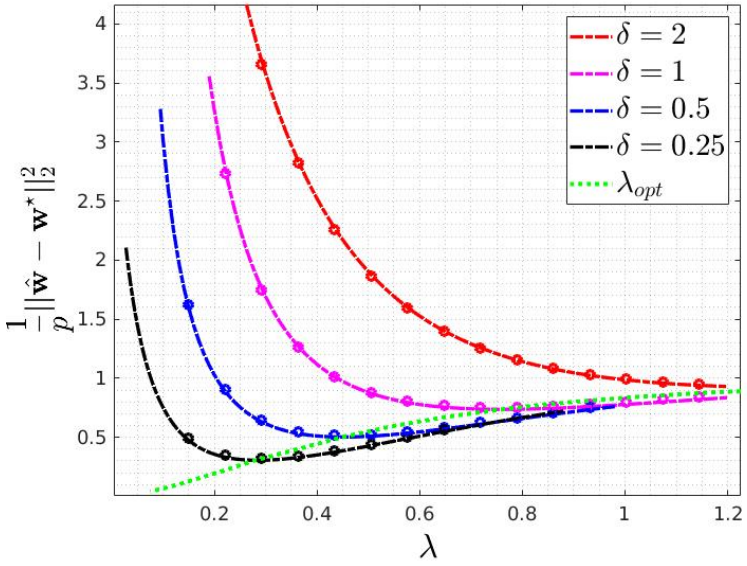


Figure 8.6: The mean-squared error $\frac{1}{p}\|\hat{\mathbf{w}} - \mathbf{w}^\star\|^2$ of the solution of logistic regression with ℓ_1 penalty.

$\kappa = 1$. It can be seen that the bound derived from our theoretical result matches the empirical simulations. Also, it can be inferred from Figure 8.6 that the optimal value of λ (λ_{opt} that achieves the minimum mean-squared error) is a decreasing function of δ.

Support recovery

We next study the support recovery in sparse logistic regression. As mentioned earlier, sparse LR is often used when the underlying parameter has few non-zero entries. We define the support of $\mathbf{w}^\star$ as $\Omega := \{j | 1 \leq j \leq p, \mathbf{w}_j^\star \neq 0\}$. Here, we would like to compute the probability of success in recovery of the support of $\mathbf{w}^\star$.

Let $\hat{\mathbf{w}}$ denote the solution of the optimization (8.13). We fix the value $\epsilon > 0$ as a hard-threshold based on which we decide whether an entry is on the support or not. In other words, we form the following set as our estimate of the support given $\hat{\mathbf{w}}$,

$$\hat{\Omega} = \{j | 1 \leq j \leq p, |\hat{\mathbf{w}}_j| > \epsilon\} \ . \tag{8.17}$$

In order to evaluate the success in support recovery, we define the following two error measures,

$$E_1(\epsilon) = \text{Prob}\{j \in \hat{\Omega} | j \notin \Omega\} \ \ , \ \ E_2(\epsilon) = \text{Prob}\{j \notin \hat{\Omega} | j \in \Omega\} \ . \tag{8.18}$$

In our estimation, E_1 represents the probability of false alarm, and E_2 is the probability of missed-detection of an entry of the support. The following lemma indicates the asymptotic behavior of both

errors as ϵ approaches zero.

Lemma 18 (Support Recovery)**.** *Let* $\hat{\mathbf{w}}$ *be the solution to the optimization* (8.13)*, and the entries of* $\mathbf{w}^\star$ *have distribution* Π *defined in* (8.14)*. Assume* λ *is chosen such that the nonlinear system* (NLS) *has a unique solution* $(\bar{\alpha}, \bar{\sigma}, \bar{\gamma}, \bar{\theta}, \bar{\tau}, \bar{r})$*. As* $p \to \infty$*, we have,*

$$\begin{aligned} \lim_{\epsilon \downarrow 0} E_1(\epsilon) &\xrightarrow{p} 2\,Q\big(\bar{t}_1\big) \;\; \textit{where,} \;\; \bar{t}_1 = \frac{\lambda}{\bar{r}\sqrt{\delta}}, \;\; \textit{and,} \\ \lim_{\epsilon \downarrow 0} E_2(\epsilon) &\xrightarrow{p} 1 - 2\,Q\big(\bar{t}_2\big) \;\; \textit{where,} \;\; \bar{t}_2 = \frac{\lambda}{\sqrt{\delta \bar{r}^2 + \frac{\bar{\theta}^2 \kappa^2}{s}}}. \end{aligned} \tag{8.19}$$

Figures 8.7 and 8.8 depict the performance of the ℓ_1-regularized logistic regression in finding the support of the underlying signal. The dashed lines are the theoretical results derived from Lemma 18, and the dots are the results of empirical simulations. For the numerical simulations, the result is the average over 100 independent trials with $p = 250$ and $\kappa = 1$ and $\epsilon = 0.001$.

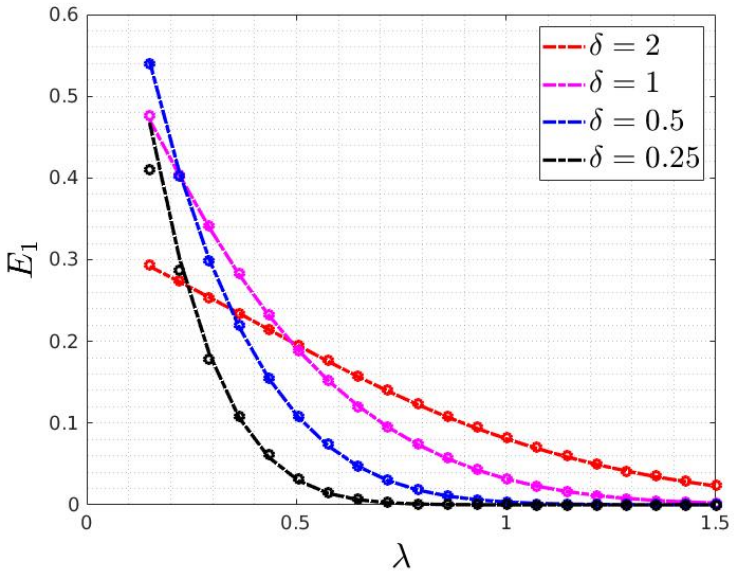


Figure 8.7: The support recovery in the regularized logistic regression with ℓ_1 penalty for E_1: the probability of false detection.

8.6 Conclusion and Future Directions

In this chapter, we analyzed the performance of the regularized logistic regression (RLR), which is often used for parameter estimation in binary classification. We considered the setting where the underlying parameter has a certain structure (e.g. sparse, group-sparse, low-rank, etc.) that can be

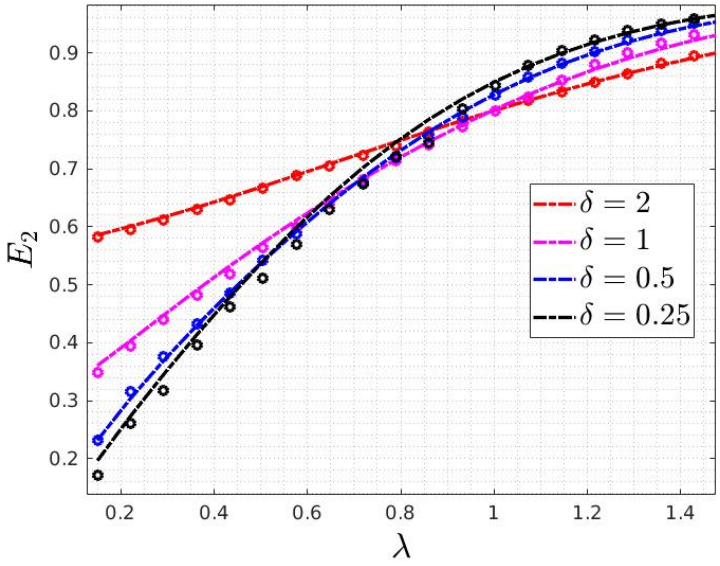


Figure 8.8: The support recovery in the regularized logistic regression with ℓ_1 penalty for E_2: the probability of missing an entry of the support.

enforced via a convex penalty function $f(\cdot)$. As mentioned earlier in Section 8.1, an advantage of RLR is that it allows parameter recovery even for instances where the (unconstrained) maximum likelihood estimate does not exist. We precisely characterized the performance of the regularized maximum likelihood estimator via the solution to a nonlinear system of equations. Our main results can be used to measure the performance of RLR for a general convex penalty function $f(\cdot)$. In particular, we apply our findings to two important special cases, i.e., ℓ_2^2-RLR and ℓ_1-RLR. When the regularizer is quadratic in parameters, we have shown that the nonlinear system can be simplified to three equations. By setting the regularization parameter, λ, to zero, which corresponds to the maximum likelihood estimator, we simply derived the results reported by Sur and Candes [123]. For sparse logistic regression, we established that the nonlinear system can be represented using two Q-functions. We further showed that these two Q-functions represent the probability of the support recovery.

For our analysis, we assumed that the data points are drawn independently from a Gaussian distribution and utilized the CGMT framework. An interesting future work is to extend our analysis to non-Gaussian distributions. To this end, we can exploit the techniques that have been used to establish the universality law (see [100, 103, 2] and the references therein).

8.7 Proof of Theorem 11

We present the proof of our main result that is a precise characterization on the performance of the optimization program (8.3) in the limit where $p, n \to \infty$ at a fixed ratio $\delta := \frac{p}{n}$. We assume that the data points are drawn independently from Gaussian distribution, $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{1}{p}\mathbf{I}_p)$. For simplicity in notations, we replace y_i with $\frac{(1+y_i)}{2}$ which results in the labels being in $\{0, 1\}$. Therefore, we can rewrite (8.3) as follows,

$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{n}\mathbf{1}^T\rho(\frac{1}{\sqrt{p}}\mathbf{H}\mathbf{w}) - \frac{1}{n\sqrt{p}}\mathbf{y}^T\mathbf{H}\mathbf{w} + \frac{\lambda}{p}f(\mathbf{w}) \tag{8.20}$$

where the action of function $\rho(\cdot)$ on a vector is considered component-wise, $\mathbf{y} \in \mathbb{R}^n$ and $\mathbf{H} \in \mathbb{R}^{n\times p}$ are defined as follows,

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \ , \ \mathbf{H} = \sqrt{p} \cdot \begin{bmatrix} -\mathbf{x}_1^T- \\ -\mathbf{x}_2^T- \\ \vdots \\ -\mathbf{x}_n^T- \end{bmatrix} \ . \tag{8.21}$$

Note that the matrix $\mathbf{H}$ is defined in such a way that its entries have i.i.d. standard normal distribution. We use the CGMT framework for our analysis. The proof strategy consists of three main steps:

1. Finding the auxiliary optimization: In order to apply the result of Theorem 15, we need to rewrite the optimization as a bilinear form and find its associated auxiliary optimization.
2. Analyzing the auxiliary optimization: The goal of this step is to simplify the auxiliary optimization in such a way that its performance can be characterized via a scalar optimization.
3. Finding the optimality condition on the scalar optimization: We investigate the solution to the resulting scalar optimization. Specifically, by writing the first-order optimality conditions, we will derive the nonlinear system of equations (NLS).

We explain each of the three steps in more details in the following subsections.

Finding the auxiliary optimization

In order to apply the CGMT, we need to have a min-max optimization. Introducing a new variable $\mathbf{u}$, we have the following optimization,

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^p,\ \mathbf{u}\in\mathbb{R}^n} \ & \frac{1}{n}\mathbf{1}^T\rho(\mathbf{u}) - \frac{1}{n}\mathbf{y}^T\mathbf{u} + \frac{\lambda}{p}f(\mathbf{w}) \\ \text{s.t.} \ & \mathbf{u} = \frac{1}{\sqrt{p}}\mathbf{H}\mathbf{w} \ . \end{aligned} \tag{8.22}$$

Next, we use the Lagrange multiplier $\mathbf{v}$ to rewrite (8.22) as a min-max optimization,

$$\min_{\mathbf{w}\in\mathbb{R}^p,\mathbf{u}\in\mathbb{R}^n}\max_{\mathbf{v}\in\mathbb{R}^n}\ \frac{1}{n}\mathbf{1}^T\rho(\mathbf{u})-\frac{1}{n}\mathbf{y}^T\mathbf{u}+\frac{\lambda}{p}f(\mathbf{w})+\frac{1}{n}\mathbf{v}^T(\mathbf{u}-\frac{1}{\sqrt{p}}\mathbf{H}\mathbf{w})\ . \tag{8.23}$$

Since $\mathbf{y}$ depends on $\mathbf{H}$, we cannot directly apply CGMT to the bilinear form $\mathbf{v}^T\mathbf{H}\mathbf{w}$. To solve this issue, we first introduce $\mathbf{P}:=\frac{1}{||\mathbf{w}^\star||_2^2}\mathbf{w}^\star\mathbf{w}^{\star T}$ and $\mathbf{P}^\perp:=\mathbf{I}_p-\mathbf{P}$, the projection matrices on the direction of $\mathbf{w}^\star$ and its orthogonal complement, respectively. We use these projections to decompose the matrix $\mathbf{H}$ as, $\mathbf{H}=\mathbf{H}_1+\mathbf{H}_2$, with $\mathbf{H}_1:=\mathbf{H}\times\mathbf{P}$, and $\mathbf{H}_2:=\mathbf{H}\times\mathbf{P}^\perp$. Rewriting (8.23) with the decomposition of $\mathbf{H}$ would give,

$$\min_{\mathbf{w}\in\mathbb{R}^p,\mathbf{u}\in\mathbb{R}^n}\max_{\mathbf{v}\in\mathbb{R}^n}\ \frac{1}{n}\mathbf{1}^T\rho(\mathbf{u})-\frac{1}{n}\mathbf{y}^T\mathbf{u}+\frac{\lambda}{p}f(\mathbf{w})+\frac{1}{n}\mathbf{v}^T(\mathbf{u}-\frac{1}{\sqrt{p}}\mathbf{H}_1\mathbf{w})-\frac{1}{n\sqrt{p}}\mathbf{v}^T\mathbf{H}_2\mathbf{w}\ . \tag{8.24}$$

It is worth noting that after performing this decomposition, the label vector ($\mathbf{y}$) would be independent of $\mathbf{H}_2$ since,

$$\mathbf{y}=Ber\big(\rho'(\frac{1}{\sqrt{p}}\mathbf{H}\mathbf{w}^\star)\big)=Ber\big(\rho'(\frac{1}{\sqrt{p}}\mathbf{H}\mathbf{P}\mathbf{w}^\star)\big)=Ber\big(\rho'(\frac{1}{\sqrt{p}}\mathbf{H}_1\mathbf{w}^\star)\big)\ , \tag{8.25}$$

where we used $\mathbf{P}\mathbf{w}^\star=\mathbf{w}^\star$. Exploiting this fact, one can check that all the additive terms in the objective function of (8.24) except the last one are independent of $\mathbf{H}_2$. Also, the objective function is convex with respect to $\mathbf{w}$ and $\mathbf{u}$ and concave with respect to $\mathbf{v}$. In order to apply the CGMT framework, we only need an extra condition which is restricting the feasible sets of $\mathbf{w}$, $\mathbf{u}$, and $\mathbf{v}$ to be compact and convex. We can introduce some artificial convex and bounded sets $\mathcal{S}_\mathbf{u}$, $S_\mathbf{v}$, and $\mathcal{S}_\mathbf{w}$, and perform the optimization over these sets. Note that these sets can be chosen large enough such that they do not affect the optimization itself. For simplicity, in our arguments here we ignore the condition on the compactness of the feasible sets and apply the CGMT whenever our feasible sets are convex.

The optimization program (8.24) is suitable to be analyzed via the CGMT as the conditions are all satisfied. Having identified (8.24) as the (PO) in our optimization, it is straightforward to write its corresponding (AO) as equation (A.1) as explained in Appendix A.1. Therefore, the Auxiliary Optimization (AO) can be written as follows,

$$\begin{aligned}\min_{\mathbf{w}\in\mathbb{R}^p,\mathbf{u}\in\mathbb{R}^n}\max_{\mathbf{v}\in\mathbb{R}^n}\ &\frac{1}{n}\mathbf{1}^T\rho(\mathbf{u})-\frac{1}{n}\mathbf{y}^T\mathbf{u}+\frac{\lambda}{p}f(\mathbf{w})+\frac{1}{n}\mathbf{v}^T(\mathbf{u}-\frac{1}{\sqrt{p}}\mathbf{H}_1\mathbf{w})\\&-\frac{1}{n\sqrt{p}}(\mathbf{v}^T\mathbf{h}||\mathbf{P}^\perp\mathbf{w}||+||\mathbf{v}||\mathbf{g}^T\mathbf{P}^\perp\mathbf{w})\ ,\end{aligned} \tag{8.26}$$

where $\mathbf{h}\in\mathbb{R}^n$ and $\mathbf{g}\in\mathbb{R}^p$ have i.i.d. standard normal entries. Next, we need to analyze the optimization (8.26) to characterize its performance.

Analyzing the auxiliary optimization

In this section, we analyze the auxiliary optimization (8.26). Ideally, we would like to solve the optimizations with respect to the direction of the vectors, in order to finally get a scalar-valued optimization over the magnitude of the variables.

Proceeding onward, we first perform the maximization with respect to the direction of $\mathbf{v}$. We can write the following maximization with respect to $\mathbf{v}$,

$$\max_{\mathbf{v}\in\mathbb{R}^n} \frac{1}{n\sqrt{p}}||\mathbf{v}||\mathbf{g}^T\mathbf{P}^{\perp}\mathbf{w} + \frac{1}{n}\mathbf{v}^T\big(\mathbf{u} - \frac{1}{\sqrt{p}}\mathbf{H}_1\mathbf{w} - \frac{||\mathbf{P}^{\perp}\mathbf{w}||}{\sqrt{p}}\mathbf{h}\big) \ . \tag{8.27}$$

In order to maximize the objective function, $\mathbf{v}$ chooses its direction to be the same as the vector it is multiplied to. Define $r := \|\mathbf{v}\|/\sqrt{n}$, then maximizing over the direction of $\mathbf{v}$ would give,

$$\max_{r\geq 0} \quad r\big(\frac{1}{\sqrt{np}}\mathbf{g}^T\mathbf{P}^{\perp}\mathbf{w} + ||\frac{1}{\sqrt{n}}\mathbf{u} - \frac{1}{\sqrt{np}}\mathbf{H}_1\mathbf{w} - \frac{||\mathbf{P}^{\perp}\mathbf{w}||}{\sqrt{np}}\mathbf{h}||\big) \ . \tag{8.28}$$

Replacing this in (8.26), we would have,

$$\begin{aligned}\min_{\mathbf{w}\in\mathbb{R}^p,\mathbf{u}\in\mathbb{R}^n} \max_{r\geq 0} \ \frac{1}{n}\mathbf{1}^T\rho(\mathbf{u}) - \frac{1}{n}\mathbf{y}^T\mathbf{u} + \frac{\lambda}{p}f(\mathbf{w}) + r\frac{1}{\sqrt{np}}\mathbf{g}^T\mathbf{P}^{\perp}\mathbf{w} \\ + r||\frac{1}{\sqrt{n}}\mathbf{u} - \frac{1}{\sqrt{np}}\mathbf{H}\ (\mathbf{P}\mathbf{w}) - \frac{||\mathbf{P}^{\perp}\mathbf{w}||}{\sqrt{np}}\mathbf{h}|| \ ,\end{aligned} \tag{8.29}$$

where we replaced $\mathbf{H}_1$ with $\mathbf{H}\times\mathbf{P}$. Next, we would like to solve the minimization with respect to $\mathbf{w}$.

Before continuing our analysis, we need to discuss an important point that would help us in the remaining of this section. It will be observed that in order to simplify the optimization, we would like to flip the orders of min and max in the (AO) optimization. Since the objective function in the optimization (8.29) is not convex-concave, we cannot appeal to the Sion's min-max theorem in order to flip min and max. However, it has been shown in [129] (see Appendix A) that flipping the order min and max in the (AO) is allowed in the asymptotic setting. This is mainly due to the fact that the original (PO) optimization was convex-concave with respect to its variables, and as the CGMT suggests, (AO) and (PO) are tightly related in the asymptotic setting; hence, flipping the order of optimizations in (AO) is justified whenever such a flipping is allowed in the (PO). We appeal to this result to flip the orders of min and max when needed.

The goal is to express the final result in terms of the *expected Moreau envelope* of the regularization function, $f(\cdot)$ and the link function, $\rho(\cdot)$. Finding the optimal direction of $\mathbf{w}$ is cumbersome due to the existence of the term $\lambda f(\mathbf{w})$ in the objective. So, we introduce new variables $\boldsymbol{\mu}, \boldsymbol{\beta} \in \mathbb{R}^p$ and rewrite the optimization as follows,

$$\min_{\substack{\mathbf{w}\in\mathbb{R}^p,\mathbf{u}\in\mathbb{R}^n\\ \boldsymbol{\mu}\in\mathbb{R}^p}} \max_{\substack{\boldsymbol{\beta}\in\mathbb{R}^p\\ r\geq 0}} \frac{1}{n}\mathbf{1}^T\rho(\mathbf{u}) - \frac{1}{n}\mathbf{y}^T\mathbf{u} + \frac{\lambda}{p}f(\boldsymbol{\mu}) + r\frac{1}{\sqrt{np}}\mathbf{g}^T\mathbf{P}^{\perp}\mathbf{w}$$
$$+r||\frac{1}{\sqrt{n}}\mathbf{u} - \frac{1}{\sqrt{np}}\mathbf{H}\,(\mathbf{P}\mathbf{w}) - \frac{||\mathbf{P}^{\perp}\mathbf{w}||}{\sqrt{np}}\mathbf{h}|| + \frac{1}{p}\boldsymbol{\beta}^T(\boldsymbol{\mu}-\mathbf{w})\,. \tag{8.30}$$

We are now able to perform the optimization with respect to $\mathbf{w}$. As explained above, we are allowed to flip the order of min and max in the asymptotic regime. We first analyze $\min_{\mathbf{w}}$ to find the optimal direction of $\mathbf{w}$. To streamline the notations, we introduce the scalars $\alpha := \frac{\mathbf{w}^T\mathbf{w}^\star}{||\mathbf{w}^\star||^2}$, and $\sigma := \frac{1}{\sqrt{p}}||\mathbf{P}^{\perp}\mathbf{w}||$. Also define $\mathbf{q} := \frac{1}{\kappa\sqrt{p}}\mathbf{H}\mathbf{w}^\star$, where $\mathbf{q}$ has i.i.d. standard normal entries (recall that $\mathbf{H}$ has i.i.d. standard normal entries). Optimizing with respect to the direction of $\mathbb{P}^{\perp}\mathbf{w}$ yields,

$$\min_{\substack{\boldsymbol{\mu}\in\mathbb{R}^p,\mathbf{u}\in\mathbb{R}^n\\ \alpha\in\mathbb{R},\sigma\geq 0}} \max_{\substack{\boldsymbol{\beta}\in\mathbb{R}^p\\ r\geq 0}} \frac{1}{n}\mathbf{1}^T\rho(\mathbf{u}) - \frac{1}{n}\mathbf{y}^T\mathbf{u} + \lambda f(\boldsymbol{\mu}) - \sigma||\frac{1}{\sqrt{p}}\mathbf{P}^{\perp}(r\sqrt{\delta}\mathbf{g}-\boldsymbol{\beta})||$$
$$+\,r||\frac{1}{\sqrt{n}}\mathbf{u} - \frac{\kappa\alpha}{\sqrt{n}}\mathbf{q} - \frac{\sigma}{\sqrt{n}}\mathbf{h}|| + \frac{1}{p}(\mathbf{P}\boldsymbol{\beta})^T\boldsymbol{\mu} + \frac{1}{p}(\mathbf{P}^{\perp}\boldsymbol{\beta})^T\boldsymbol{\mu} - \frac{1}{p}(\mathbf{P}\boldsymbol{\beta})^T\mathbf{w}\,, \tag{8.31}$$

where $\delta := \frac{p}{n}$ is the overparameterization ratio. Next, using a subtle trick by introducing two new scalar variables, namely υ and τ, we can change $||\cdot||$ to $||\cdot||^2$ which simplifies the next steps of our analysis. The new optimization would be,

$$\min_{\substack{\boldsymbol{\mu}\in\mathbb{R}^p,\mathbf{u}\in\mathbb{R}^n\\ \alpha\in\mathbb{R},\sigma\geq 0\\ \upsilon\geq 0}} \max_{\substack{\boldsymbol{\beta}\in\mathbb{R}^p\\ r,\tau\geq 0}} \frac{1}{n}\mathbf{1}^T\rho(\mathbf{u}) - \frac{1}{n}\mathbf{y}^T\mathbf{u} + \frac{\lambda}{p}f(\boldsymbol{\mu}) - \frac{\sigma}{2\tau} - \frac{\sigma\tau}{2}||\frac{1}{\sqrt{p}}\mathbf{P}^{\perp}(r\sqrt{\delta}\mathbf{g}-\boldsymbol{\beta})||^2 + \frac{r}{2\upsilon}$$
$$+\frac{r\upsilon}{2}||\frac{1}{\sqrt{n}}\mathbf{u} - \frac{\kappa\alpha}{\sqrt{n}}\mathbf{q} - \frac{\sigma}{\sqrt{n}}\mathbf{h}||^2 + \frac{1}{p}(\mathbf{P}\boldsymbol{\beta})^T\boldsymbol{\mu} + \frac{1}{p}(\mathbf{P}^{\perp}\boldsymbol{\beta})^T\boldsymbol{\mu} - \frac{1}{p}(\mathbf{P}\boldsymbol{\beta})^T\mathbf{w}\,. \tag{8.32}$$

Next, in order to compute the optimal $\boldsymbol{\beta}$, we use the following completion of squares,

$$-\frac{\sigma\tau}{2}||\frac{1}{\sqrt{p}}\mathbf{P}^{\perp}(r\sqrt{\delta}\mathbf{g}-\boldsymbol{\beta})||^2 + \frac{1}{p}(\mathbf{P}^{\perp}\boldsymbol{\beta})^T\boldsymbol{\mu} = -\frac{\sigma\tau}{2}||\frac{1}{\sqrt{p}}\mathbf{P}^{\perp}(r\sqrt{\delta}\mathbf{g}-\boldsymbol{\beta}+\frac{1}{\sigma\tau}\boldsymbol{\mu})||^2$$
$$+\frac{1}{2p\sigma\tau}||\mathbf{P}^{\perp}\boldsymbol{\mu}+\sigma\tau r\sqrt{\delta}\mathbf{P}^{\perp}\mathbf{g}||^2 - \frac{\sigma\tau r^2}{2n}||\mathbf{P}^{\perp}\mathbf{g}||^2\,. \tag{8.33}$$

Since $\mathbf{g} \in \mathbb{R}^p$ has standard normal entries, we can approximate $\frac{\sigma\tau r^2}{2n}||\mathbf{P}^\perp \mathbf{g}||^2$ with $\frac{\delta\sigma\tau r^2}{2}$. We exploit (8.33) to solve the inner optimization with respect to $\boldsymbol{\beta}$ which gives,

$$\begin{aligned}\min_{\substack{\boldsymbol{\mu}\in\mathbb{R}^p,\mathbf{u}\in\mathbb{R}^n \\ \alpha\in\mathbb{R},\sigma,\upsilon\geq 0 \\ \frac{1}{p}\mathbf{w}^{\star T}\boldsymbol{\mu}=\alpha\kappa^2}} \max_{r,\tau\geq 0} \; & \frac{1}{n}\mathbf{1}^T\rho(\mathbf{u}) - \frac{1}{n}\mathbf{y}^T\mathbf{u} - \frac{\sigma}{2\tau} - \frac{\delta\sigma\tau r^2}{2} + \frac{r}{2\upsilon} - \frac{\kappa^2\alpha^2}{2\sigma\tau} \\ & + \frac{r\upsilon}{2}||\frac{1}{\sqrt{n}}\mathbf{u} - \frac{\kappa\alpha}{\sqrt{n}}\mathbf{q} - \frac{\sigma}{\sqrt{n}}\mathbf{h}||^2 + \frac{1}{2p\sigma\tau}||\boldsymbol{\mu} + \sigma\tau r\sqrt{\delta}\mathbf{g}||^2 + \frac{\lambda}{p}f(\boldsymbol{\mu}) \; ,\end{aligned} \tag{8.34}$$

where we also used the following equality:

$$\begin{aligned}\frac{1}{p}||\mathbf{P}^\perp\boldsymbol{\mu} + \sigma\tau r\sqrt{\delta}\mathbf{P}^\perp\mathbf{g}||^2 &= \frac{1}{p}||\boldsymbol{\mu} + \sigma\tau r\sqrt{\delta}\mathbf{g}||^2 - \frac{1}{p}||\mathbf{P}\boldsymbol{\mu}||^2 - (\sigma\tau r)^2\frac{||\mathbf{P}\mathbf{g}||^2}{n} - \frac{2\sigma\tau r\sqrt{\delta}}{p}(\mathbf{P}\mathbf{g})^T\mu \\ \boxed{p\to+\infty} \qquad &= \frac{1}{p}||\boldsymbol{\mu} + \sigma\tau r\sqrt{\delta}\mathbf{g}||^2 - \frac{1}{p}||\mathbf{P}\boldsymbol{\mu}||^2 = \frac{1}{p}||\boldsymbol{\mu} + \sigma\tau r\sqrt{\delta}\mathbf{g}||^2 - \kappa^2\alpha^2 \; .\end{aligned} \tag{8.35}$$

Consequently, by flipping the order of min and max, we first compute the minimization with respect to $\boldsymbol{\mu}$. Hence, the optimal $\boldsymbol{\mu}$ would be the solution to the following optimization:

$$\begin{aligned}\min_{\boldsymbol{\mu}\in\mathbb{R}^p} \quad & \frac{1}{2p\sigma\tau}||\boldsymbol{\mu} - \sigma\tau r\sqrt{\delta}\mathbf{g}||^2 + \frac{\lambda}{p}f(\boldsymbol{\mu}) \\ \text{s.t.} \quad & \frac{1}{p}\mathbf{w}^{\star T}\boldsymbol{\mu} = \alpha\kappa^2 \; .\end{aligned} \tag{8.36}$$

Using the Lagrange multiplier θ, we can rewrite this optimization as,

$$\min_{\boldsymbol{\mu}\in\mathbb{R}^p} \max_{\theta\in\mathbb{R}} \quad \frac{1}{2p\sigma\tau}||\boldsymbol{\mu} - \sigma\tau r\sqrt{\delta}\mathbf{g}||^2 + \frac{\lambda}{p}f(\boldsymbol{\mu}) - \frac{\theta}{p}\mathbf{w}^{\star T}\boldsymbol{\mu} + \alpha\theta\kappa^2 \; . \tag{8.37}$$

Applying yet another completion of squares, we have,

$$\frac{1}{2p\sigma\tau}||\boldsymbol{\mu} - \sigma\tau r\sqrt{\delta}\mathbf{g}||^2 - \frac{\theta}{p}\mathbf{w}^{\star T}\boldsymbol{\mu} = \frac{1}{2p\sigma\tau}||\boldsymbol{\mu} - \sigma\tau r\sqrt{\delta}\mathbf{g} - \theta\sigma\tau\mathbf{w}^\star||^2 - \frac{\sigma\tau\theta^2\kappa^2}{2} \; , \tag{8.38}$$

where we omit the term $\frac{1}{p}\mathbf{g}^T\mathbf{w}^\star = O(\frac{1}{\sqrt{p}})$ as it is negligible compared to the other terms (which are of constant orders). We are able to represent the solution of (8.36) in terms of the Moreau envelope of the function $f(\cdot)$ as follows,

$$\min_{\substack{\boldsymbol{\mu}\in\mathbb{R}^p \\ \frac{1}{p}\mathbf{w}^{\star T}\boldsymbol{\mu}=\alpha\kappa^2}} \frac{1}{2p\sigma\tau}||\boldsymbol{\mu}-\sigma\tau r\sqrt{\delta}\mathbf{g}||^2+\frac{\lambda}{p}f(\boldsymbol{\mu}) \;=\; \max_{\theta\in\mathbb{R}} \frac{1}{p}M_{\lambda f}\big(\sigma\tau(r\sqrt{\delta}\mathbf{g}+\theta\mathbf{w}^\star),\sigma\tau\big)+\alpha\theta\kappa^2-\frac{\sigma\tau\theta^2\kappa^2}{2}\,. \tag{8.39}$$

Substituting (8.39) in (8.34), we have the following optimization:

$$\begin{aligned}\min_{\substack{\mathbf{u}\in\mathbb{R}^n \\ \alpha\in\mathbb{R},\sigma,\upsilon\geq 0}} \max_{\substack{r,\tau\geq 0 \\ \theta\in\mathbb{R}}} \frac{1}{n}\mathbf{1}^T\rho(\mathbf{u})-\frac{1}{n}\mathbf{y}^T\mathbf{u}+\frac{r\upsilon}{2}||\frac{1}{\sqrt{n}}\mathbf{u}-\frac{\kappa\alpha}{\sqrt{n}}\mathbf{q}-\frac{\sigma}{\sqrt{n}}\mathbf{h}||^2-\frac{\sigma}{2\tau}-\frac{\delta\sigma\tau r^2}{2}+\frac{r}{2\upsilon} \\ -\frac{\kappa^2\alpha^2}{2\sigma\tau}+\kappa^2\alpha\theta-\frac{\kappa^2\sigma\tau\theta^2}{2}+\frac{1}{p}M_{\lambda f(\cdot)}\big(\sigma\tau(r\sqrt{\delta}\mathbf{g}+\theta\mathbf{w}^\star),\sigma\tau\big)\,.\end{aligned} \tag{8.40}$$

We now focus on the optimization with respect to $\mathbf{u}$. Recall that $\mathbf{y} = Ber\big(\rho'(\frac{1}{\sqrt{p}}\mathbf{H}\mathbf{w}^\star)\big) = Ber\big(\rho'(\kappa\mathbf{q})\big)$. We are interested in solving the following optimization:

$$\min_{\mathbf{u}\in\mathbb{R}^n} \frac{1}{n}\mathbf{1}^T\rho(\mathbf{u})-\frac{1}{n}\mathbf{y}^T\mathbf{u}+\frac{r\upsilon}{2}||\frac{1}{\sqrt{n}}\mathbf{u}-\frac{\kappa\alpha}{\sqrt{n}}\mathbf{q}-\frac{\sigma}{\sqrt{n}}\mathbf{h}||^2\,. \tag{8.41}$$

Similarly to the previous steps, we first do a completion of squares as follows,

$$\begin{aligned}-\frac{1}{n}\mathbf{y}^T\mathbf{u}+\frac{r\upsilon}{2}||\frac{1}{\sqrt{n}}\mathbf{u}-\frac{\kappa\alpha}{\sqrt{n}}\mathbf{q}-\frac{\sigma}{\sqrt{n}}\mathbf{h}||^2 = \frac{r\upsilon}{2}||\frac{1}{\sqrt{n}}\mathbf{u}-\frac{\kappa\alpha}{\sqrt{n}}\mathbf{q}-\frac{\sigma}{\sqrt{n}}\mathbf{h}-\frac{1}{r\upsilon\sqrt{n}}\mathbf{y}||^2 \\ -\frac{1}{2r\upsilon}||\mathbf{y}||^2-\frac{k\alpha}{n}\mathbf{y}^T\mathbf{q}-\frac{\sigma}{n}\mathbf{y}^T\mathbf{h}\,.\end{aligned} \tag{8.42}$$

Next, we use the distribution of $\mathbf{y}$ to simplify the expressions in the right-hand side of (8.42). We can write,

$$\frac{1}{n}||\mathbf{y}||^2=\frac{1}{n}\sum_{i=1}^n y_i^2 \underset{n\to\infty}{\overset{\text{WLLN}}{\Longrightarrow}} \mathbb{E}[y_i^2]=\mathbb{E}[y_i]=\mathbb{E}_Z[\rho'(\kappa Z)]=\frac{1}{2}\,, \tag{8.43}$$

and,

$$\frac{1}{n}\mathbf{y}^T\mathbf{q}=\frac{1}{n}\sum_{i=1}^n y_i q_i=\frac{1}{n}\sum_{i=1}^n Ber(\rho'(\kappa q_i))q_i \underset{n\to\infty}{\overset{\text{WLLN}}{\Longrightarrow}} \mathbb{E}_Z[Z\cdot\rho'(\kappa Z)]=\kappa\,\mathbb{E}_Z[\rho''(\kappa Z)]\,, \tag{8.44}$$

where $Z\sim\mathcal{N}(0,1)$, and for the last equality, we used the Stein's lemma (Lemma 35 in Appendix A.2). Also note that we can ignore the term $\frac{\sigma}{n}\mathbf{y}^T\mathbf{h}$ since it is $O\left(\frac{1}{\sqrt{n}}\right)$ (which goes to zero in the

asymptotic regime while the other terms are of constant orders). Hence, we are able to rewrite the optimization (8.41) with respect to $\mathbf{u}$ in the following form:

$$\min_{\mathbf{u}\in\mathbb{R}^n} \frac{1}{n}\mathbf{1}^T\rho(\mathbf{u}) + \frac{r\upsilon}{2n}||\mathbf{u} - \kappa\alpha\mathbf{q} - \sigma\mathbf{h} - \frac{1}{r\upsilon}\mathbf{y}||^2 - \frac{1}{4r\upsilon} - \kappa^2\alpha\mathbb{E}_Z[\rho''(\kappa Z)] \ . \tag{8.45}$$

We can rewrite the equation (8.45) in terms of the Moreau envelope, $M_{\rho(\cdot)}$, as follows,

$$\begin{aligned}\min_{\substack{\sigma,\upsilon\geq 0\\ \alpha\in\mathbb{R}}} \max_{\substack{r,\tau\geq 0\\ \theta\in\mathbb{R}}} \ & -\frac{\sigma}{2\tau} - \frac{\delta\sigma\tau r^2}{2} + \frac{r}{2\upsilon} - \frac{\kappa^2\alpha^2}{2\sigma\tau} + \kappa^2\alpha\theta - \frac{\kappa^2\sigma\tau\theta^2}{2} - \frac{1}{4r\upsilon} - \kappa^2\alpha\mathbb{E}_Z[\rho''(\kappa Z)] \\ & + \frac{1}{p}M_{\lambda f(\cdot)}\big(\sigma\tau(r\sqrt{\delta}\mathbf{g} + \theta\mathbf{w}^\star), \sigma\tau\big) + \frac{1}{n}\, M_{\rho(\cdot)}\big(\kappa\alpha\mathbf{q} + \sigma\mathbf{h} + \frac{1}{r\upsilon}\mathbf{y}, \frac{1}{r\upsilon}\big) \ .\end{aligned} \tag{8.46}$$

As the last act in this step, we want to analyze the convergence properties of (AO). Recall that $f(\cdot)$ is a separable function. Therefore, using the result of Lemma 36 (see Appendix A.2), we have:

$$M_{\lambda f(\cdot)}\big(\sigma\tau(\frac{r}{\sqrt{\delta}}\mathbf{g} + \theta\mathbf{w}^\star), \sigma\tau\big) = \sum_{i=1}^{p} M_{\lambda\tilde{f}(\cdot)}\big(\sigma\tau(r\sqrt{\delta}\mathbf{g}_i + \theta\mathbf{w}_i^\star), \sigma\tau\big) \ . \tag{8.47}$$

Using the strong law of large numbers, we have,

$$\frac{1}{p}\, M_{\lambda f(\cdot)}\big(\sigma\tau(r\sqrt{\delta}\mathbf{g} + \theta\mathbf{w}^\star), \sigma\tau\big) \xrightarrow{a.s.} \mathbb{E}\big[M_{\lambda\tilde{f}(\cdot)}\big(\sigma\tau(r\sqrt{\delta}Z + \theta W), \sigma\tau\big)\big] \ , \tag{8.48}$$

where Z is a standard normal random variable and $W \sim \Pi$ is independent of Z. Similarly, we can write,

$$\frac{1}{n}\, M_{\rho(\cdot)}\big(\kappa\alpha\mathbf{q} + \sigma\mathbf{h} + \frac{1}{r\upsilon}\mathbf{y}, \frac{1}{r\upsilon}\big) \xrightarrow{a.s.} \mathbb{E}\big[M_{\rho(\cdot)}\big(\kappa\alpha Z_1 + \sigma Z_2 + \frac{1}{r\upsilon}Ber(\kappa Z_1), \frac{1}{r\upsilon}\big)\big] \ . \tag{8.49}$$

We appeal to Lemma 33 to analyze the convergence properties of (AO). Due to the convergence we are getting from the LLN, applying this lemma enables us to replace the Moreau envelopes with their expected value. Hence, we need to analyze the following optimization,

$$\begin{aligned}\min_{\substack{\sigma,\upsilon\geq 0\\ \alpha\in\mathbb{R}}} \max_{\substack{r,\tau\geq 0\\ \theta\in\mathbb{R}}} \ & -\frac{\sigma}{2\tau} - \frac{\delta\sigma\tau r^2}{2} + \frac{r}{2\upsilon} - \frac{\kappa^2\alpha^2}{2\sigma\tau} + \kappa^2\alpha\theta - \frac{\kappa^2\sigma\tau\theta^2}{2} - \frac{1}{4r\upsilon} - \kappa^2\alpha\mathbb{E}_Z[\rho''(\kappa Z)] \\ & + \mathbb{E}\big[M_{\lambda\tilde{f}(\cdot)}\big(\sigma\tau(r\sqrt{\delta}Z + \theta W), \sigma\tau\big)\big] + \mathbb{E}\big[M_{\rho(\cdot)}\big(\kappa\alpha Z_1 + \sigma Z_2 + \frac{1}{r\upsilon}Ber(\kappa Z_1), \frac{1}{r\upsilon}\big)\big] \ .\end{aligned} \tag{8.50}$$

Finding the optimality condition of the scalar optimization

In this section, we conclude the proof of the main result of the paper. For this, we need to show that the optimizer of the optimization (8.50) can be found by solving the nonlinear system of equations (NLS). Before analyzing the auxiliary optimization, we state two lemmas that will be used in our analysis.

The next two lemmas present some properties of the proximal operator of the function $\rho(z) = \log(1 + e^z)$.

Lemma 19. *Let $\rho(z) = \log(1 + e^z)$, then the following identity holds,*

$$Prox_{t\rho}(x + t) = -Prox_{t\rho}(-x) \ . \tag{8.51}$$

Proof. Since the function $\rho(\cdot)$ is differentiable, the proximal operator satisfies the following equation,

$$\frac{1}{t}(\text{Prox}_{t\rho(\cdot)}(x) - x) + \rho'(\text{Prox}_{t\rho(\cdot)}(x)) = 0 \ . \tag{8.52}$$

Next we use the fact that $\rho'(-z) = 1 - \rho'(z)$ for $z \in \mathbb{R}$, to rewrite the equation as follows,

$$\frac{1}{t}\big(- \text{Prox}_{t\rho(\cdot)}(-x) - (x + t)\big) + \rho'(-\text{Prox}_{t\rho(\cdot)}(-x)) = 0 \ , \tag{8.53}$$

which gives the desired identity. □

Lemma 20. *The derivative of the proximal operator of the function $\rho(\cdot)$ can be computed as follows,*

$$\frac{d}{dx}Prox_{t\rho(\cdot)}(x) = \frac{1}{1 + t\rho''\big(Prox_{t\rho(\cdot)}(x)\big)} \ . \tag{8.54}$$

Proof. Taking derivative with respect to x of (8.52),

$$\frac{1}{t}(\frac{d}{dx}\text{Prox}_{t\rho(\cdot)}(x) - 1) + \frac{d}{dx}\text{Prox}_{t\rho(\cdot)}(x) \times \rho''\big(\text{Prox}_{t\rho(\cdot)}(x)\big) = 0 \ , \tag{8.55}$$

which can be written as in (8.54). □

Let $C(\alpha, \sigma.r, \tau, \upsilon, \theta)$ denote the objective function in (8.50). We want to find the optimizer of $C(\cdot)$, i.e., the point $(\alpha^\star, \sigma^\star, r^\star, \tau^\star, \upsilon^\star, \theta^\star)$. Since the objective function is smooth, when the optimal values are all non-zero, they should satisfy the first order optimality condition, i.e.,

$$\nabla C = \mathbf{0} \ . \tag{8.56}$$

We will show that the (8.56) would simplify to our system of nonlinear equations. We start by putting the derivative w.r.t. θ equal to zero. We have the following,

$$\frac{\partial C}{\partial \theta} = 0 \Rightarrow \kappa^2\alpha - \kappa^2\sigma\tau\theta + \frac{1}{p}\mathbb{E}\big[\mathbf{w}^{\star T}\big(\tau\sigma(r\sqrt{\delta}\mathbf{g} + \theta\mathbf{w}^\star) - \text{Prox}_{\sigma\tau\lambda f(\cdot)}(\sigma\tau(r\sqrt{\delta}\mathbf{g} + \theta\mathbf{w}^\star))\big)\big] = 0\,, \tag{8.57}$$

where we used Lemma 34 (in Appendix A.2) for taking the derivative of the Moreau envelope, $M_{\lambda f(\cdot)}$. We can simplify (8.57) and rewrite it as follows,

$$\kappa^2\alpha = \frac{1}{p}\mathbb{E}\big[\mathbf{w}^{\star T}\text{Prox}_{\sigma\tau\lambda f(\cdot)}(\sigma\tau(r\sqrt{\delta}\mathbf{g} + \theta\mathbf{w}^\star))\big]\,. \tag{8.58}$$

Next, we take derivative of the objective function $C(\cdot)$ w.r.t. r and υ and put that equal to zero. We state the following lemma which will be exploited in taking the derivatives.

Lemma 21. *For fixed values of $\kappa, \alpha,$ and σ, let the function $F : \mathbb{R}_+ \to \mathbb{R}$ be defined as follows,*

$$F(\gamma) = -\frac{1}{4}\gamma + \mathbb{E}_{Z_1,Z_2}\big[M_{\rho(\cdot)}\big(\kappa\alpha Z_1 + \sigma Z_2 + \gamma Ber(\rho'(\kappa Z_1)), \gamma\big)\big], \tag{8.59}$$

then the derivative of $F(\cdot)$ would be as follows:

$$F'(\gamma) = -\frac{1}{\gamma^2}\mathbb{E}\big[\rho'(-\kappa Z_1)\big(\kappa\alpha Z_1 + \sigma Z_2 - Prox_{\gamma\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2)\big)^2\big]\,. \tag{8.60}$$

Proof. We have,

$$F'(\gamma) = -\frac{1}{4} + \frac{d}{d\gamma}\,\mathbb{E}_{Z_1,Z_2}\big[M_{\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2 + \gamma Ber(\rho'(\kappa Z_1)), \gamma)\big]\,. \tag{8.61}$$

In order to compute the last derivative, we exploit Lemma 34. We have,

$$\begin{aligned}\frac{d}{d\gamma}\,\mathbb{E}\big[M_{\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2 + \gamma Ber(\rho'(\kappa Z_1)), \gamma)\big] &= -\mathbb{E}\big[\frac{\rho'(-\kappa Z_1)}{2\gamma^2}\big(\kappa\alpha Z_1 + \sigma Z_2 - \text{Prox}_{\gamma\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2)\big)^2\big] \\ &\quad - \mathbb{E}\big[\frac{\rho'(\kappa Z_1)}{2\gamma^2}\big(\kappa\alpha Z_1 + \sigma Z_2 + \gamma - \text{Prox}_{\gamma\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2 + \gamma)\big)^2\big] \\ &\quad + \mathbb{E}\big[\frac{\rho'(\kappa Z_1)}{\gamma}\big(\kappa\alpha Z_1 + \sigma Z_2 + \gamma - \text{Prox}_{\gamma\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2 + \gamma)\big)\big]\,,\end{aligned} \tag{8.62}$$

where we used the fact that for $x \in \mathbb{R}$, $\rho'(-x) = 1 - \rho'(x)$. To derive (8.60), we appeal to the result of Lemma 19 which gives the following identity,

$$\text{Prox}_{\gamma\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2 + \gamma) = -\text{Prox}_{\gamma\rho(\cdot)}(-\kappa\alpha Z_1 - \sigma Z_2)\,. \tag{8.63}$$

□

Proceeding onward, we use the result of Lemma 21 to find the optimality conditions with respect to r and υ. We have,

$$\begin{cases} \frac{\partial}{\partial r}C = 0 \Rightarrow -\delta\sigma\tau r + \frac{1}{2\upsilon} - \frac{1}{\upsilon r^2}F'(\frac{1}{\upsilon r}) + \frac{1}{p}\mathbb{E}\big[\sqrt{\delta}\mathbf{g}^T\big(\sigma\tau r\sqrt{\delta}\mathbf{g} - \text{Prox}_{\sigma\tau\lambda f(\cdot)}(\sigma\tau(r\sqrt{\delta}\mathbf{g} + \theta\mathbf{w}^\star))\big)\big] = 0\,, \\ \frac{\partial}{\partial \upsilon}C = 0 \Rightarrow \frac{-r}{2\upsilon^2} - \frac{1}{r\upsilon^2}F'(\frac{1}{r\upsilon}) = 0\,. \end{cases} \tag{8.64}$$

In order to simplify the equations, we define a new variable $\gamma := \frac{1}{r\upsilon}$. We can rewrite the equations (8.64) as follows,

$$\begin{cases} \gamma = \frac{1}{p}\mathbb{E}\big[\frac{\sqrt{\delta}\mathbf{g}^T}{r}\text{Prox}_{\sigma\tau\lambda f(\cdot)}\big(\sigma\tau(r\sqrt{\delta}\mathbf{g} + \theta\mathbf{w}^\star)\big)\big]\,, \\ \gamma^2 = \mathbb{E}\big[\frac{2\rho'(-\kappa Z_1)}{r^2}\big(\kappa\alpha Z_1 + \sigma Z_2 - \text{Prox}_{\gamma\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2)\big)^2\big]\,. \end{cases} \tag{8.65}$$

So far we have shown that three of the optimality conditions are the same as the nonlinear equations 1,2, and 5 in (NLS). Next, we take the derivative w.r.t. τ. We have,

$$\frac{\partial}{\partial\tau}C = 0 \Rightarrow \frac{\sigma}{2\tau^2} - \frac{\delta\sigma r^2}{2} + \frac{\kappa^2\alpha^2}{2\sigma\tau^2} - \frac{\kappa^2\sigma\theta^2}{2} + \frac{1}{p}\,\frac{\partial}{\partial\tau}\mathbb{E}[M_{\lambda f(\cdot)}\big(\sigma\tau(r\sqrt{\delta}\mathbf{g} + \theta\mathbf{w}^\star), \sigma\tau\big)] = 0\,. \tag{8.66}$$

The derivative of the expected Moreau envelope can be computed as follows,

$$\frac{1}{p}\,\frac{\partial}{\partial\tau}\mathbb{E}[M_{\lambda f(\cdot)}\big(\sigma\tau(r\sqrt{\delta}\mathbf{g}+\theta\mathbf{w}^\star), \sigma\tau\big)] = \frac{\sigma}{2}(\delta r^2+\theta^2\kappa) - \frac{1}{2\sigma\tau^2}\mathbb{E}\big[\|\text{Prox}_{\sigma\tau\lambda f(\cdot)}\big(\sigma\tau(r\sqrt{\delta}\mathbf{g}+\theta\mathbf{w}^\star)\big)\|_2^2\big]\,. \tag{8.67}$$

Replacing (8.67) in (8.66) would result in,

$$(\kappa\alpha)^2 + \sigma^2 = \mathbb{E}\big[\|\text{Prox}_{\sigma\tau\lambda f(\cdot)}\big(\sigma\tau(r\sqrt{\delta}\mathbf{g} + \theta\mathbf{w}^\star)\big)\|_2^2\big]\,. \tag{8.68}$$

(8.68) is the third equation in the nonlinear system (NLS). Next, putting the derivative w.r.t. σ to equal zero gives the following,

$$\begin{aligned} -\frac{1}{2\tau} - \frac{\delta\tau r^2}{2} + \frac{\kappa^2\alpha^2}{2\sigma^2\tau} - \frac{\kappa^2\tau\theta^2}{2} + \frac{1}{p}\,\frac{\partial}{\partial\sigma}\mathbb{E}[M_{\lambda f(\cdot)}\big(\sigma\tau(r\sqrt{\delta}\mathbf{g} + \theta\mathbf{w}^\star), \sigma\tau\big)] \\ + \frac{\partial}{\partial\sigma}\mathbb{E}\big[M_{\rho(\cdot)}\big(\kappa\alpha Z_1 + \sigma Z_2 + \gamma Ber(\kappa Z_1), \gamma\big)\big] = 0\,. \end{aligned} \tag{8.69}$$

We can compute the partial derivative of the expected Moreau envelopes as follows,

$$\frac{1}{p}\frac{\partial}{\partial\sigma}\mathbb{E}[M_{\lambda f(\cdot)}\big(\sigma\tau(r\sqrt{\delta}\mathbf{g}+\theta\mathbf{w}^\star),\sigma\tau\big)] = \frac{\tau}{2}(\delta r^2+\theta^2\kappa)-\frac{1}{2\sigma^2\tau}\mathbb{E}\big[||\text{Prox}_{\sigma\tau\lambda f(\cdot)}\big(\sigma\tau(r\sqrt{\delta}\mathbf{g}+\theta\mathbf{w}^\star)\big)||_2^2\big]\,, \tag{8.70}$$

and,

$$\begin{aligned}\frac{\partial}{\partial\sigma}\mathbb{E}\big[M_{\rho(\cdot)}\big(\kappa\alpha Z_1+\sigma Z_2+\gamma Ber(\kappa Z_1),\gamma\big)\big] &= \frac{\sigma}{\gamma}-\frac{2}{\gamma}\mathbb{E}\big[Z_2\rho'(-\kappa Z_1)\text{Prox}_{\gamma\rho(\cdot)}\big(\kappa\alpha Z_1+\sigma Z_2\big)\big]\,,\\ &= \frac{\sigma}{\gamma}\big(1-2\mathbb{E}\big[\frac{\rho'(-\kappa Z_1)}{1+\gamma\rho''\big(\text{Prox}_{\gamma\rho(\cdot)}(\kappa\alpha Z_1+\sigma Z_2)\big)}\big]\big)\,.\end{aligned} \tag{8.71}$$

To derive the last equality, we used Lemma 20 as well as Stein's lemma.

Replacing (8.70) and (8.71) in (8.69) gives,

$$1-\frac{\gamma}{\tau\sigma} = \mathbb{E}\big[\frac{2\rho'(-\kappa Z_1)}{1+\gamma\rho''\big(\text{Prox}_{\gamma\rho(\cdot)}(\kappa\alpha Z_1+\sigma Z_2)\big)}\big]\,. \tag{8.72}$$

As the last step in deriving the first-order optimality conditions, we take the derivative with respect to α in order to derive the fourth equation in the nonlinear system (NLS). We have,

$$\frac{\partial C}{\partial\alpha} = \frac{-\kappa^2\alpha}{\sigma\tau}+\kappa^2\theta-\kappa^2\mathbb{E}[\rho''(\kappa Z)]+\frac{\partial}{\partial\alpha}\mathbb{E}[M_{\rho(\cdot)}\big(\kappa\alpha Z_1+\sigma Z_2+\gamma Ber(\rho'(\kappa Z_1)),\gamma\big)] = 0\,. \tag{8.73}$$

To simplify this equation, we write,

$$-\kappa^2\mathbb{E}[\rho''(\kappa Z)]+\frac{\partial}{\partial\alpha}\mathbb{E}[M_{\rho(\cdot)}\big(\kappa\alpha Z_1+\sigma Z_2+\gamma Ber(\rho'(\kappa Z_1)),\gamma\big)] = \frac{\kappa^2\alpha}{\gamma}-2\mathbb{E}\big[\frac{\kappa}{\gamma}Z_1\rho'(-\kappa Z_1)\text{Prox}_{\gamma\rho(\cdot)}\big(\kappa\alpha Z_1+\sigma Z_2\big)\big]\,. \tag{8.74}$$

Replacing (8.74) in (8.73) would result in,

$$\frac{\gamma\kappa}{2}(\theta-\frac{\alpha}{\sigma\tau})+\frac{\kappa\alpha}{2} = \mathbb{E}\big[Z_1\rho'(-\kappa Z_1)\text{Prox}_{\gamma\rho(\cdot)}\big(\kappa\alpha Z_1+\sigma Z_2\big)\big]\,. \tag{8.75}$$

Using Stein's lemma, we can rewrite the right-hand-side as,

$$\begin{aligned}\text{RHS} &= -\mathbb{E}\big[\kappa\rho''(-\kappa Z_1)\text{Prox}_{\gamma\rho(\cdot)}\big(\kappa\alpha Z_1+\sigma Z_2\big)\big]+\kappa\alpha\mathbb{E}[\frac{\rho'(-\kappa Z_1)}{1+\gamma\rho''\big(\text{Prox}_{\gamma\rho(\cdot)}(\kappa\alpha Z_1+\sigma Z_2)\big)}],\\ &= -\mathbb{E}\big[\kappa\rho''(-\kappa Z_1)\text{Prox}_{\gamma\rho(\cdot)}\big(\kappa\alpha Z_1+\sigma Z_2\big)\big]+\frac{\kappa\alpha}{2}-\frac{\kappa\alpha\gamma}{2\tau\sigma}\,,\end{aligned} \tag{8.76}$$

where we exploited (8.72) to derive the last equation. Substituting in (8.75) would give,

$$\gamma\theta = -2\mathbb{E}\big[\rho''(-\kappa Z_1)\text{Prox}_{\gamma\rho(\cdot)}\big(\kappa\alpha Z_1 + \sigma Z_2\big)\big] \ . \tag{8.77}$$

Therefore, we have shown that the nonlinear system (NLS) is equivalent to the optimality condition in (8.50).

Recall that in the process of simplifying (AO), we introduced the Moreau envelope of $f(\cdot)$ in (8.39). The optimizer of that Moreau envelope gives the solution of the Auxiliary optimization. Let $(\bar{\alpha}, \bar{\sigma}, \bar{\gamma}, \bar{\theta}, \bar{\tau}, \bar{r})$ be the unique solution of the nonlinear system. Hence, we can present the solution of the (AO) in terms of the proximal operator as follows,

$$\hat{\boldsymbol{\beta}}_i^{AO} = \Gamma(\boldsymbol{\beta}_i^*, Z) = \text{Prox}_{\lambda\bar{\sigma}\bar{\tau}\tilde{f}(\cdot)}\big(\bar{\sigma}\bar{\tau}(\bar{\theta}\mathbf{w}_i^\star + \frac{\bar{r}}{\sqrt{\delta}}Z)\big), \quad \text{for } \ i = 1, 2, \dots p. \tag{8.78}$$

As the last step, we want to show the convergence of the locally-Lipschitz function $\Psi(\cdot,\cdot)$. Earlier in this section, in the process of applying the CGMT, we have introduced some artificial bounded sets on the optimization variables and stated that we can perform the optimization over these sets. Considering that the variables belong to those bounded sets, we can state that the function $\Psi(\cdot,\cdot)$ is Lipschitz, as constraining a locally-Lipschitz function to a bounded set gives a Lipschitz function. Next, using the strong law of large numbers along with the fact that the entries of $\mathbf{w}^*$ are i.i.d. and drawn from distribution Π, we have,

$$\frac{1}{p}\sum_{i=1}^{p}\Psi(\hat{\mathbf{w}}_i^{AO}, \mathbf{w}_i^\star) \xrightarrow{a.s.} \mathbb{E}\big[\Psi(\Gamma(W, Z), W)\big] \ , \tag{8.79}$$

where Z is a standard normal random variable and $W \sim \Pi$ is independent of Z.

Exploiting the asymptotic convergence of CGMT (Lemma 32), we can introduce the set $\mathcal{S}$ as follows,

$$\mathcal{S} = \{\mathbf{w} \in \mathbb{R}^p : |\frac{1}{p}\sum_{i=1}^{p}\Psi(\mathbf{w}, \mathbf{w}_i^*) - \mathbb{E}\big[\Psi(\Gamma(W, Z), W)\big]| > \epsilon\} \ . \tag{8.80}$$

The convergence in (8.79) would establish that as $p \to \infty$, $\hat{\mathbf{w}}^{AO} \in \mathcal{S}$ with probability approaching 1. Therefore, as the result of Lemma 32, $\hat{\mathbf{w}} = \hat{\mathbf{w}}^{PO} \in \mathcal{S}$ with probability approaching 1. This concludes the proof of Theorem 11.

8.8 Proof of Theorem 12

This result can be derived using the result of Theorem 11. We just need to show that the parameters θ, r, and τ can be explicitly computed from the first three equations in the nonlinear system (NLS).

Recall that we characterize the performance of the RLR in terms of the solution of the following nonlinear equation,

$$\begin{cases} \kappa^2\alpha = \mathbb{E}\big[W\ \mathrm{Prox}_{\lambda\sigma\tau\tilde{f}(\cdot)}\big(\sigma\tau(\theta W + r\sqrt{\delta}Z)\big)\big]\ , \\ \gamma = \dfrac{\sqrt{\delta}}{r}\,\mathbb{E}\big[Z\ \mathrm{Prox}_{\lambda\sigma\tau\tilde{f}(\cdot)}\big(\sigma\tau(\theta W + r\sqrt{\delta}Z)\big)\big]\ , \\ \kappa^2\alpha^2 + \sigma^2 = \mathbb{E}\ \big[\mathrm{Prox}_{\lambda\sigma\tau\tilde{f}(\cdot)}\big(\sigma\tau(\theta W + r\sqrt{\delta}Z)\big)^2\big]\ , \\ \gamma^2 = \dfrac{2}{r^2}\,\mathbb{E}\big[\rho'(-\kappa Z_1)\big(\kappa\alpha Z_1 + \sigma Z_2 - \mathrm{Prox}_{\gamma\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2)\big)^2\big]\ , \\ \theta\gamma = -2\,\mathbb{E}\big[\rho''(-\kappa Z_1)\mathrm{Prox}_{\gamma\rho(\cdot)}\big(\kappa\alpha Z_1 + \sigma Z_2\big)\big]\ , \\ 1 - \dfrac{\gamma}{\sigma\tau} = \mathbb{E}\big[\dfrac{2\rho'(-\kappa Z_1)}{1 + \gamma\rho''\big(\mathrm{Prox}_{\gamma\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2)\big)}\big]\ . \end{cases} \tag{8.81}$$

In the ℓ_2^2-regularization, we have $\tilde{f}(\cdot) = \frac{1}{2}(\cdot)^2$, for which the proximal operator can be computed in closed-form, i.e., we have,

$$\mathrm{Prox}_{t\tilde{f}}(x) = \frac{x}{1+t}\ . \tag{8.82}$$

Replacing in the first equation of (8.81) gives,

$$\begin{aligned} \kappa^2\alpha &= \mathbb{E}\big[W\ \mathrm{Prox}_{\lambda\sigma\tau\tilde{f}(\cdot)}\big(\sigma\tau(\theta W + r\sqrt{\delta}Z)\big)\big] \\ &= \mathbb{E}\big[W \times \frac{\sigma\tau(\theta W + r\sqrt{\delta}Z)}{1+\lambda\sigma\tau}\big] = \frac{\sigma\tau\theta\kappa^2}{1+\lambda\sigma\tau}\ , \end{aligned} \tag{8.83}$$

where we used the fact that $\mathbb{E}[W^2] = \kappa^2$ and $\mathbb{E}[W \cdot Z] = 0$ due to the independence of W and Z. Next, from the second equation in (8.81), we have,

$$\begin{aligned} \gamma &= \frac{\sqrt{\delta}}{r}\mathbb{E}\big[Z\ \mathrm{Prox}_{\lambda\sigma\tau\tilde{f}(\cdot)}\big(\sigma\tau(\theta W + r\sqrt{\delta}Z)\big)\big] \\ &= \frac{\sqrt{\delta}}{r}\mathbb{E}\big[Z \times \frac{\sigma\tau(\theta W + r\sqrt{\delta}Z)}{1+\lambda\sigma\tau}\big] = \frac{\delta\sigma\tau}{(1+\lambda\sigma\tau)}\ , \end{aligned} \tag{8.84}$$

and finally from the third equation in (8.81), we can compute,

$$\begin{aligned} \kappa^2\alpha^2 + \sigma^2 &= \mathbb{E}\big[\big(\ \mathrm{Prox}_{\lambda\sigma\tau\tilde{f}(\cdot)}\big(\sigma\tau(\theta W + r\sqrt{\delta}Z)\big)\big)^2\big] \\ &= \frac{\sigma^2\tau^2}{(1+\lambda\sigma\tau)^2}(\theta^2\kappa^2 + \delta r^2) \\ &= \kappa^2\alpha^2 + \frac{\delta\sigma^2\tau^2 r^2}{(1+\lambda\sigma\tau)^2}\ . \end{aligned} \tag{8.85}$$

We can rewrite the equations (8.83), (8.84), and (8.85) as follows,

$$\begin{cases} \theta = \dfrac{\alpha\delta}{\gamma}\,, \\ \tau = \dfrac{\gamma}{\sigma\big(\delta - \lambda\gamma\big)}\,, \\ r = \dfrac{\sigma\sqrt{\delta}}{\gamma}\,. \end{cases} \tag{8.86}$$

Replacing the derived expressions in (8.86) for θ, r, and τ in the last three equations of (8.81) gives the following system of three equations with three unknowns,

$$\begin{cases} \dfrac{\delta\sigma^2}{2} = \mathbb{E}\big[\rho'(-\kappa Z_1)\big(\kappa\alpha Z_1 + \sigma Z_2 - \mathrm{Prox}_{\gamma\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2)\big)^2\big]\,, \\ -\dfrac{\delta\alpha}{2} = \mathbb{E}\big[\rho''(-\kappa Z_1)\mathrm{Prox}_{\gamma\rho(\cdot)}\big(\kappa\alpha Z_1 + \sigma Z_2\big)\big]\,, \\ 1 - \delta + \lambda\gamma = \mathbb{E}\big[\dfrac{2\rho'(-\kappa Z_1)}{1 + \gamma\rho''\big(\mathrm{Prox}_{\gamma\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2)\big)}\big]\,. \end{cases} \tag{8.87}$$

This concludes the proof of Theorem 12.

PERFORMANCE OF EXTENDED MARGIN MAXIMIZERS ON SEPARABLE DATA

[1] F. Salehi et al. "The Performance Analysis of Generalized Margin Maximizers on Separable Data". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 8417–8426.

Logistic models are commonly used for binary classification tasks. The success of such models has often been attributed to their connection to maximum-likelihood estimators. It has been shown that gradient descent algorithm, when applied on the logistic loss, converges to the max-margin classifier (a.k.a. hard-margin SVM). The performance of the max-margin classifier has been recently analyzed in [95, 39]. Inspired by these results, in this chapter, we present and study a more general setting, where the underlying parameters of the logistic model possess certain structures (sparse, block-sparse, low-rank, etc.) and introduce a more general framework (which is referred to as "Extended Margin Maximizer", EMM). While classical max-margin classifiers minimize the 2-norm of the parameter vector subject to linearly separating the data, EMM minimizes any arbitrary convex function of the parameter vector. We provide a precise analysis of the performance of EMM via the solution of a system of nonlinear equations. We also provide a detailed study for three special cases: (1) ℓ_2-EMM that is the max-margin classifier, (2) ℓ_1-EMM which encourages sparsity, and (3) ℓ_∞-EMM which is often used when the parameter vector has binary entries. Our theoretical results are validated by extensive simulation results across a range of parameter values, problem instances, and model structures.

9.1 Motivations and Background

Machine learning models have been very successful in many applications, ranging from spam detection, face and pattern recognition, to the analysis of genome sequencing and financial markets. However, despite this indisputable success, our knowledge on why the various machine learning methods exhibit the performances they do is still at a very early stage. To make this gap between the theory and the practice narrower, researchers have recently begun to revisit simple machine learning models with the hope that understanding their performance will lead the way to understanding the performance of more complex machine learning methods.

More specifically, studies on the performance of different classifiers for binary classification dates

back to the seminal work of Vapnik in the 1980's [140]. In an effort to find the "optimal" hyperplane that separates the data, he presented an upper bound on the test error which is inversely proportional to the margin (minimum distance of the data points to the separating hyperplane), and concluded that the max-margin classifier is indeed the desired classifier. It has also been observed that to construct such optimal hyperplanes, one only has to take into account a small amount of the training data, the so-called support vectors [37].

In this chapter, we challenge the conventional wisdom by showing that when the underlying parameter has a certain structure, one can come up with classifiers that outperform the max-margin classifier. We introduce the **E**xtended **M**argin **M**aximizer (**EMM**) which takes into account the structure of the underlying parameter as well as the minimum distance of the data points to the separating hyperplane. We provide sharp asymptotic results on various performance measures (such as the generalization error) of EMM and show that an appropriate choice of the potential function can in fact improve the resulting estimator.

Prior work

There have been many recent attempts to understand the generalization behavior of simple machine learning models [12, 92, 149, 14, 62]. Most of these studies focus on the least-squares/ridge regression, where the loss function is the squared ℓ_2-norm, and derive sharp asymptotic results on the performance of the estimator. In particular, in [62, 79] the authors have shown that the minimum-norm least square solution demonstrates the so-called "double-descent" behavior [15].

A more recent line of research studies the generalization performance of gradient descent (GD) for binary classification. It has been shown [120]) that for a separable data set, GD (when applied on the logistic loss) converges in the direction to the max-margin classifier (a.k.a. hard-margin SVM). The performance of max-margin classifier has been recently analyzed in two independent works [95, 39]. It is worth noting that understanding the implicit bias of optimization algorithms in more complex machine learning models has recently gained a lot of attention [98, 50]. These understandings can justify interesting properties of machine learning models observed in practice.

Summary of contributions

Inspired by the recent results in understanding the performance of the max-margin classifier, in this chapter we introduce and study a more general framework. We assume that the underlying parameters possess certain structures (e.g. sparse) and introduce the extended margin maximizer (EMM) as the solution of a convex optimization problem whose objective function encourages the structure.

We analyze the performance of EMM in the high-dimensional regime where both the number of parameters, p, and the number of samples n grows, and analyze the asymptotic performance as a function of the overparameterization ratio $\delta := \frac{p}{n} > 0$.

In Chapter 7, we provided the phase transition condition for the separability of data, i.e., we derived the exact value of δ^* such that the data is separable for all $\delta > \delta^*$.

Here, we analyze the performance in the interpolating regime ($\delta > \delta^*$). To the best of our knowledge, our result presented here is the first in the literature that introduces the extended margin maximizers, and provides sharp asymptotic results on the performance of EMM classifiers on separable data.

Recently, there have been a series of works by multiple groups of researchers to characterize the performance of the logistic loss minimizer in binary classification [113, 126] (see Chapter 8 for more details). Furthermore, in an analogous avenue of research, the CGMT framework has been utilized to study the generalization behavior of the gradient descent algorithm in the interpolating regime, where there exists a (nonempty) set of parameters that perfectly fit the training data [95, 39].

The organization of this chapter is as follows: In Section 9.2, we mathematically introduce the problem. Section 9.3 contains the main results of this chapter where we present the precise performance analysis of EMM, which then will be used to compute the generalization error. We investigate our theoretical findings for three specific cases of potential functions in Section 9.4. Numerical simulations for the generalization error of the EMM classifiers are presented in Section 9.5. We should note that most technical derivations of the results presented in the chapter are deferred to the end of the chapter.

9.2 Problem Setup

We consider the problem of binary classification, having a set of training data, $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where each of the sample points consists of a p-dimensional feature vector, $\mathbf{x}_i$, and a binary label, $y_i \in \{\pm 1\}$. We assume that the data set $\mathcal{D}$ is generated from a logistic-type model with the underlying parameter $\mathbf{w}^\star \in \mathbb{R}^p$. This means that

$$y_i \sim \textsc{Rad}(\rho(\mathbf{x}_i^T \mathbf{w}^\star)) \ , \quad i = 1, \dots, n \ , \tag{9.1}$$

where $\rho : \mathbb{R} \to [0, 1]$ is a non-decreasing function and is often referred to as the link function. A commonly-used instance of the link function is the standard logistic function defined as $\rho(t) := \frac{1}{1+e^{-t}}$.

When n/p is sufficiently large, i.e., when we have access to a sufficiently large number of samples, the maximum-likelihood estimator($\hat{\mathbf{w}}_{ML}$) is well-defined. In such settings, the MLE is often the estimator of choice due to its desirable properties in the classical statistics. Sur and Candès [123]

have recently studied the performance of the MLE in logistic regression in the high-dimensional regime, where the number of observations and parameters are comparable, and show, among other things, that the maximum likelihood estimator is biased. Their results have been extended to regularized logistic regression [113], assuming some prior knowledge on the structure of the data. In particular, it has been observed that, when the regularization parameter is tuned properly, the regularized logistic regression can outperform the MLE.

Inspired by the recent results on analyzing the generalization error of machine learning models, in this chapter, we study the generalization error of binary classification in a regime of parameters known as the interpolating regime. Here, the assumption is that there exists a parameter vector that can perfectly fit (interpolate) the data, i.e.,

$$\exists \mathbf{w}_0 \;\; \text{s.t.} \;\; \text{SIGN}(\mathbf{w}_0^T \mathbf{x}_i) = y_i, \;\; \text{for } i = 1, 2, \ldots, n. \tag{9.2}$$

Let $\mathcal{W}$ denote the set of all the parameters that interpolate the data.

$$\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^p : \text{SIGN}(\mathbf{w}^T \mathbf{x}_i) = y_i \;, \;\; \text{for } 1 \leq i \leq n.\}. \tag{9.3}$$

It has been observed that in many machine learning tasks, the iterative solvers that minimize the loss function often converge to one of the points in the set $\mathcal{W}$ (the training error converges to zero). Therefore, one can (qualitatively) pose the following important (yet still mysterious) question:

Which point(s) in $\mathcal{W}$ is (are) "better" estimator(s) of the actual parameter, $\mathbf{w}^\star$?

In an attempt to find an answer to this question, we focus on the simple (yet fundamental) model of binary classification. We assume that the underlying parameter, $\mathbf{w}^\star$ possesses a certain structure (sparse, low-rank, block-sparse, etc.), and consider a locally-Lipschitz and convex function $\psi : \mathbb{R}^p \to \mathbb{R}$ which encourages this structure. We introduce the *Extended Margin Maximizer* (EMM) as the solution to the following optimization:

$$\begin{aligned} \min_{\mathbf{w} \in \mathbb{R}^p} \quad & \psi(\mathbf{w}) \\ \text{s.t.} \;\; & y_i(\mathbf{x}_i^T \mathbf{w}) \geq 1, \;\; \text{for } 1 \leq i \leq n. \end{aligned} \tag{9.4}$$

It is worth noting that the condition on the separability of the data set is crucial for the optimization program (9.4) to have a feasible point.

Remark 13. *It can be shown that when $\psi(\cdot)$ is absolutely scalable*[1]*, the EMM can be found by solving the following equivalent optimization program,*

$$\max_{\mathbf{w}\in\mathbb{R}^d} \frac{\min\limits_{1\le i\le n} y_i(\mathbf{x}_i^T\mathbf{w})}{\psi(\mathbf{w})} = \max_{\mathbf{w}\in\mathbb{R}^d} \frac{\min\limits_{1\le i\le n} y_i(\mathbf{x}_i^T\mathbf{w})}{\|\mathbf{w}\|} \times \frac{\|\mathbf{w}\|}{\psi(\mathbf{w})}. \tag{9.5}$$

The first multiplicative term on the right indicates the margin associated with the separator $\mathbf{w}$*, and the second term,* $\frac{\|\mathbf{w}\|}{\psi(\mathbf{w})}$ *takes into account the structure of the model. Hence, we refer to the objective function in the optimization* (9.5) *as the extended margin, and the solution to this optimization is called the extended margin maximizer (EMM).*

In this chapter, we study the linear asymptotic regime in which the problem dimensions $p,\ n$ grow to infinity at a proportional rate, $\delta := \frac{p}{n} > 0$. Our main result characterizes the performance of the solution of (9.4), $\hat{\mathbf{w}}$, in terms of the ratio, δ, and the signal strength, $\kappa := \frac{\|\mathbf{w}^\star\|}{\sqrt{p}}$. We assume that the data points, $\{\mathbf{x}_i\}_{i=1}^n$, are drawn independently from the Gaussian distribution. Our main result characterizes the performance of the resulting estimator through the solution of a system of five nonlinear equations with five unknowns. In particular, as an application of our main result, we can accurately predict the generalization error of the resulting estimator.

9.3 Main Results

In this section, we present our main result, that is the characterization of the performance of the extended margin maximizers. Our results are represented in terms of a summary functional, $c_t(\cdot,\cdot)$, which incorporates the information about the underlying model. (Recall that this function has been defined earlier in Chapter 7.)

Definition 12. *For the parameter $t > 0$, the function $c_t : \mathbb{R}\times\mathbb{R}_+ \to \mathbb{R}_+$ is defined as,*

$$c_t(s,r) = \mathbb{E}\big[(1 - tsZ_1Y - rZ_2)_+^2\big], \tag{9.6}$$

where $Z_1, Z_2 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, and $Y \sim R_{AD}(\rho(tZ_1))$.

As discussed in Theorem 10 in Chapter 7, in the Gaussian setting, the data is separable iff,

[1] A function $f : \mathbb{R}^d \to \mathbb{R}$ is absolutely scalable when,

$$\forall \mathbf{v}\in\mathbb{R}^d, \forall\alpha\in\mathbb{R},\ \ f(\alpha\mathbf{v}) = |\alpha| f(\mathbf{v}).$$

All ℓ_p norms, for example, are absolutely scalable.

$$\delta > \delta^* = \delta^*(\kappa) := \inf_{s,r\geq 0} \frac{c_\kappa(s,r)}{r^2} \ . \tag{9.7}$$

It is worth emphasizing again that this condition, which is simply the condition on separability of the data set $\mathcal{D}$, does not depend on the choice of the potential function $\psi(\cdot)$.

A nonlinear system of equations

Our main result in the next section precisely characterizes the performance of EMM in terms of a system of 5 nonlinear equations with 5 unknowns, $(\alpha, \sigma, \beta, \gamma, \tau)$, defined as follows,

$$\begin{cases} \frac{1}{p}\,\mathbb{E}\left[\mathbf{w}^{\star T}\mathbf{P}\right] = \alpha\kappa^2, \\ \frac{1}{p}\,\mathbb{E}\left[\mathbf{h}^{T}\mathbf{P}\right] = \sqrt{\frac{c_\kappa(\alpha,\sigma)}{\delta}}, \\ \frac{1}{p}\,\mathbb{E}\,\|\mathbf{P}\|^2 = \alpha^2\kappa^2 + \sigma^2, \\ \frac{\partial c_\kappa(\alpha,\sigma)}{\partial \alpha} = \frac{2\kappa^2\gamma}{\beta}\sqrt{c_\kappa(\alpha,\sigma)}, \\ \frac{\partial c_\kappa(\alpha,\sigma)}{\partial \sigma} = \frac{2\sqrt{c_\kappa(\alpha,\sigma)}}{\beta\tau}, \end{cases} \tag{9.8}$$

where $\mathbf{P}$ is defined as,

$$\mathbf{P} = \mathrm{Prox}_{\sigma\tau\psi(\cdot)}\left((\alpha - \sigma\tau\gamma)\mathbf{w}^\star + \beta\sigma\tau\sqrt{\delta}\mathbf{h}\right) . \tag{9.9}$$

Remark 14. *The first three equations in the nonlinear system* (9.8) *capture the role of the potential function via its proximal operator. When $\psi(\cdot)$ is separable, these functions can further be reduced to the proximal operator of a real-valued function. For instance, when $\psi(\cdot) = \|\cdot\|_1$, the proximal operator is simply equivalent to applying the well known shrinkage (defined as $\eta(x,t) = \frac{x}{|x|}(|x| - t)_+$) on each entry. For more information on the proximal operators, please refer to [104].*

Asymptotic performance of EMM

We are now ready to present the main result of the chapter. Theorem 13 characterizes the asymptotic behavior of EMM, that is the solution to the optimization program (9.4). It connects the performance of EMM to the solution of the nonlinear system of equations (9.8), and informally states that,

$$\hat{\mathbf{w}} \xrightarrow{D} \Gamma(\mathbf{w}^\star, \mathbf{h}), \text{ as } p \to \infty, \tag{9.10}$$

where $\mathbf{h} \in \mathbb{R}^p$ has standard normal entries, and $\Gamma : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^p$ is defined as,

$$\Gamma(\mathbf{v}_1, \mathbf{v}_2) = \text{Prox}_{\bar{\sigma}\bar{\tau}\psi(\cdot)}\big((\bar{\alpha} - \bar{\sigma}\bar{\tau}\bar{\gamma})\mathbf{v}_1 + \bar{\beta}\bar{\sigma}\bar{\tau}\sqrt{\delta}\mathbf{v}_2\big), \tag{9.11}$$

where $(\bar{\alpha}, \bar{\sigma}, \bar{\beta}, \bar{\gamma}, \bar{\tau})$ is the solution to the nonlinear system (9.8).

Theorem 13. *Let $\hat{\mathbf{w}}$ be the solution of the EMM optimization* (9.4)*, where for $i = 1, 2, \ldots, n$, $\mathbf{x}_i$ has the multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}, \frac{1}{p}\mathbf{I}_p)$, and $y_i \sim \text{RAD}(\rho(\mathbf{x}_i^T\mathbf{w}^\star))$, and $\mathbf{w}^\star$ is drawn from a distribution Π with $\kappa = \frac{\|\mathbf{w}^\star\|}{\sqrt{p}}$. As $n, p \to \infty$ at a fixed overparameterization ratio $\delta = \frac{p}{n} > \delta^*(\kappa)$, then,*

(i) The nonlinear system (9.8) *has a unique solution $(\bar{\alpha}, \bar{\sigma}, \bar{\beta}, \bar{\gamma}, \bar{\tau})$.*

(ii) For any locally-Lipschitz function $F : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$, we have,

$$F(\hat{\mathbf{w}}, \mathbf{w}^\star) \xrightarrow{P} \mathbb{E}[F(\Gamma(\mathbf{w}, \mathbf{h}), \mathbf{w})], \tag{9.12}$$

where $\mathbf{h} \in \mathbb{R}^p$ has standard normal entries, $\mathbf{w} \sim \Pi$ is independent of $\mathbf{h}$, and the function $\Gamma(\cdot, \cdot)$ is defined in (9.11)*.*

The detailed proof of this result is deferred to Section 9.6. In short, we introduce dual variables and write down the Lagrangian which contains a bilinear form with respect to a matrix with i.i.d. Gaussian entries. Exploiting the CGMT framework, we then analyze the nearly-separable auxiliary optimization to find its optimal value, and show that the nonlinear system (9.8) corresponds to its optimality condition.

Remark 15. *The result in Theorem 13 is stated for a general locally-Lipschitz function $F(\cdot, \cdot)$. To evaluate a specific performance measure, one can appeal to this theorem with an appropriate choice of F. As an example, the function $F(\mathbf{u}, \mathbf{v}) = \frac{1}{p}\|\mathbf{u} - \mathbf{v}\|^2$ gives the mean-squared error (MSE).*

Generalization error

Theorem 13 can be utilized to derive useful information on the performance of the classifier. In fact, using this theorem, one can show that the parameters $\bar{\alpha}$ and $\bar{\sigma}$, respectively, correspond to the correlation (to the underlying parameter) and the mean-squared error of the resulting estimator.

An important measure of performance is the generalization error, which indicates the success of the trained model on unseen data. Here, we compute the generalization error of the EMM classifier. We do so by appealing to the result of Theorem 13.

Definition 13. *The generalization error for a binary classifier with parameter* $\hat{\mathbf{w}}$ *is defined as,*

$$GE_{\hat{\mathbf{w}}} = \mathbb{P}_{\mathbf{x}}\{SIGN(\mathbf{x}^T\hat{\mathbf{w}}) \neq SIGN(\mathbf{x}^T\mathbf{w}^\star)\}, \tag{9.13}$$

where the probability is computed with respect to the distribution of the test data.

It can be shown that when the distribution of the test data is rotationally invariant (e.g., Gaussian, uniform dist. on the unit-sphere), GE only depends on the angle between $\hat{\mathbf{w}}$ and $\mathbf{w}^\star$. The following proposition provides sharp asymptotic results on the generalization error of the EMM classifier.

Proposition 4 (Generalization Error)**.** *Let* $\hat{\mathbf{w}}$ *be the EMM classifier defined in Section 9.2. Assume* $\delta > \delta^*$*, and the (test) data is distributed according to the multivariate Gaussian distribution* $\mathcal{N}(\mathbf{0}, \frac{1}{p}\mathbf{I}_p)$*. Then, as* $p \to \infty$*, we have,*

$$GE_{\hat{\mathbf{w}}} \xrightarrow{P} \frac{1}{\pi} acos\left(\frac{\kappa\bar{\alpha}}{\sqrt{\kappa^2\bar{\alpha}^2 + \bar{\sigma}^2}}\right), \tag{9.14}$$

where $\bar{\alpha}$ *and* $\bar{\sigma}$ *are derived by solving the nonlinear system* (9.8)*.*

Proof. We first note that when the data is normally distributed, the generalization error for $\hat{\mathbf{w}}$ is defined as,

$$GE_{\hat{\mathbf{w}}} = \frac{1}{\pi}\text{acos}\left(\frac{\hat{\mathbf{w}}^T\mathbf{w}^\star}{\|\mathbf{w}^\star\|\,\|\hat{\mathbf{w}}\|}\right). \tag{9.15}$$

We appeal to the result of Theorem 13 with two different functions. Using $F_1(\mathbf{u}, \mathbf{v}) = \frac{1}{p}\mathbf{v}^T\mathbf{u}$ in (9.12) will give,

$$\frac{1}{p}\hat{\mathbf{w}}^T\mathbf{w}^\star \xrightarrow{P} \frac{1}{p}\,\mathbb{E}\left[\mathbf{w}^{\star T}\text{Prox}_{\bar{\sigma}\bar{\tau}\psi(\cdot)}\big((\bar{\alpha} - \bar{\sigma}\bar{\tau}\bar{\gamma})\mathbf{w}^\star + \bar{\beta}\bar{\sigma}\bar{\tau}\sqrt{\delta}\mathbf{h}\big)\right]. \tag{9.16}$$

Since $(\bar{\alpha}, \bar{\sigma}, \bar{\beta}, \bar{\gamma}, \bar{\tau})$ is the solution to the nonlinear system, we can replace the expectation from the first equation in (9.8), which gives the following,

$$\frac{1}{p}\hat{\mathbf{w}}^T\mathbf{w}^\star \xrightarrow{P} \kappa^2\bar{\alpha}. \tag{9.17}$$

Similarly, using the result of Theorem 13 for the measure function $F_2(\mathbf{u}, \mathbf{v}) = \frac{1}{p}\|\mathbf{u}\|^2$, along with the third equation in (9.8) gives,

$$\frac{1}{\sqrt{p}}\|\hat{\mathbf{w}}\| \xrightarrow{P} \sqrt{\kappa^2\bar{\alpha}^2 + \bar{\sigma}^2}\,. \tag{9.18}$$

The proof is the consequence of (9.15), (9.17), and (9.18), along with the continuity of the function $\operatorname{acos}(\cdot)$. □

9.4 EMM for Various Structures

As explained earlier, the potential function $\psi(\cdot)$ is chosen to encourage the structure of the underlying parameter. In this section, we investigate the performance of the EMM classifier for some common structures and the corresponding choices of the potential function.

Max-margin classifier (ℓ_2-EMM)

The ℓ_2-norm regularization is commonly used in machine learning applications to stabilize the model. Here, we study the performance of the EMM classifier when $\psi(\cdot) = \frac{1}{2}\|\cdot\|_2^2$, i.e., the solution to the following optimization program,

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^p} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 \\ \text{s.t. } & y_i(\mathbf{x}_i^T\mathbf{w}) \geq 1, \quad \text{for } 1 \leq i \leq n. \end{aligned} \tag{9.19}$$

The optimization program (9.19) is called the hard-margin SVM and the corresponding solution is the max-margin classifier, as it maximizes the minimum distance (margin) of the data points from the separating hyperplane. As mentioned earlier in Section 9.1, the conventional justification for using such a classifier is that the risk of a classifier is inversely proportional to its margin. The performance of ℓ_2-EMM (9.19), has been earlier analyzed in [95].

When $\psi(\cdot) = \frac{1}{2}\|\cdot\|_2^2$, the proximal operator has the following closed-form,

$$\operatorname{Prox}_{\frac{t}{2}\|\cdot\|^2}(\mathbf{u}) = \frac{1}{1+t}\mathbf{u}. \tag{9.20}$$

By replacing the proximal operator in the nonlinear system (9.8), we can explicitly find two of the variables (β, and γ) and reduce it to the following system of three nonlinear equations in three unknowns,

$$\begin{cases} \sqrt{c_\kappa(\alpha,\sigma)} = \sigma\sqrt{\delta}, \\ \dfrac{\partial c_\kappa(\alpha,\sigma)}{\partial\alpha} = \dfrac{-2\kappa^2\alpha\tau\sigma\delta}{1+\sigma\tau}, \\ \dfrac{\partial c_\kappa(\alpha,\sigma)}{\partial\sigma} = \dfrac{2\sigma\delta}{1+\sigma\tau}. \end{cases} \tag{9.21}$$

Sparse classifier (ℓ_1-EMM)

In today's machine learning applications, typically the number of available features, p, is overwhelmingly large. To reduce the risk of overfitting in such settings, feature selection methods are often performed to exclude irrelevant variables from the model [71]. Adding an ℓ_1 penalty is the most popular approach for feature selection.

As a natural consequence of our main result in Theorem 13, here we analyze the asymptotic performance of EMM when the potential function is the ℓ_1 norm, and evaluate its success on the unseen data (i.e., the test error) when the underlying parameter, $\mathbf{w}^\star$, is sparse.

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^p} \quad & \|\mathbf{w}\|_1 \\ \text{s.t. } & y_i(\mathbf{x}_i^T\mathbf{w}) \geq 1, \quad \text{for } 1 \leq i \leq n. \end{aligned} \tag{9.22}$$

In this case, the proximal operator of the potential function ($\|\cdot\|_1$) is basically equivalent to applying the soft-thresholding operator, on each entry, i.e.,

$$\text{Prox}_{t\|\cdot\|_1}(\mathbf{u}) = \eta(\mathbf{u}, t), \tag{9.23}$$

where $\eta(x,t) := \frac{x}{|x|}(|x|-t)_+$ is the soft-thresholding operator. It is worth noting that the analysis presented here is similar to our analysis in Section 8.5 of the previous chapter. Here, for a sparsity factor $s \in (0,1]$, we assume the entries of $\mathbf{w}^\star$ are sampled i.i.d. from the following distribution,

$$\Pi_s(w) = (1-s)\cdot\delta_0(w) + s\cdot\Big(\frac{\phi(\frac{w}{\frac{\kappa}{\sqrt{s}}})}{\frac{\kappa}{\sqrt{s}}}\Big), \tag{9.24}$$

where $\delta_0(\cdot)$ is the Dirac delta function, and $\phi(t) := \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}}$ is the density of the standard normal random variable. This means that each of the entries of $\mathbf{w}^\star$ are zero with probability $1-s$, and the nonzero entries have independent Gaussian distribution with variance $\frac{\kappa^2}{s}$. Having this assumption, we can further simplify the first three equations in the nonlinear system (9.8), and present them in terms of Q-functions. To streamline our representation, we introduce the following proxies,

$$t_1 = \frac{\sigma\tau}{\sqrt{\frac{\kappa^2}{s}(\alpha-\sigma\tau\gamma)^2 + \beta^2\sigma^2\tau^2\delta}}, \quad t_2 = \frac{1}{\beta\sqrt{\delta}}. \tag{9.25}$$

We also define the function $\chi : \mathbb{R} \to \mathbb{R}_+$ as,

$$\begin{aligned}\chi(t) &= \mathbb{E}\left[(Z-t)_+^2\right], \quad Z \sim \mathcal{N}(0,1) \\ &= Q(t)(1+t^2) - t\phi(t),\end{aligned} \tag{9.26}$$

where $Q(t) := \int_t^\infty \phi(x)dx$ denotes the tail distribution of a standard normal random variable. We are now able to simplify the first three equations in (9.8) and derive the following nonlinear system,

$$\begin{cases} Q(t_1) = \frac{\alpha}{2(\alpha-\sigma\tau\gamma)}, \\ s\cdot Q(t_1) + (1-s)\cdot Q(t_2) = \frac{\sqrt{c_\kappa(\alpha,\sigma)}}{2\beta\sigma\tau\delta}, \\ \frac{s}{t_1^2}\cdot\chi(t_1) + \frac{(1-s)}{t_2^2}\cdot\chi(t_2) = \frac{\kappa^2\alpha^2}{2\sigma^2\tau^2} + \frac{1}{2\tau^2}, \\ \frac{\partial c_\kappa(\alpha,\sigma)}{\partial\alpha} = \frac{2\kappa^2\gamma}{\beta}\sqrt{c_\kappa(\alpha,\sigma)}, \\ \frac{\partial c_\kappa(\alpha,\sigma)}{\partial\sigma} = \frac{2\sqrt{c_\kappa(\alpha,\sigma)}}{\beta\tau}. \end{cases} \tag{9.27}$$

The nonlinear system (9.27) can be solved via numerical methods. For our numerical simulations in Section 9.5, we exploit accelerated fixed-point methods to solve the nonlinear system. Using the result of Lemma 4, we can compute the generalization error.

Another important measure in this setting (when $\mathbf{w}^\star$ is sparse) is the probability of error in support recovery. Let $\Omega \subseteq [p]$ denote the support of $\mathbf{w}^\star$ (i.e. $\Omega = \{j : \mathbf{w}_j^\star \neq 0\}$.) For a predefined threshold $\epsilon > 0$, we form the following estimate of the support,

$$\hat{\Omega}_\epsilon = \{j : 1 \le j \le p, |\hat{\mathbf{w}}_j| > \epsilon\}. \tag{9.28}$$

The following lemma establishes the success in the support recovery:

Lemma 22 (Support Recovery)**.** *For a sparsity factor $s \in (0,1]$, let the entries of $\mathbf{w}^\star$ have distribution Π_s defined in* (9.24)*, and $\hat{\mathbf{w}}$ be the solution to the optimization* (9.22)*. Then, as $p \to \infty$, we have,*

$$\begin{aligned}\lim_{\epsilon\downarrow 0} P_1(\epsilon) &:= \mathbb{P}\left\{j \notin \hat{\Omega}_\epsilon | j \in \Omega\right\} \xrightarrow{P} 1 - 2Q(\bar{t}_1) \\ \lim_{\epsilon\downarrow 0} P_2(\epsilon) &:= \mathbb{P}\left\{j \in \hat{\Omega}_\epsilon | j \notin \Omega\right\} \xrightarrow{P} 2Q(\bar{t}_2)\,,\end{aligned} \tag{9.29}$$

where $\bar{t}_1$ and $\bar{t}_2$ are defined as in (9.25)*, with variables derived from solving the nonlinear system* (9.27)*.*

Binary classifier (ℓ_∞-EMM)

As the last example of structured classifiers, here we study the case where $\mathbf{w}^\star \in \{\pm\}^p$. To encourage this structure, the potential function is chosen to be the ℓ_∞ norm. In linear regression, $\|\cdot\|_\infty$ is used to recover the binary signals, i.e., when $\mathbf{w}^\star \in \{\pm 1\}^p$ [33]. This problem arises in integer programming and has some connections to the Knapsack problem [91]. Here, we consider analyzing the performance of the solution of the following optimization program,

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^p} \quad & \|\mathbf{w}\|_\infty \\ \text{s.t. } & y_i(\mathbf{x}_i^T\mathbf{w}) \geq 1, \quad \text{for } 1 \leq i \leq n. \end{aligned} \tag{9.30}$$

It can be shown that the proximal operator of the ℓ_∞-norm can be derived by projecting the points onto the ℓ_1-ball. We use this connection to present the proximal operator in this case in terms of the soft-thresholding operator $\eta(\cdot,\cdot)$.

For a vector $\mathbf{w}$ whose entries are drawn independently from a distribution Π, we can present the following formula for the proximal operator:

$$\text{Prox}_{tp\|\cdot\|_\infty}(\mathbf{w}) = \mathbf{w} - \text{Prox}_{\lambda\|\cdot\|_1}(\mathbf{w}), \tag{9.31}$$

where $\lambda := \lambda(t)$ is the smallest non-negative number that satisfies,

$$\mathbb{E}\left[|\eta(W,\lambda)|\right] = \mathbb{E}\left[(|W|-\lambda)_+\right] \leq t. \tag{9.32}$$

Here, the expectation is with respect to $W \sim \Pi$. Note that λ is a non-increasing function of t, and $\lambda = 0$ whenever $t \geq \mathbb{E}\,|W|$.

Similar to the case of ℓ_1-EMM, here we can use the closed-form of the proximal operator to simplify the first three equations in the nonlinear system (9.8).

For our numerical simulations in the next section, we have done the computations for three different distributions: (1) The i.i.d. Gaussian distribution, (2) the sparse distribution defined in (9.24), and (3) the uniform binary distribution, $\Pi = \text{Unif}\big(\{\pm 1\}^p\big)$.

9.5 Numerical Simulations

In this section, we investigate the validity of our theoretical results with multiple numerical simulations applied to the three different cases of EMM classifiers elaborated in Section 9.4. For each of the three potential functions discussed in the paper (i.e., ℓ_1, ℓ_2, and ℓ_∞ norms), we perform

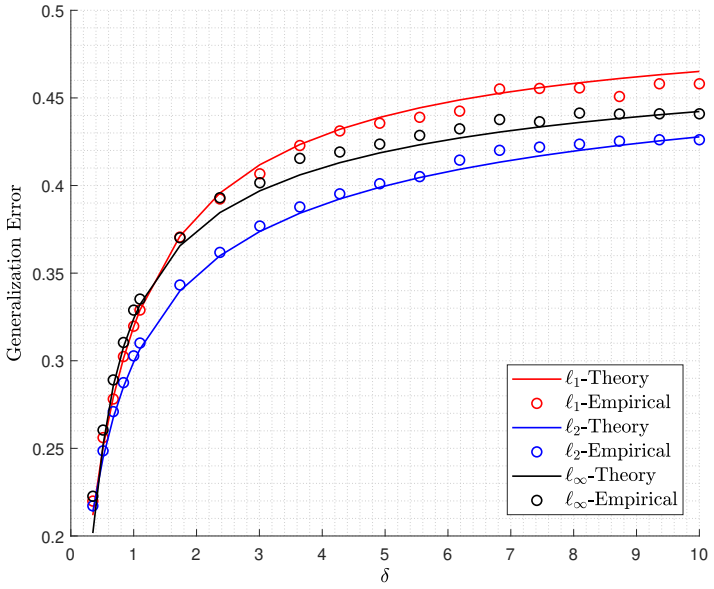


Figure 9.1: Generalization error of the EMM classifier under three potential functions, ℓ_1 norm with the red line (ℓ_1-EMM), ℓ_2 norm with the blue line (ℓ_2-EMM), and ℓ_∞ norm with the black line (ℓ_∞-EMM). The entries of $\mathbf{w}^\star$ are drawn independently from $\mathcal{N}(0, \kappa^2)$ Gaussian distribution.

numerical simulations for three different models on the distribution of $\mathbf{w}^\star$. In other words, we change the distribution of the entries of $\mathbf{w}^\star$ and evaluate the performance of the aforementioned classifiers on each model. It will be observed in our numerical simulations that the appropriate choice of the potential function in the EMM optimization (9.4) has an impact on the generalization error of the resulting classifier. The three different distributions that we choose for the underlying parameter are as follows:

Gaussian: in the first model, we assume that the entries of $\mathbf{w}^\star$ are drawn from a zero-mean Gaussian distribution, $\mathcal{N}(0, \kappa^2)$. In this model, the direction of $\mathbf{w}^\star$ (which indicates the separating hyperplane) is distributed uniformly on the unit sphere. Figure 9.1 gives the generalization error when $\mathbf{w}^\star$ has Gaussian distribution. The solid lines show the theoretical results derived from Theorem 13 and Proposition 4. The circles depict empirical results that are computed by taking the average over 100 trials with $p = 200$ and $\kappa = 2$. Although our theory provides the generalization error in the asymptotic regime, it appropriately matches the result of empirical simulations in our simulations in finite dimensions. It can be observed in this figure that the max-margin classifier (ℓ_2-EMM) outperforms the other two classifiers. We should also note that as the overparameterization ratio, δ, grows, the generalization error increases which indicates that the larger the value of δ, the less

reliable our classifiers become.

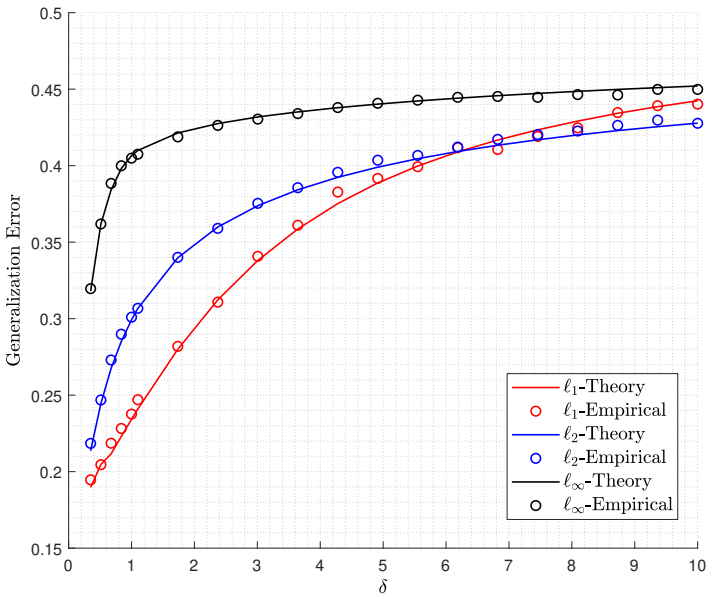


Figure 9.2: Generalization error of the EMM classifier under three potential functions, ℓ_1 norm with the red line (ℓ_1-EMM), ℓ_2 norm with the blue line (ℓ_2-EMM), and ℓ_∞ norm with the black line (ℓ_∞-EMM). The underlying vector $\mathbf{w}^\star$ is s-sparse with the non-zero entries drawn independently from $\mathcal{N}(0, \kappa^2/s)$ Gaussian distribution.

Sparse: here, we assume that the entries of $\mathbf{w}^\star$ are drawn from the sparse distribution represented in (9.24), i.e., each entry is nonzero with probability s, and the nonzero entries have i.i.d. Gaussian distribution with appropriately-defined variance. Figure 9.2 demonstrates the result of the numerical simulations for this model for the three different classifiers of interest. The empirical result is the average over 100 trials with $p = 200$, $s = 0.1$, and $\kappa = 2$. Similarly to the previous case, the empirical results match the theory. Also, it can be observed that the ℓ_1-EMM outperforms the two other classifiers in the regime of δ where the classifiers performs well (i.e. $\delta \lessapprox 6$.) Similarly, we can observe that for large values of δ, all the classifiers perform poorly.

Binary: in this model, the entries of $\mathbf{w}^\star$ are independently drawn from $\{+\kappa, -\kappa\}$, i.e., $\mathbf{w}^\star$ is uniformly chosen on the discrete set $\{\pm\kappa\}^p$. Figure 9.3 shows the result of numerical simulations under this model. Similarly to previous cases, the empirical results ($\kappa = 2$, $p = 200$) match the theory. Also, the ℓ_∞-EMM classifier outperforms the other two classifiers for $\delta < 1$ (which corresponds to the under-parameterized setting). However, the max-margin classifier performs better for larger values of δ.

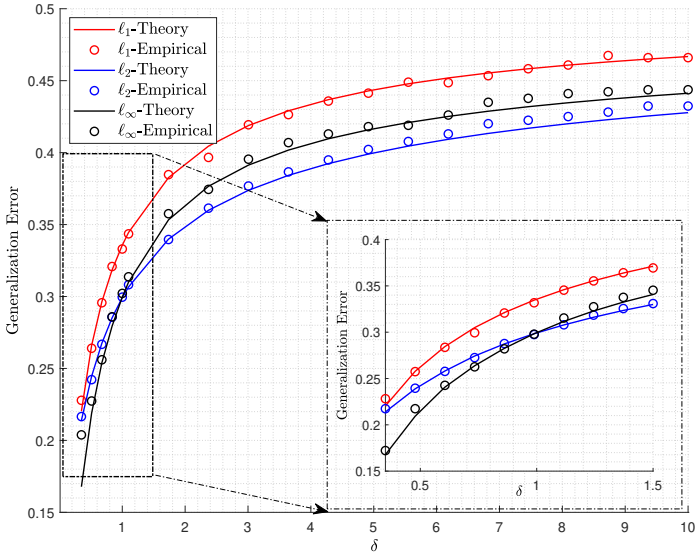


Figure 9.3: Generalization error of the EMM classifier under three potential functions, ℓ_1 norm with the red line (ℓ_1-EMM), ℓ_2 norm with the blue line (ℓ_2-EMM), and ℓ_∞ norm with the black line (ℓ_∞-EMM). The entries of $\mathbf{w}^\star$ are drawn independently from $\kappa * \textsc{Rad}(0.5)$ Rademacher distribution.

9.6 Proof of Theorem 13

Here, we present the proof of the main result of this chapter. Recall that the extended margin maximizer is defined as the solution to the following optimization program,

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^p} \quad & \psi(\mathbf{w}) \\ \text{s.t.} \quad & y_i(\mathbf{x}_i^T\mathbf{w}) \geq 1, \quad \text{for } 1 \leq i \leq n. \end{aligned} \tag{9.33}$$

Theorem 13 provides a precise characterization on the performance of this optimization program in the asymptotic regime, where $n, p \to \infty$ at a fixed ratio $\delta := n/p$. We assume that the data points are drawn independently from the multivariate Gaussian distribution, i.e., $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \frac{1}{p}\mathbb{I}_p)$.

For our analysis, we utilize the CGMT framework (see Appendix A.1), which will provide us with a nearly-separable optimization program that has the same performance as (9.33). To simplify the presentation, we are breaking down the proof into the following three main steps:

1. Finding the auxiliary optimization: By introducing dual variables, we present the optimization (9.33) as a bilinear form with respect to a Gaussian matrix. Consequently, we use the result of Lemma 32 to find the auxiliary optimization.

2. Analyzing the auxiliary optimization: The first step provides a nearly-separable optimization. The purpose of this step is to simplify this optimization and present it in terms of an optimization program with respect to scalar variables.

3. Optimality condition of the auxiliary optimization: By taking the derivatives with respect to various scalars, we present the first-order optimality condition on the solution of the (simplified) auxiliary optimization. Further simplification gives the nonlinear system (9.8).

It is worth noting that the steps explained above resemble our proof in the previous chapter.

We explain each of the three steps in more details in the following subsections.

Finding the auxiliary optimization

The following lemma presents the auxiliary optimization associated with the EMM optimization (9.33).

Lemma 23. *Let* $\hat{\mathbf{w}}$ *be the solution to the optimization* (9.33)*. Consider the following optimization:*

$$\begin{aligned} \min_{\substack{\alpha\in\mathbb{R} \\ \tilde{\mathbf{w}}\in\mathbb{R}^p \\ \tilde{\mathbf{w}}\perp\mathbf{w}^\star}} \ & \frac{1}{p}\psi(\alpha\mathbf{w}^\star+\tilde{\mathbf{w}}) \\ & s.t. \quad \frac{1}{p}(\mathbf{h}^T\tilde{\mathbf{w}})^2 \geq n\cdot c_\kappa\big(\alpha, \frac{\|\tilde{\mathbf{w}}\|}{\sqrt{p}}\big), \end{aligned} \tag{9.34}$$

where $\mathbf{h}\in\mathbb{R}^p$ *has i.i.d. standard normal entries. Assume* $(\bar{\alpha},\bar{\tilde{\mathbf{w}}})\in\mathbb{R}\times\mathbb{R}^p$ *be the solution to this optimization program. Then, as* $p\to\infty$*, we have:*

$$\big\|\hat{\mathbf{w}}-(\bar{\alpha}\mathbf{w}^\star+\bar{\tilde{\mathbf{w}}})\big\| \xrightarrow{P} 0. \tag{9.35}$$

Proof. In order to apply the CGMT, we need to have a min-max optimization. Introducing the Lagrange variable, $\boldsymbol{\lambda} := \Big[\lambda_1,\lambda_2,\ldots,\lambda_n\Big]^T \in\mathbb{R}^n_+$, we can rewrite the optimization program as follows,

$$\min_{\mathbf{w}\in\mathbb{R}^p}\max_{\boldsymbol{\lambda}\in\mathbb{R}^n_+} \frac{1}{p}\psi(\mathbf{w})+\frac{1}{n}\sum_{i=1}^n\lambda_i\big(1-y_i(\mathbf{x}_i^T\mathbf{w})\big). \tag{9.36}$$

Note that the scaling has been performed in such a way that all the terms in the objective be of constant order. We define the matrix $\mathbf{H}\in\mathbb{R}^{n\times p}$ as,

$$\mathbf{H} := -\sqrt{p} \cdot \begin{bmatrix} -\mathbf{x}_1^T - \\ -\mathbf{x}_2^T - \\ \vdots \\ -\mathbf{x}_n^T - \end{bmatrix}. \tag{9.37}$$

Based on the assumption on the distribution of data points, this matrix has i.i.d. standard normal $\mathcal{N}(0,1)$ entries. To ease the notation, we also define a new variable $\bar{\boldsymbol{\lambda}} = \boldsymbol{\lambda} \odot \mathbf{y}$ (i.e., $\bar{\lambda}_i = \lambda_i y_i$) and reformulate the optimization (9.36) as,

$$\min_{\mathbf{w}\in\mathbb{R}^p} \max_{\substack{\bar{\boldsymbol{\lambda}}\in\mathbb{R}^n \\ \bar{\lambda}_i y_i \geq 0}} \frac{1}{p}\psi(\mathbf{w}) + \frac{1}{n}\bar{\boldsymbol{\lambda}}^T\mathbf{y} + \frac{1}{n\sqrt{p}}\bar{\boldsymbol{\lambda}}^T\mathbf{H}\mathbf{w}. \tag{9.38}$$

We proceed by analyzing the optimization program (9.38). In order to apply the CGMT, we need an additive bilinear form that is statistically independent of other functions that appear in the objective. Note that the label vector $\mathbf{y} \in \{\pm 1\}^n$ is a random vector that depends on $\mathbf{H}\mathbf{w}^\star$, as $\mathbf{y} = \textsc{Rad}\big(\rho(-\frac{1}{\sqrt{p}}\mathbf{H}\mathbf{w}^\star)\big)$. Therefore, to remove this independence between $\mathbf{y}$ and the bilinear form, we use the projection onto $\mathbf{w}^\star$. Let $\mathbf{P}$ be the matrix of orthogonal projection onto $\text{span}(\mathbf{w}^\star)$, i.e., $\mathbf{P} = \frac{1}{\|\mathbf{w}^\star\|^2}\mathbf{w}^\star\mathbf{w}^{\star T}$, and $\mathbf{P}^\perp$ be its orthogonal complement, $\mathbf{P}^\perp = \mathbf{I}_p - \mathbf{P}$. We use these projection matrices to decompose the Gaussian matrix $\mathbf{H}$ as $\mathbf{H} = \mathbf{H}_1 + \mathbf{H}_2$ with $\mathbf{H}_1 := \mathbf{H}\times\mathbf{P}$, and $\mathbf{H}_2 := \mathbf{H}\times\mathbf{P}^\perp$. This gives the following equivalent optimization,

$$\min_{\mathbf{w}\in\mathbb{R}^p} \max_{\substack{\bar{\boldsymbol{\lambda}}\in\mathbb{R}^n \\ \bar{\lambda}_i y_i \geq 0}} \frac{1}{p}\psi(\mathbf{w}) + \frac{1}{n}\bar{\boldsymbol{\lambda}}^T\mathbf{y} + \frac{1}{n\sqrt{p}}\bar{\boldsymbol{\lambda}}^T\mathbf{H}_1\mathbf{w} + \frac{1}{n\sqrt{p}}\bar{\boldsymbol{\lambda}}^T\mathbf{H}_2\mathbf{w}. \tag{PO}$$

It is worth noting that the projections of a Gaussian matrix (or vector) onto orthogonal subspaces are statistically independent. Also, the label vector $\mathbf{y}$ would be independent of $\mathbf{H}_2$ since,

$$\mathbf{y} = \textsc{Rad}\big(\rho(-\frac{1}{\sqrt{p}}\mathbf{H}\mathbf{w}^\star)\big) = \textsc{Rad}\big(\rho(-\frac{1}{\sqrt{p}}\mathbf{H}\mathbf{P}\mathbf{w}^\star)\big) = \textsc{Rad}\big(\rho(-\frac{1}{\sqrt{p}}\mathbf{H}_1\mathbf{w}^\star)\big), \tag{9.39}$$

where we used $\mathbf{P}\mathbf{w}^\star = \mathbf{w}^\star$. Therefore, all the additive terms in the objective function of (PO) except the last one are independent of $\mathbf{H}_2$. Also, the objective function is convex with respect to $\mathbf{w}$ and concave(linear) with respect to $\bar{\boldsymbol{\lambda}}$. In order to apply the CGMT framework, we only need an extra condition which is restricting the feasibility sets of $\mathbf{w}$ and $\bar{\boldsymbol{\lambda}}$ to be compact and convex. We can introduce some artificial convex and bounded sets $\mathcal{S}_{\mathbf{w}}$ and $\mathcal{S}_{\bar{\boldsymbol{\lambda}}}$, and perform the optimization

over these sets. Note that these sets can be chosen large enough such that they do not affect the optimization itself. For simplicity, in our arguments here we ignore the condition on the compactness of the feasible sets and apply the CGMT whenever the variables are defined on a convex domain.

The optimization program (PO) is suitable to be analyzed via the CGMT as the conditions are all satisfied. Having identified (PO) as the primary optimization, it is straightforward to write its corresponding auxiliary optimization (AO) [as in (A.1), c.f. Appendix A.1]. The Auxiliary Optimization (AO) can be written as follows,

$$\min_{\mathbf{w}\in\mathbb{R}^p} \max_{\substack{\bar{\boldsymbol{\lambda}}\in\mathbb{R}^n \\ \bar{\lambda}_i y_i \geq 0}} \frac{1}{p}\psi(\mathbf{w}) + \frac{1}{n}\bar{\boldsymbol{\lambda}}^T\mathbf{y} + \frac{1}{n\sqrt{p}}\bar{\boldsymbol{\lambda}}^T\mathbf{H}_1\mathbf{w} + \frac{1}{n\sqrt{p}}\big(\|\bar{\boldsymbol{\lambda}}\|\,\mathbf{h}^T\mathbf{P}^{\perp}\mathbf{w} + \bar{\boldsymbol{\lambda}}^T\mathbf{g}\,\|\mathbf{P}^{\perp}\mathbf{w}\|\big), \tag{AO}$$

where $\mathbf{h} \in \mathbb{R}^p$ and $\mathbf{g} \in \mathbb{R}^n$ have i.i.d. standard normal entries. Next, we decompose $\mathbf{w}$ as $\mathbf{w} := \mathbf{P}\mathbf{w} + \mathbf{P}^{\perp}\mathbf{w} = \alpha\mathbf{w}^\star + \tilde{\mathbf{w}}$, where $\alpha \in \mathbb{R}$, and $\tilde{\mathbf{w}} \in \mathbb{R}^p$ is such that $\tilde{\mathbf{w}} \perp \mathbf{w}^\star$. We also define the vector $\mathbf{q} := -\frac{1}{\kappa\sqrt{p}}\mathbf{H}\mathbf{w}^\star$. Note that since $\|\mathbf{w}\| = \kappa\sqrt{p}$, the entries of $\mathbf{q}$ have standard normal distribution. Therefore, we have the following equivalent optimization,

$$\min_{\substack{\alpha\in\mathbb{R} \\ \tilde{\mathbf{w}}\in\mathbb{R}^p \\ \tilde{\mathbf{w}}\perp\mathbf{w}^\star}} \max_{\substack{\bar{\boldsymbol{\lambda}}\in\mathbb{R}^n \\ \bar{\lambda}_i y_i \geq 0}} \frac{1}{p}\psi(\alpha\mathbf{w}^\star + \tilde{\mathbf{w}}) + \frac{1}{n}\bar{\boldsymbol{\lambda}}^T\mathbf{y} - \frac{\alpha\kappa}{n}\bar{\boldsymbol{\lambda}}^T\mathbf{q} + \frac{1}{n\sqrt{p}}\big(\|\bar{\boldsymbol{\lambda}}\|\,\mathbf{h}^T\tilde{\mathbf{w}} + \bar{\boldsymbol{\lambda}}^T\mathbf{g}\,\|\tilde{\mathbf{w}}\|\big)\,. \tag{9.40}$$

Proceeding onward, we solve the inner optimization ($\max_{\bar{\boldsymbol{\lambda}}}$) with respect to the direction of $\bar{\boldsymbol{\lambda}}$. We have:

$$\max_{\substack{\bar{\boldsymbol{\lambda}}\in\mathbb{R}^n \\ \bar{\lambda}_i y_i \geq 0}} \frac{1}{n}\bar{\boldsymbol{\lambda}}^T\mathbf{y} - \frac{\alpha\kappa}{n}\bar{\boldsymbol{\lambda}}^T\mathbf{q} + \frac{1}{n\sqrt{p}}\big(\|\bar{\boldsymbol{\lambda}}\|\,\mathbf{h}^T\tilde{\mathbf{w}} + \bar{\boldsymbol{\lambda}}^T\mathbf{g}\,\|\tilde{\mathbf{w}}\|\big) = \max_{\substack{\bar{\boldsymbol{\lambda}}\in\mathbb{R}^n \\ \bar{\lambda}_i y_i \geq 0}} \frac{\|\bar{\boldsymbol{\lambda}}\|}{\sqrt{n}}\Big(\frac{1}{\sqrt{np}}\mathbf{h}^T\tilde{\mathbf{w}} + \frac{1}{\sqrt{n}}\,\|\boldsymbol{\mu}\|\Big) \tag{9.41}$$

$$\text{s.t. } \mu_i = \Big(1 - \alpha\kappa q_i y_i + \frac{\|\tilde{\mathbf{w}}\|}{\sqrt{p}} g_i y_i\Big)_+,\ 1 \leq i \leq n.$$

Recall that the function $c_\kappa : \mathbb{R}\times\mathbb{R}_+ \to \mathbb{R}$ is defined (c.f. Definition 12) as follows:

$$c_\kappa(t_1, t_2) = \mathbb{E}\left(1 - \kappa t_1 Z_1 Y + t_2 Z_2\right)_+^2, \tag{9.42}$$

where Z_1, Z_2 are independent standard normal random variables, and $Y \sim \textsc{Rad}\big(\rho(\kappa Z_1)\big)$. Therefore, we have $\mathbb{E}\,\mu_i^2 = c_\kappa(\alpha, \frac{\|\tilde{\mathbf{w}}\|}{\sqrt{p}})$, and using the SLLN as $p, n \to \infty$, we can replace $\|\boldsymbol{\mu}\|$ with

$\sqrt{n \cdot c_\kappa(\alpha, \frac{\|\tilde{\mathbf{w}}\|}{\sqrt{p}})}$ due to the almost sure convergence. Introducing the positive variable $\beta = \frac{\|\bar{\lambda}\|}{\sqrt{n}}$, we have the following reformulation of the auxiliary optimization,

$$\min_{\substack{\alpha\in\mathbb{R}\\ \tilde{\mathbf{w}}\in\mathbb{R}^p\\ \tilde{\mathbf{w}}\perp\mathbf{w}^\star}} \max_{\beta\geq 0} \ \frac{1}{p}\psi(\alpha\mathbf{w}^\star+\tilde{\mathbf{w}}) + \frac{\beta}{\sqrt{np}}\mathbf{h}^T\tilde{\mathbf{w}} + \beta\cdot\sqrt{c_\kappa\big(\alpha, \frac{\|\tilde{\mathbf{w}}\|}{\sqrt{p}}\big)} \ . \tag{9.43}$$

We can write the inner maximization (with respect to β) as a constraint for the optimization, which gives the same formulation as (9.34), i.e.,

$$\begin{aligned}\min_{\substack{\alpha\in\mathbb{R}\\ \tilde{\mathbf{w}}\in\mathbb{R}^p\\ \tilde{\mathbf{w}}\perp\mathbf{w}^\star}} \ & \frac{1}{p}\psi(\alpha\mathbf{w}^\star+\tilde{\mathbf{w}}) \\ \text{s.t.} \ & \frac{1}{p}(\mathbf{h}^T\tilde{\mathbf{w}})^2 \geq n\cdot c_\kappa\big(\alpha, \frac{\|\tilde{\mathbf{w}}\|}{\sqrt{p}}\big) \ .\end{aligned} \tag{9.44}$$

Using the result of Lemma 32, we have that when the solution of the primary optimization converges as the problem dimensions grow ($p\to\infty$), the solution of the auxiliary optimization converges to the same set (point). This concludes the proof. □

Analyzing the auxiliary optimization

In this section, we analyze the performance of the refined version of the auxiliary optimization in (9.43). Although this optimization program is (nearly) separable, it is still a high-dimensional optimization. Ideally, one would like to simplify this optimization to obtain another optimization program in lower dimensions (with respect to a few scalar variables) where the performance can be numerically computed. To do so, in this section we exploit some tools from convex analysis along with some tricks from calculus to further simplify the optimization program (9.43).

The goal is to express the final result in terms of the *expected Moreau envelope* of the regularization function. To better understand the behavior of the solution in (9.43), we first introduce some new variables, $\mathbf{u}, \mathbf{v}\in\mathbb{R}^p$, and $\gamma\in\mathbb{R}$ and write the optimization as follows,

$$\min_{\substack{\alpha\in\mathbb{R}\\ \mathbf{u},\tilde{\mathbf{w}}\in\mathbb{R}^p}} \max_{\substack{\beta\geq 0,\gamma\\ \mathbf{v}\in\mathbb{R}^p}} \frac{1}{p}\psi(\mathbf{u}) + \frac{\beta}{\sqrt{np}}\mathbf{h}^T\tilde{\mathbf{w}} + \beta\cdot\sqrt{c_\kappa\big(\alpha, \frac{\|\tilde{\mathbf{w}}\|}{\sqrt{p}}\big)} + \frac{1}{p}\mathbf{v}^T\big(\mathbf{u}-\alpha\mathbf{w}^\star-\tilde{\mathbf{w}}\big) + \frac{\gamma}{p}\mathbf{w}^{\star T}\tilde{\mathbf{w}}. \tag{9.45}$$

The variable $\mathbf{u}$ has been introduced to detach the impact of s and $\tilde{\mathbf{w}}$ from $\psi(\cdot)$. The variables $\mathbf{v}$ and γ are Lagrange dual variables to remove the constraints from the optimization. We shall emphasize

again that the normalization has been performed to ensure that all the terms in the objective are of constant order. Next, we would like to solve the minimization with respect to $\tilde{\mathbf{w}}$.

Before continuing our analysis, we need to discuss an important point that would help us in the remaining of this section. It will be observed that in order to simplify the optimization, we would like to flip the orders of min and max in the (AO) optimization. Since the objective function in the auxiliary optimization is not convex-concave, we cannot appeal to the Sion's min-max theorem in order to flip min and max. However, it has been shown (see Appendix A in [129]) that flipping the orders of min and max in the (AO) is allowed in the asymptotic setting. This is mainly due to the fact that the original (PO) optimization was convex-concave with respect to its variables, and as the CGMT suggests (AO) and (PO) are tightly related in the asymptotic setting; hence, flipping the order of optimizations in (AO) is justified whenever such a flipping is allowed in the (PO). We appeal to this result to flip the orders of min and max when needed.

Next, we solve the optimization with respect to the direction of $\tilde{\mathbf{w}}$. Defining $\sigma := \|\tilde{\mathbf{w}}\| / \sqrt{p}$ and solving the optimization with respect to the direction of $\tilde{\mathbf{w}}$ leads to,

$$\min_{\substack{\sigma\geq 0,\alpha\\ \mathbf{u}\in\mathbb{R}^p}} \max_{\substack{\beta\geq 0,\gamma\\ \mathbf{v}\in\mathbb{R}^p}} \frac{1}{p}\psi(\mathbf{u}) + \beta\cdot\sqrt{c_\kappa(\alpha,\sigma)} + \frac{1}{p}\mathbf{v}^T\left(\mathbf{u}-\alpha\mathbf{w}^\star\right) - \sigma\cdot\left\|\frac{\beta}{\sqrt{n}}\mathbf{h} - \frac{1}{\sqrt{p}}\mathbf{v} + \frac{\gamma}{\sqrt{p}}\mathbf{w}^\star\right\|. \tag{9.46}$$

Consequently, we are considering the maximization with respect to the vector variable $\mathbf{v}\in\mathbb{R}^p$. As seen in (9.46), this variable appears in the last two additive terms in the objective function. To find the optimal value for $\mathbf{v}$, we introduce a new scalar variable $\tau > 0$ [2], which simplifies the optimization by changing $\|\cdot\|$ to $\|\cdot\|^2$. The new optimization would be,

$$\min_{\substack{\sigma\geq 0,\alpha\\ \mathbf{u}\in\mathbb{R}^p}} \max_{\substack{\beta\geq 0,\tau>0,\gamma\\ \mathbf{v}\in\mathbb{R}^p}} \frac{1}{p}\psi(\mathbf{u}) + \beta\cdot\sqrt{c_\kappa(\alpha,\sigma)} + \frac{1}{p}\mathbf{v}^T\left(\mathbf{u}-\alpha\mathbf{w}^\star\right) - \frac{\sigma\tau}{2} - \frac{\sigma}{2\tau}\cdot\left\|\frac{\beta}{\sqrt{n}}\mathbf{h} - \frac{1}{\sqrt{p}}\mathbf{v} + \frac{\gamma}{\sqrt{p}}\mathbf{w}^\star\right\|^2. \tag{9.47}$$

It can be easily checked that the optimization programs (9.46) and (9.47) are equivalent by simply solving the inner optimization with respect to the variable τ. We are now ready to solve the optimization with respect to $\mathbf{v}$. To do so, we continue by making a completion of squares as follows,

[2] The square-root trick: it was first proposed in the analysis of the auxiliary optimization in regularized M-estimators, and the idea is to use the following equivalence (which is derived immediately from AM-GM inequality):

$$\sqrt{x} = \min_{\tau>0}\ \frac{1}{2\tau}x + \frac{\tau}{2}\,,\quad \forall x > 0.$$

$$\frac{1}{p}\mathbf{v}^T\left(\mathbf{u}-\alpha\mathbf{w}^\star\right)-\frac{\sigma}{2\tau}\cdot\left\|\frac{\beta}{\sqrt{n}}\mathbf{h}-\frac{1}{\sqrt{p}}\mathbf{v}+\frac{\gamma}{\sqrt{p}}\mathbf{w}^\star\right\|^2=-\frac{\sigma}{2\tau}\cdot\left\|\frac{\beta}{\sqrt{n}}\mathbf{h}-\frac{1}{\sqrt{p}}\mathbf{v}+\frac{\gamma}{\sqrt{p}}\mathbf{w}^\star+\frac{\tau}{\sigma\sqrt{p}}\mathbf{u}-\frac{\alpha\tau}{\sigma\sqrt{p}}\mathbf{w}^\star\right\|^2$$
$$+\frac{\tau}{2\sigma p}\left\|\mathbf{u}-\alpha\mathbf{w}^\star\right\|^2+\frac{\beta}{\sqrt{np}}\mathbf{u}^T\mathbf{h}+\frac{\gamma}{p}\mathbf{u}^T\mathbf{w}^\star$$
$$-\frac{\alpha\beta\sqrt{\delta}}{p}\mathbf{h}^T\mathbf{w}^\star-\alpha\gamma\kappa^2,$$
$$\boxed{p,n\longrightarrow+\infty}\qquad=-\frac{\sigma}{2\tau}\cdot\left\|\frac{\beta}{\sqrt{n}}\mathbf{h}-\frac{1}{\sqrt{p}}\mathbf{v}+\frac{\gamma}{\sqrt{p}}\mathbf{w}^\star+\frac{\tau}{\sigma\sqrt{p}}\mathbf{u}-\frac{\alpha\tau}{\sigma\sqrt{p}}\mathbf{w}^\star\right\|^2$$
$$+\frac{\tau}{2\sigma p}\left\|\mathbf{u}+\frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}+(\frac{\sigma\gamma}{\tau}-\alpha)\mathbf{w}^\star\right\|^2$$
$$-\frac{\sigma}{2\tau}(\delta\beta^2+\gamma^2\kappa^2),\tag{9.48}$$

where we exploit the fact that, as $p\to\infty$, we can replace $\frac{1}{p}\left\|\mathbf{w}^\star\right\|^2$ and $\frac{1}{p}\left\|\mathbf{h}\right\|^2$ with κ^2 and 1, respectively. Furthermore, we omit the term $\frac{1}{p}\mathbf{h}^T\mathbf{w}^\star=O(\frac{1}{\sqrt{p}})$ as it is negligible compared to other terms in the optimization (which are of constant $O(1)$ orders.) Using the above completion-of-squares, $\mathbf{v}$ is now appearing in only one quadratic term in (9.48). Hence, to maximize the objective, $\mathbf{v}$ chooses itself in such a way that it makes the quadratic term equal to zero. This gives the following optimization,

$$\min_{\substack{\sigma\geq0,\alpha\\ \mathbf{u}\in\mathbb{R}^p}}\max_{\substack{\beta\geq0,\tau>0,\gamma\\ \mathbf{v}\in\mathbb{R}^p}}\beta\cdot\sqrt{c_\kappa(\alpha,\sigma)}-\frac{\sigma\tau}{2}-\frac{\sigma}{2\tau}\left(\delta\beta^2+\gamma^2\kappa^2\right)+\frac{1}{p}\Big[\psi(\mathbf{u})+\frac{\tau}{2\sigma}\left\|\mathbf{u}+\frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}+(\frac{\sigma\gamma}{\tau}-\alpha)\mathbf{w}^\star\right\|^2\Big].\tag{9.49}$$

We now switch the order of min and max (similar to what we did earlier for $\tilde{\mathbf{w}}$) and perform the minimization with respect to $\mathbf{u}$. Using the definition of the Moreau envelope, we can write down this optimization in terms of the Moreau envelope of the potential function. We have,

$$M_{\psi(\cdot)}\big((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^\star-\frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h},\frac{\sigma}{\tau}\big)\;=\;\min_{u\in\mathbb{R}^p}\;\psi(\mathbf{u})+\frac{\tau}{2\sigma}\left\|\mathbf{u}+\frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}+(\frac{\sigma\gamma}{\tau}-\alpha)\mathbf{w}^\star\right\|^2.\tag{9.50}$$

Using the result of Lemma 37 (see Appendix A.2), we have that the Moreau envelope is a Lipschitz function as $\psi(\cdot)$ is Lipschitz. Therefore, we can exploit the Gaussian concentration of Lipschitz functions (see Theorem 5.22 in [141]) which gives,

$$\frac{1}{p}M_{\psi(\cdot)}\big((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^\star-\frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h},\frac{\sigma}{\tau}\big)\xrightarrow{P}\frac{1}{p}\,\mathbb{E}\left[M_{\psi(\cdot)}\big((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^\star-\frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h},\frac{\sigma}{\tau}\big)\right],\ \text{as } p\to\infty. \tag{9.51}$$

We now appeal to Lemma 33 in Appendix A.2 which allows us to replace the Moreau envelope with their expected value due to the convergence we are getting in (9.51). Hence, by replacing the expected value of the Moreau envelope function, we are getting the following optimization, to be analyzed in the next section.

$$\min_{\sigma\geq 0,\alpha}\max_{\substack{\gamma\\ \beta\geq 0,\tau>0}}\ \frac{1}{p}\,\mathbb{E}\left[M_{\psi(\cdot)}\big((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^\star-\frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h},\frac{\sigma}{\tau}\big)\right]+\beta\sqrt{c_\kappa(\alpha,\sigma)}-\frac{\sigma\tau}{2}-\frac{\sigma}{2\tau}(\delta\beta^2+\gamma^2\kappa^2). \tag{9.52}$$

Optimality conditions of the auxiliary optimization

In this section, we conclude the proof of the main result of the paper by showing that (when $\delta>\delta^*$) the optimizer to the scalar optimization (9.52) can be derived by solving the nonlinear system of equations (9.8).

Here, we investigate the optimality condition for the solution of the auxiliary optimization. So far, we simplified the (AO) and after some algebra, we got the scalar optimization (9.52) with respect to five variables. Here, we would like to present the solution to this optimization. Let $C(\alpha,\sigma,\gamma,\beta,\tau)$ denote the objective function in the scalar optimization. In other words, the function C is defined as:

$$C(\alpha,\sigma,\gamma,\beta,\tau)=\frac{1}{p}\,\mathbb{E}\left[M_{\psi(\cdot)}\big((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^\star+\frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h},\frac{\sigma}{\tau}\big)\right]+\beta\sqrt{c_\kappa(\alpha,\sigma)}-\frac{\sigma\tau}{2}-\frac{\sigma}{2\tau}(\delta\beta^2+\gamma^2\kappa^2). \tag{9.53}$$

The following lemma describes the behavior of the function C with respect to its variables.

Lemma 24. *The function $C:\mathbb{R}^5\to\mathbb{R}$ defined in* (9.53) *is (jointly) convex with respect to the variables (α,σ), and (jointly) concave with respect to the variables (γ,β,τ).*

The proof of this lemma is provided in Appendix 9.7. Using the result of Theorem 10 in Chapter 7, the objective function, C, will diverge when $\delta<\delta^*$. For $\delta>\delta^*$, Lemma 24 states that the function C is convex-concave. The following remark indicates that the optimal solution of the optimization problem does not happen at the boundary values.

Remark 16. *We need to show that the optimal solution does not happen at the boundary, i.e., at* $\beta = 0$*, or* $\sigma = 0$*. Taking the derivative with respect to* β *at the objective function in* (9.47)*, we will have* $\frac{\partial}{\partial\beta}|_{\beta=0} = \sqrt{c_\kappa(\alpha,\sigma)} > 0$*. Therefore, the optimal* β *is nonzero. It can also be seen in the same optimization program that when* $\sigma = 0$*,* β *can choose its value arbitrarily large and the optimal value would be* $+\infty$*. Hence, the optimal* σ *is also nonzero as we have a minimization w.r.t.* σ*.*

Let $(\bar{\alpha}, \bar{\sigma}, \bar{\gamma}, \bar{\beta}, \bar{\tau})$ denote the solution to the optimization (9.52). Since the objective function is smooth with respect to its variables and the optimal values do not coincide with the boundaries, its solution must satisfy the first-order optimality condition, i.e., $\nabla C\big(\bar{\alpha}, \bar{\sigma}, \bar{\gamma}, \bar{\beta}, \bar{\tau}\big) = \mathbf{0}_{5\times 1}$. We will show that this would simplify to our system of nonlinear equations (9.8).

We start by setting the derivative with respect to α to zero. We have,

$$\frac{\partial C}{\partial \alpha} = 0 \Rightarrow \frac{1}{p}\,\mathbb{E}\,\big[\frac{\partial}{\partial \alpha} M_{\psi(\cdot)}\big((\alpha - \frac{\sigma\gamma}{\tau})\mathbf{w}^\star + \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}, \frac{\sigma}{\tau}\big)\big] + \frac{\beta}{2\sqrt{c_\kappa(\alpha,\sigma)}} \cdot \frac{\partial c_\kappa(\alpha,\sigma)}{\partial \alpha} \;= 0\,, \quad (9.54)$$

where we used the Leibniz integral rule to bring the derivative inside the expectation. Using the result of Lemma 34, we can write the following,

$$\frac{1}{p}\,\mathbb{E}\,\big[\frac{\partial}{\partial \alpha} M_{\psi(\cdot)}\big((\alpha - \frac{\sigma\gamma}{\tau})\mathbf{w}^\star + \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}, \frac{\sigma}{\tau}\big)\big] = \frac{\kappa^2\alpha\tau}{\sigma} - \kappa^2\gamma - \frac{\tau}{p\sigma}\,\mathbb{E}\,\big[\mathbf{w}^{\star T}\mathrm{Prox}_{\frac{\sigma}{\tau}\psi(\cdot)}((\alpha - \frac{\sigma\gamma}{\tau})\mathbf{w}^\star + \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h})\big]. \quad (9.55)$$

Replacing (9.55) in (9.54) gives the following nonlinear equation,

$$\frac{\tau}{p\sigma}\,\mathbb{E}\,\big[\mathbf{w}^{\star T}\mathrm{Prox}_{\frac{\sigma}{\tau}\psi(\cdot)}((\alpha - \frac{\sigma\gamma}{\tau})\mathbf{w}^\star + \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h})\big] = \frac{\kappa^2\alpha\tau}{\sigma} - \kappa^2\gamma + \frac{\beta}{2\sqrt{c_\kappa(\alpha,\sigma)}} \cdot \frac{\partial c_\kappa(\alpha,\sigma)}{\partial \alpha}\,. \quad (9.56)$$

Next, we find another optimality condition by setting the derivative with respect to β to zero. We have,

$$\frac{\partial C}{\partial \beta} = 0 \Rightarrow \frac{1}{p}\,\mathbb{E}\,\big[\frac{\partial}{\partial \beta} M_{\psi(\cdot)}\big((\alpha - \frac{\sigma\gamma}{\tau})\mathbf{w}^\star + \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}, \frac{\sigma}{\tau}\big)\big] + \sqrt{c_\kappa(\alpha,\sigma)} - \frac{\sigma\delta}{\tau}\beta \;= 0\,. \quad (9.57)$$

Similarly to (9.55), we can compute the expected derivative of the Moreau envelope function by appealing to Lemma 34,

$$\frac{1}{p}\mathbb{E}\left[\frac{\partial}{\partial\beta}M_{\psi(\cdot)}\big((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^{\star}+\frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h},\frac{\sigma}{\tau}\big)\right]=\frac{\beta\sigma\delta}{\tau}-\frac{\sqrt{\delta}}{p}\,\mathbb{E}\left[\mathbf{h}^{T}\mathrm{Prox}_{\frac{\sigma}{\tau}\psi(\cdot)}((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^{\star}+\frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h})\right]. \tag{9.58}$$

Replacing (9.58) in (9.57) will give the following nonlinear equation:

$$\boxed{\frac{1}{p}\,\mathbb{E}\left[\mathbf{h}^{T}\mathrm{Prox}_{\frac{\sigma}{\tau}\psi(\cdot)}((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^{\star}+\frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h})\right]=\sqrt{\frac{c_{\kappa}(\alpha,\sigma)}{\delta}}\,.} \tag{E2}$$

Next, we compute the derivative with respect to γ and set it to zero. We have,

$$\frac{\partial C}{\partial\gamma}=0\Rightarrow\frac{1}{p}\,\mathbb{E}\left[\frac{\partial}{\partial\gamma}M_{\psi(\cdot)}\big((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^{\star}+\frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h},\frac{\sigma}{\tau}\big)\right]-\frac{\sigma\kappa^{2}\gamma}{\tau}\;=0\,, \tag{9.59}$$

$$\frac{1}{p}\,\mathbb{E}\left[\frac{\partial}{\partial\gamma}M_{\psi(\cdot)}\big((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^{\star}+\frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h},\frac{\sigma}{\tau}\big)\right]=\frac{\sigma\kappa^{2}\gamma}{\tau}-\kappa^{2}\alpha-\frac{1}{p}\,\mathbb{E}\left[\mathbf{w}^{\star T}\mathrm{Prox}_{\frac{\sigma}{\tau}\psi(\cdot)}((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^{\star}+\frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h})\right]. \tag{9.60}$$

Replacing (9.60) in (9.59) will give the following nonlinear equation:

$$\boxed{\frac{1}{p}\,\mathbb{E}\left[\mathbf{w}^{\star T}\mathrm{Prox}_{\frac{\sigma}{\tau}\psi(\cdot)}((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^{\star}+\frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h})\right]=\kappa^{2}\alpha\,.} \tag{E1}$$

Also, replacing (E1) in the nonlinear equation (9.56) gives the following nonlinear equation:

$$\boxed{\frac{\partial c_{\kappa}(\alpha,\sigma)}{\partial\alpha}=\frac{2\kappa^{2}\gamma}{\beta}\sqrt{c_{\kappa}(\alpha,\sigma)}\,.} \tag{E4}$$

Next, we take the derivative with respect to σ. We have:

$$\frac{\partial C}{\partial\sigma}=0\Rightarrow\frac{1}{p}\,\mathbb{E}\left[\frac{\partial}{\partial\sigma}M_{\psi(\cdot)}\big((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^{\star}+\frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h},\frac{\sigma}{\tau}\big)\right]+\frac{\beta}{2\sqrt{c_{\kappa}(\alpha,\sigma)}}\frac{\partial c_{\kappa}(\alpha,\sigma)}{\partial\sigma}-\frac{\tau}{2}-\frac{1}{2\tau}(\delta\beta^{2}+\gamma^{2}\kappa^{2})=0\,. \tag{9.61}$$

We use the result of the Lemma 34 to compute the derivative of $M_\psi(\cdot,\cdot)$ with respect to σ. We have,

$$\frac{1}{p}\,\mathbb{E}\,\big[\frac{\partial}{\partial\sigma}M_{\psi(.)}\big((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^\star+\frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h},\frac{\sigma}{\tau}\big)\big]=\frac{1}{2\tau}\big(\gamma^2\kappa^2+\delta\beta^2+\frac{\alpha^2\kappa^2\tau^2}{\sigma^2} \tag{9.62}$$

$$-\frac{\tau^2}{p\sigma^2}\,\mathbb{E}\left\|\mathrm{Prox}_{\frac{\sigma}{\tau}\psi(.)}\big((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^\star+\frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}\big)\right\|^2\big)\,. \tag{9.63}$$

Replacing this into (9.61) will give the following equation,

$$\frac{1}{p}\,\mathbb{E}\left\|\mathrm{Prox}_{\frac{\sigma}{\tau}\psi(.)}\big((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^\star+\frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}\big)\right\|^2=\alpha^2\kappa^2+\frac{\beta\sigma^2}{\tau\sqrt{c_\kappa(\alpha,\sigma)}}\frac{\partial c_\kappa(\alpha,\sigma)}{\partial\sigma}-\sigma^2\,. \tag{9.64}$$

Similarly, by taking the derivative with respect to τ, we have:

$$\boxed{\frac{1}{p}\,\mathbb{E}\left\|\mathrm{Prox}_{\frac{\sigma}{\tau}\psi(.)}\big((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^\star+\frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}\big)\right\|^2=\alpha^2\kappa^2+\sigma^2\,.} \tag{E3}$$

We can now simplify (9.64) to get the following equation:

$$\boxed{\frac{\partial c_\kappa(\alpha,\sigma)}{\partial\sigma}=\frac{2\tau\sqrt{c_\kappa(\alpha,\sigma)}}{\beta}\,.} \tag{E5}$$

Finally, we make a change of variable by replacing τ with $\frac{1}{\tau}$ in the equations (E1), (E2), (E3),(E4), and (E5) will respectively give the desired equations in the system of nonlinear equations (9.8) as the optimality condition on the solution of the optimization (9.52). This concludes the proof.

9.7 Proof of Lemma 24

We first state the following lemma which will be useful in our proof.

Lemma 25. *The function $f(s,r):=\sqrt{c_\kappa(s,r)}$ is (jointly) convex in (s,r).*

Proof. First, note that for $\mathbf{x}\in\mathbb{R}^n$, the function $\mathbf{x}\mapsto\|(\mathbf{x})_+\|$ is a convex function as it can be written as a sup of convex(linear) functions.

$$\|(\mathbf{x})_+\|=\sup_{\substack{\mathbf{u}\in\mathbb{R}^n_+\\ \|\mathbf{u}\|\le 1}}\mathbf{u}^T\mathbf{x}\,.. \tag{9.65}$$

For $n \in \mathbb{N}$, define the function $f_\kappa^{(n)}(s,r)$ as:

$$f_\kappa^{(n)}(s,r) = \frac{1}{\sqrt{n}} \left\| \left(\mathbf{1}_n - s\kappa\mathbf{h}\mathbf{y} + r\mathbf{g}\mathbf{y}\right)_+ \right\|, \tag{9.66}$$

where $\mathbf{1}_n$ denotes the all-one vector, $\mathbf{h}, \mathbf{g} \in \mathbb{R}^n$ have i.i.d. $\mathcal{N}(0,1)$ entries, and $Y \sim \textsc{Rad}\big(\rho(\kappa\mathbf{h})\big)$. It is readily seen that $f_\kappa^{(n)}(s,r)$ is jointly convex with respect to s and r as it is a combination of a convex function and a linear function. Using the LLN, we also have that,

$$f_\kappa^{(n)}(s,r) \xrightarrow{P} f(s,r) = \sqrt{c_\kappa(s,r)}. \tag{9.67}$$

Therefore, $f(s,r)$ is a convex function as it is a point-wise limit of convex functions. □

Consider the objective function in the optimization program 9.47, i.e.,

$$f^{(p)}\big(\alpha,\sigma,\mathbf{u};\gamma,\beta,\tau,\mathbf{v}\big) = \frac{1}{p}\psi(\mathbf{u}) + \beta\cdot\sqrt{c_\kappa\big(\alpha,\sigma\big)} + \frac{1}{p}\mathbf{v}^T\big(\mathbf{u}-\alpha\mathbf{w}^\star\big) - \frac{\sigma\tau}{2} - \frac{\sigma}{2\tau}\cdot\left\|\frac{\beta}{\sqrt{n}}\mathbf{h} - \frac{1}{\sqrt{p}}\mathbf{v} + \frac{\gamma}{\sqrt{p}}\mathbf{w}^\star\right\|^2. \tag{9.68}$$

First, we would like to show that $f^{(p)}$ is jointly convex with respect to α, σ, and $\mathbf{u}$. From Lemma (25), we know that $\sqrt{c_\kappa\big(\alpha,\sigma\big)}$ is jointly convex with respect to α and σ. The function $\psi(\cdot)$ is also convex and the remaining terms are all linear with respect to these three variables. Hence, $f^{(p)}$ is convex with respect to $\mathbf{u}$, α, and σ.

Next, we show that this function is jointly concave with respect to the remaining variables. We note that the function $\left\|\frac{\beta}{\sqrt{n}}\mathbf{h} - \frac{1}{\sqrt{p}}\mathbf{v} + \frac{\gamma}{\sqrt{p}}\mathbf{w}^\star\right\|^2$ is convex with respect to variables $\mathbf{v}$, γ, and β. The perspective of this function $\frac{1}{\tau}\left\|\frac{\beta}{\sqrt{n}}\mathbf{h} - \frac{1}{\sqrt{p}}\mathbf{v} + \frac{\gamma}{\sqrt{p}}\mathbf{w}^\star\right\|^2$ is (jointly) convex with respect to $(\gamma,\beta,\tau,\mathbf{v})$. Therefore, $f^{(p)}$ is jointly convex with respect to these variables as the remaining terms are affine with respect to $(\gamma,\beta,\tau,\mathbf{v})$. Next, we define the function $C^{(p)}$ by maximizing $f^{(p)}$ with respect to $\mathbf{v}$, i.e.,

$$C^{(p)}\big(\alpha,\sigma,\mathbf{u};\gamma,\beta,\tau\big) = \max_{\mathbf{v}\in\mathbb{R}^p} f^{(p)}\big(\alpha,\sigma,\mathbf{u};\gamma,\beta,\tau,\mathbf{v}\big)\ . \tag{9.69}$$

This function is also jointly convex-concave, since it is a point-wise maximum of concave function with respect to $\mathbf{v}$. The result is the consequence of the fact that $C^{(p)}$ converges to C, i.e.,

$$C^{(p)}\big(\alpha,\sigma,\mathbf{u};\gamma,\beta,\tau\big) \xrightarrow{P} C(\alpha,\sigma,\gamma,\beta,\tau)\ ,\ \text{ as }\ p\to\infty. \tag{9.70}$$

9.8 Proof of Theorem 10 in Chapter 7

In this section, we prove the result presented in Theorem 10 which identifies the phase transition on the separability of the data. To this end, we exploit the result of Lemma 23 which associates the following optimization to the EMM optimization (9.33).

$$\begin{aligned} \min_{\substack{\alpha\in\mathbb{R}\\ \tilde{\mathbf{w}}\in\mathbb{R}^p\\ \tilde{\mathbf{w}}\perp\mathbf{w}^\star}} \; & \frac{1}{p}\psi(\alpha\mathbf{w}^\star+\tilde{\mathbf{w}}) \\ & \text{s.t.} \;\; \frac{1}{p}(\mathbf{h}^T\tilde{\mathbf{w}})^2 \geq n\cdot c_\kappa\big(\alpha, \frac{\|\tilde{\mathbf{w}}\|}{\sqrt{p}}\big). \end{aligned} \tag{9.71}$$

We first show that, as $p, n \to \infty$, $\delta > \delta^* = \delta^*(\kappa)$ is the necessary and sufficient condition for the optimization program (9.71) to have a feasible solution. Define $\sigma := \|\tilde{\mathbf{w}}\|\sqrt{p}$, and write the following:

$$\sup_{\substack{\alpha\in\mathbb{R}\\ \tilde{\mathbf{w}}\perp\mathbf{w}^\star}} \frac{1}{p}(\mathbf{h}^T\tilde{\mathbf{w}})^2 - n\cdot c_\kappa\big(\alpha, \frac{\|\tilde{\mathbf{w}}\|}{\sqrt{p}}\big) = \sup_{\sigma\geq 0,\alpha} \sigma^2\cdot\left\|\mathbf{P}^\perp\mathbf{h}\right\|^2 - n\cdot c_\kappa\big(\alpha,\sigma\big)\;. \tag{9.72}$$

Note that we used the fact that the $\mathbf{P}^\perp$ is the projection onto the hyperplane orthogonal to $\mathbf{w}^\star$. The supremum is achieved iff $\tilde{\mathbf{w}}$ chooses its direction to be the same as $\mathbf{P}^\perp\mathbf{h}$. The optimization program has a feasible point if and only if the optimal value in (9.72) is non-negative. In other words, the necessary and sufficient condition on the separability of the data is:

$$\exists\, r\geq 0,\; s\;,\;\; \text{s.t.}\;\; r^2\cdot\left\|\mathbf{P}^\perp\mathbf{h}\right\|^2 - n\cdot c_\kappa\big(s,r\big)\geq 0 \iff \frac{1}{n}\left\|\mathbf{P}^\perp\mathbf{h}\right\|^2 \geq \delta^* = \inf_{s,r\geq 0}\frac{c_\kappa(s,r)}{r^2}. \tag{9.73}$$

Next we note that $\mathbf{h}$ has i.i.d. $\mathcal{N}(0,1)$ entries, therefore, SLLN asserts that,

$$\frac{1}{n}\left\|\mathbf{P}^\perp\mathbf{h}\right\| \xrightarrow{a.s.} \frac{p-1}{n}. \tag{9.74}$$

Therefore, as $n, p\to\infty$ with $\delta := \frac{p}{n}$, the optimization program (9.71) is feasible if and only if $\delta > \delta^*$. As Lemma 23 states that the solution to the EMM optimization (9.33) converges in probability to the solution of (9.71). Therefore, $\delta > \delta^*$ indicates the phase transition for the existence of the EMM classifier.

We would also want to refer the interested reader to [38] for an astute geometric/combinatorial perspective on the phase transition behavior in binary classification.

9.9 EMM for Various Structures

In this section, we provide some technical details on how to characterize the performance of the classifiers introduced in Section 9.4. For each of the three classifiers, depending on the distribution of the underlying parameter ($\mathbf{w}^\star$), we simplify the nonlinear system (9.8) by explicitly evaluating the expected values.

Max-margin classifier (ℓ_2-EMM)

As mentioned earlier, when $\psi(\cdot) = \frac{1}{2}\|\cdot\|_2^2$, the EMM classifier will become the well-known max-margin classifier. In this case, we can find the following closed-form for the proximal operator:

$$\text{Prox}_{\frac{t}{2}\|\cdot\|^2}(\mathbf{v}) = \frac{1}{1+t}\mathbf{v}. \tag{9.75}$$

Therefore, the expectations in the nonlinear system (9.8) can be computed explicitly as follows:

$$\begin{cases} \frac{1}{p}\,\mathbb{E}\left[\mathbf{w}^{\star T}\text{Prox}_{\frac{\sigma\tau}{2}\|\cdot\|^2}\big((\alpha-\sigma\tau\gamma)\mathbf{w}^\star+\beta\sigma\tau\sqrt{\delta}\mathbf{h}\big)\right] = \dfrac{\kappa^2\big(\alpha-\sigma\tau\gamma\big)}{1+\sigma\tau}, \\ \frac{1}{p}\,\mathbb{E}\left[\mathbf{h}^{T}\text{Prox}_{\sigma\tau\psi(\cdot)}\big((\alpha-\sigma\tau\gamma)\mathbf{w}^\star+\beta\sigma\tau\sqrt{\delta}\mathbf{h}\big)\right] = \dfrac{\beta\sigma\tau\sqrt{\delta}}{1+\sigma\tau}, \\ \frac{1}{p}\,\mathbb{E}\left\|\text{Prox}_{\sigma\tau\psi(\cdot)}\big((\alpha-\sigma\tau\gamma)\mathbf{w}^\star+\beta\sigma\tau\sqrt{\delta}\mathbf{h}\big)\right\|^2 = \dfrac{\kappa^2\big(\alpha-\sigma\tau\gamma\big)^2+\beta^2\sigma^2\tau^2\delta}{\big(1+\sigma\tau\big)^2}. \end{cases} \tag{9.76}$$

Replacing these evaluations into the first three equations in the nonlinear system (9.8) will explicitly give two of the variables in terms of the other three variables. More specifically, we get $\gamma = -\alpha$ from the first equation and $\beta = \frac{1+\sigma\tau}{\tau\sqrt{\delta}}$ from the third equation in the nonlinear system (9.8). Hence, the nonlinear system would reduce to solving the following system of 3 nonlinear equations with 3 unknowns:

$$\begin{cases} \sqrt{c_\kappa(\alpha,\sigma)} = \sigma\sqrt{\delta}, \\ \dfrac{\partial c_\kappa(\alpha,\sigma)}{\partial\alpha} = \dfrac{-2\kappa^2\alpha\tau\sigma\delta}{1+\sigma\tau}, \\ \dfrac{\partial c_\kappa(\alpha,\sigma)}{\partial\sigma} = \dfrac{2\sigma\delta}{1+\sigma\tau}. \end{cases} \tag{9.77}$$

Sparse classifier (ℓ_1-EMM)

The second choice for the potential function is $\psi(\cdot) = \|\cdot\|_1$, which is used to promote sparsity in the underlying parameter. Here, we assume that the entries of the underlying parameter are generated

independently from the distribution Π_s introduced in (9.24), where $s \in (0,1)$ denotes the sparsity factor which indicates the probability of an entry being nonzero. The nonzero entries have Gaussian distribution with variance κ^2/s. The proximal operator for ℓ_1 norm can be computed explicitly as,

$$\text{Prox}_{t\|\cdot\|_1}(\mathbf{u}) = \eta(\mathbf{u}, t), \tag{9.78}$$

where $\eta(x,t) = \frac{x}{|x|}\big(|x| - t\big)_+$ is the soft thresholding function that has been applied entrywise. The expectations that appear in the first three equations in the nonlinear system (9.8) can be presented as follows:

$$\begin{cases} \frac{1}{p}\,\mathbb{E}\left[\mathbf{w}^{\star T}\text{Prox}_{\frac{\sigma\tau}{2}\|\cdot\|^2}\big((\alpha-\sigma\tau\gamma)\mathbf{w}^\star + \beta\sigma\tau\sqrt{\delta}\mathbf{h}\big)\right] = 2\kappa^2\cdot Q(t_1)\cdot\big(\alpha-\sigma\tau\gamma\big), \\ \frac{1}{p}\,\mathbb{E}\left[\mathbf{h}^{T}\text{Prox}_{\sigma\tau\psi(\cdot)}\big((\alpha-\sigma\tau\gamma)\mathbf{w}^\star + \beta\sigma\tau\sqrt{\delta}\mathbf{h}\big)\right] = \big[2sQ(t_1) + 2(1-s)Q(t_2)\big]\cdot\beta\sigma\tau\sqrt{\delta}, \\ \frac{1}{p}\,\mathbb{E}\left\|\text{Prox}_{\sigma\tau\psi(\cdot)}\big((\alpha-\sigma\tau\gamma)\mathbf{w}^\star + \beta\sigma\tau\sqrt{\delta}\mathbf{h}\big)\right\|^2 = 2\sigma^2\tau^2\big(\frac{s}{t_1^2}\cdot\chi(t_1) + \frac{1-s}{t_2^2}\cdot\chi(t_2)\big), \end{cases} \tag{9.79}$$

where t_1 and t_2 are defined as,

$$t_1 = \frac{\sigma\tau}{\sqrt{\frac{\kappa^2}{s}(\alpha-\sigma\tau\gamma)^2 + \beta^2\sigma^2\tau^2\delta}}\,, \quad t_2 = \frac{1}{\beta\sqrt{\delta}}, \tag{9.80}$$

and the function $\chi : \mathbb{R} \to \mathbb{R}_+$ is defined as:

$$\chi(t) = \mathbb{E}[\big(Z-t\big)_+^2] = Q(t)\big(1+t^2\big) - t\phi(t)\,, \tag{9.81}$$

where the random variable Z in the above expectation have standard normal distribution, and $\phi(x) = \frac{1}{\sqrt{2\pi}}\exp(-x^2/2)$ denotes the density of the standard normal distribution. Replacing the computed expectations in (9.79) in the nonlinear system (9.8) gives the sparse nonlinear system presented in (9.27).

It is worth mentioning that the sparse nonlinear system (9.27) can be solved efficiently via iterative numerical methods. A main advantage of the sparse nonlinear system is that it has been presented in terms of the $Q(\cdot)$ function which can be computed quickly in most numerical softwares (e.g. MATLAB). For our numerical simulations in Section 9.5, we used an accelerated fixed-point iterative method to find the solution of the nonlinear system.

Binary classifier (ℓ_∞-EMM)

The third and last choice of the potential function is the ℓ_∞ norm. In this case, the potential function is defined as $\psi(\cdot) = p\,\|\cdot\|_\infty$[3]. The following lemma determines how to compute the proximal operator in this case.

Lemma 26. *Let* $\mathbf{u} \in \mathbb{R}^p$ *have i.i.d. entries from a distribution* Π. *Then, for* $t > 0$, *we have:*

$$Prox_{tp\|\cdot\|_\infty}(\mathbf{u}) = \mathbf{u} - Prox_{\lambda\|\cdot\|_1}(\mathbf{u}), \tag{9.82}$$

where λ *is defined as,*

1. *for* $t \le \mathbb{E}\,|W|$, λ *is the unique solution of* $\mathbb{E}\left[(|W| - \lambda)_+\right] = t$.
2. *for* $t \ge \mathbb{E}\,|W|$, *then* $\lambda = 0$.

In the following subsections, we use the result of Lemma 26 to compute the proximal operator for two different models (i.e., two different distributions on the entries of $\mathbf{w}^\star$).

ℓ_∞-EMM with sparse parameter

Here, we consider the case where the entries of $\mathbf{w}^\star$ are drawn independently from the distribution Π_s defined in (9.24). Note that when we set s to 1, this distribution will be the same as i.i.d. Gaussian entries. Hence, the result in this section can be applied to the non-sparse setting (when the underlying parameter has i.i.d. Gaussian entries.)

Using the result of Lemma 26, in this case the proximal operator can be computed as follows,

$$\mathrm{Prox}_{\sigma\tau p\|\cdot\|_\infty}\big((\alpha-\sigma\tau\gamma)\mathbf{w}^\star+\beta\sigma\tau\sqrt{\delta}\mathbf{h}\big) = (\alpha-\sigma\tau\gamma)\mathbf{w}^\star+\beta\sigma\tau\sqrt{\delta}\mathbf{h}-\mathrm{Prox}_{\lambda\sigma\tau\|\cdot\|_1}\big((\alpha-\sigma\tau\gamma)\mathbf{w}^\star+\beta\sigma\tau\sqrt{\delta}\mathbf{h}\big), \tag{9.83}$$

where λ is defined in terms of the proxies t_1 and t_2 (defined in (9.80)):

1. If $\frac{s}{t_1} + \frac{1-s}{t_2} > \sqrt{\frac{\pi}{2}}$, then λ is the unique solution of the following nonlinear equation:

$$2s \cdot \Big[\frac{1}{t_1}\phi(\lambda t_1) - \lambda Q(\lambda t_1)\Big] + 2(1-s)\Big[\frac{1}{t_2}\phi(\lambda t_2) - \lambda Q(\lambda t_2)\Big] = 1\ . \tag{9.84}$$

[3]The multiplication by the dimension, p, is necessary to ensure that all the terms in the optimization have constant ($O(1)$) order.

2. If $\frac{s}{t_1} + \frac{1-s}{t_2} \leq \sqrt{\frac{\pi}{2}}$, then $\lambda = 0$.

Therefore, after finding the value of λ by solving equation (9.84), the proximal operator which appears in the first three equations of the nonlinear system (9.8) can be written explicitly in terms of the proximal operator of the ℓ_1 norm which was illustrated in the previous part. Also, similarly to the case of ℓ_1-EMM, the expectations are written in terms of the functions $Q(\cdot)$ and $\phi(\cdot)$. Therefore, the solution to the nonlinear system can be found efficiently using numerical solvers.

ℓ_∞-EMM with binary parameter

Here, we consider the case where $\mathbf{w}^\star$ has i.i.d. entries with distribution $\Pi = \kappa \cdot \textsc{Rad}(\frac{1}{2})$. To simplify our presentation, we define the following proxy:

$$t_3 = \big(\frac{\alpha}{\sigma\tau} - \gamma\big) \cdot \kappa \; .$$

Using the result of Lemma 26, in this case the proximal operator can be computed as follows,

$$\text{Prox}_{\sigma\tau p\|\cdot\|_\infty}\big((\alpha - \sigma\tau\gamma)\mathbf{w}^\star + \beta\sigma\tau\sqrt{\delta}\mathbf{h}\big) = (\alpha - \sigma\tau\gamma)\mathbf{w}^\star + \beta\sigma\tau\sqrt{\delta}\mathbf{h} - \text{Prox}_{\lambda\sigma\tau\|\cdot\|_1}\big((\alpha - \sigma\tau\gamma)\mathbf{w}^\star + \beta\sigma\tau\sqrt{\delta}\mathbf{h}\big), \tag{9.85}$$

where λ is defined as:

1. When $\beta\sqrt{\delta} \cdot \phi(-\frac{t_3}{\beta\sqrt{\delta}}) + t_3 \cdot Q(-\frac{t_3}{\beta\sqrt{\delta}}) > \frac{1}{2}$, λ is defined as the unique solution of the following equations:
$$\beta\sqrt{\delta} \cdot \phi(\frac{\lambda - t_3}{\beta\sqrt{\delta}}) + (t_3 - \lambda) \cdot Q(\frac{\lambda - t_3}{\beta\sqrt{\delta}}) = \frac{1}{2} \; . \tag{9.86}$$
2. Otherwise, $\lambda = 0$.

Hence, λ can be computed by solving the equation (9.86), and consequently the proximal operator which appears in the first three equations of the nonlinear system (9.8) can be written explicitly in terms of the proximal operator of the ℓ_1 norm which was illustrated in the previous part.

ROBUSTIFYING BINARY CLASSIFICATION TO ADVERSARIAL PERTURBATION

[1] F. Salehi and B. Hassibi. "Robustifying Binary Classification to Adversarial Perturbation". In: *arXiv preprint arXiv:2010.15391* (2020).

Despite the enormous success of machine learning models in various applications, most of these models lack resilience to (even small) perturbations in their input data. Hence, new methods to robustify machine learning models seem very essential.

In this chapter, we consider the problem of binary classification with adversarial perturbations. By investigating the solution to a min-max optimization (which considers the worst-case loss in the presence of adversarial perturbations), we introduce a generalization to the max-margin classifier which takes into account the power of the adversary in manipulating the data. We refer to this classifier as the "Robust Max-margin" (RM) classifier. Under some mild assumptions on the loss function, we theoretically show that the gradient descent iterates (with sufficiently small step size) converge to the RM classifier in its direction. Therefore, the RM classifier can be studied to compute various performance measures (e.g. generalization error) of binary classification with adversarial perturbations.

10.1 Motivation and Background

Machine learning models have been very successful in many applications, ranging from spam detection, speech and visual recognition, to the analysis of genome sequencing and financial markets. Yet, despite this indisputable success, it has been observed that commonly used machine learning models (such as deep neural networks) are very unstable in the presence of non-random perturbations [125, 17, 32].

In this chapter, we study the simple (yet fundamental) problem of binary classification where the goal is to find a classifier that has a high accuracy in predicting the binary labels when having feature vectors as its input. When the clean data is available, max-margin classifier [140] is the model of choice as maximizing the margin is interpreted as minimizing the risk of misclassification [37]. Recently, it was shown in [120] that for a broad class of loss functions, including the well-known logistic loss, the gradient descent iterates converge to the max-margin classifier. More recently, the

asymptotic performance of this classifiers has been characterized in [95, 39, 114]. In Chapter 9, we have provided a detailed discussion on the performance of the max-margin classifier (as a special member of extended margin maximizers) under Gaussian data.

Here, we consider the case where the training data is perturbed by an adversary and introduce the "Robust Max-margin" (RM) classifier as a generalization of max-margin to perturbed input data. We then consider the adversarial training method, in which the optimal parameter is a solution to a saddle-point optimization. We show that the gradient descent algorithm with properly-tuned step sizes converges in its direction to the RM classifier. A significant consequence of this result is that one can characterize various performance measures (e.g. generalization error) of adversarial training in binary classification by analyzing the performance of the RM classifier.

To the extent of our knowledge, this is the first work that introduces the robust max-margin classifier and proves the convergence of gradient descent iterates to this classifier. Very recently, the authors of [72] have analyzed the performance of robust max-margin classifier (referred to as the "robust separation") under i.i.d. Gaussian training data. Their analysis on the performance of the resulting estimator is similar to what we showed in the previous chapter and is based on the Convex Gaussian Min-max Theorem [121, 130]. Similar analyses have been recently provided for the performance of max-margin classifiers as well as other generalized linear models [113, 95, 39, 49, 114, 55].

The organization of this chapter is as follows: inn Section 10.2, we provide some background on the binary classification problem and how it connects with the max-margin classifier. The mathematical setup for the problem of binary classification with perturbed training data is provided in Section 10.3. The main result of this chapter is presented in Section 10.4, and the proofs are provided in Sections 10.5 and 10.6.

10.2 Preliminaries

We start with some notations that are used throughout this chapter. For any vector $\mathbf{w} \in \mathbb{R}^p$, the binary classifier associated with $\mathbf{w}$ is defined as: $C_{\mathbf{w}} : \mathbb{R}^p \to \{\pm 1\}$, such that $C_{\mathbf{w}}(\mathbf{x}) = \text{Sign}(\mathbf{w}^T\mathbf{x})$. $\mathbb{N}$ denotes the set of non-negative integers. $\sigma_{\max}(\mathbf{M})$ denotes the maximum singular value of the matrix $\mathbf{M}$. $\mathbf{0}_d$ and $\mathbf{1}_d$, respectively, represent the all-one and all-zero vectors in dimension d. A function $f(\cdot)$ is said to be L-smooth if its derivative, $f'(\cdot)$, is L-Lipschitz.

Background: binary classification with unperturbed data

Here, we review some of the main ideas of binary classification when the adversary is not present. Let $\mathcal{D} = \{(\mathbf{x}_i, y_i) : 1 \le i \le n\}$ denote a set of data points, where for $i = 1, \dots, n$, $\mathbf{x}_i \in \mathbb{R}^p$ is the feature vector, and $y_i \in \{\pm 1\}$ is the binary label. We assume that $\mathcal{D}$ is linearly separable, i.e., there

exist $\mathbf{w}^\star \in \mathbb{R}^p$ such that:

$$y_i = \text{SIGN}(\mathbf{x}_i^T \mathbf{w}^\star) \ , \text{ for } i = 1, 2, \ldots, n. \tag{10.1}$$

When the training data has no perturbation, one can attempt to find a classifier by minimizing the empirical loss on data set $\mathcal{D}$. In the setting of binary classification, the loss function is usually formed as,

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{n} \ell(y_i \mathbf{x}_i^T \mathbf{w}) \tag{10.2}$$

where the function $\ell(\cdot) : \mathbb{R} \to \mathbb{R}_+$ is a decreasing function that approaches 0 as its input approaches infinity. A typical approach to find the minimizer of the loss function $\mathcal{L}(\mathbf{w})$ is through the iterative algorithms, such as the gradient descent (GD) algorithm. The convergence of the GD iterates on separable data sets has been studied in recent papers [73, 120], where it was shown, among others, that while the norm of the iterates approaches infinity, their direction would approach to the direction of the well-known L_2 max-margin classifier defined as,

$$\begin{aligned} \mathbf{w}_M = \arg \min_{\mathbf{w} \in \mathbb{R}^p} \quad & \|\mathbf{w}\| \\ \text{s.t.} \quad & y_i \mathbf{x}_i^T \mathbf{w} \geq 1 \ , \quad 1 \leq i \leq n. \end{aligned} \tag{10.3}$$

In other words, their result states that for almost every $\mathbf{x} \in \mathbb{R}^p$, $C_{\mathbf{w}_t}(\mathbf{x}) \to C_{\mathbf{w}_M}(\mathbf{x})$ as t grows, where $\mathbf{w}_t$ denotes the result of GD after t steps. The max-margin classifier (10.3) (a.k.a. hard-margin SVM [37]) has been extensively studied in the machine learning community (see Chapter 9 for a more detailed discussion). This classifier simply maximizes the smallest distance of the data points to the separating hyper-plane (referred to as the margin).

The above mentioned result, i.e., convergence of the GD iterates to the max-margin classifier, has significant consequences as the max-margin classifier can then be studied to compute various performance measures (such as the generalization error) of the resulting estimator. Very recently, researchers have exploited this result to accurately compute the generalization error of GD over the logistic loss [95, 114].

10.3 Binary Classification with Adversarial Perturbations

As explained earlier in Section 10.1, understanding the behavior of machine learning models under perturbed input is very essential with the goal of improving the robustness of these models. Inspired by recent advances in understanding the behavior of machine learning models under adversarial perturbation, here we study the problem of binary classification with perturbed data.

We assume that the training data is a perturbed version of the underlying data set, $\mathcal{D}$. Let $\mathcal{D}' = \{(\mathbf{x}_i + \mathbf{z}_i, y_i) : 1 \leq i \leq n\}$ denote the set of training data, where, for $i = 1, 2, \ldots, n$, $\mathbf{z}_i \in \mathcal{S}_i$

is the unknown perturbation, and the set $\mathcal{S}_i$ consists of all the allowed perturbation vector. In the adversarial setting, it is often assumed that the perturbation vectors, $\{\mathbf{z}_i\}_{i=1}^n$, are chosen in such a way that the training algorithm is beguiled into generating a wrong solution.

Throughout this chapter, for our analysis purposes, we assume that the perturbation vectors have bounded norms by defining $\mathcal{S}_i = \epsilon_i \mathcal{B}_p$, where $\mathcal{B}_p$ denotes the unit ball in $\mathbb{R}^p$, and $\epsilon_i \geq 0$, for $1 \leq i \leq n$, indicates the maximum allowed norm for the i^{th} perturbation vector, $\mathbf{z}_i$. While the perturbation vectors are hidden to us, we assume having knowledge of $\{\epsilon_i\}_{i=1}^n$.

Note that the set of allowed perturbations can be different for different data points. This includes certain special cases such as: (1) only a subset of the data is perturbed ($\epsilon_i = 0$ if the i^{th} data point is not perturbed), and (2) all the data points have the same perturbation set, i.e., for some $\epsilon \geq 0$, we have $\epsilon_i = \epsilon$ for $1 \leq i \leq n$.

Saddle-point optimization

The parameters of the desired model are often derived by forming a loss function and solving an optimization problem to find a minimizer of the loss. In adversarial training, one should also consider the manipulative power of the adversary where the adversary attempts to misguide the training algorithm. When the goal of a training algorithm is to minimize a loss function, one can view the adversary as an entity which attempts to maximize the loss. The following min-max optimization problem incorporates the contrary behaviors of the adversary and the training algorithm with respect to the loss function.

$$\min_{\mathbf{w}\in\mathbb{R}^p} \max_{\mathbf{z}_i\in\mathcal{S}_i, 1\leq i\leq n} \mathcal{L}(\mathbf{w}) := \sum_{i=1}^n \ell\big(y_i(\mathbf{x}_i + \mathbf{z}_i)^T\mathbf{w}\big). \tag{10.4}$$

In order to find a robust model, we should solve this saddle-point optimization. Under our assumptions on the perturbation sets, we can introduce the function $\mathcal{L}_{\boldsymbol{\epsilon}}(\mathbf{w})$ which is the result of the inner maximization in (10.4), i.e.,

$$\mathcal{L}_{\boldsymbol{\epsilon}}(\mathbf{w}) = \sum_{i=1}^n \max_{\|\mathbf{z}_i\|\leq\epsilon_i} \ell\big(y_i(\mathbf{x}_i + \mathbf{z}_i)^T\mathbf{w}\big), \tag{10.5}$$

where $\boldsymbol{\epsilon} = \Big[\epsilon_1, \epsilon_2, \ldots, \epsilon_n\Big]^T$. Therefore, the robust classifier is defined as a minimizer of $\mathcal{L}_{\boldsymbol{\epsilon}}(\mathbf{w})$.

10.4 Main Results

In this section, we present the main results of this chapter, that is the convergence of the gradient descent iterates to the robust max-margin classifier. We first introduce the **R**obust **M**ax-margin

(**RM**) classifier as an extension of the max-margin classifier when the training data is perturbed. Consequently, in our main result we show that, under some conditions on the function $\ell(\cdot)$, the gradient descent algorithm (with sufficiently small step size) converges in its direction to the RM classifier.

Robust Max-margin (RM) Classifier

The max-margin classifier is a classifier that maximizes the minimum distance of the data points to the separating hyperplane (also known as the margin). In our setting where the training data is perturbed, we should modify the notion of the margin to incorporate various perturbations across data points. More specifically, in order to get a robust classifier, we would like the data points with higher perturbations to be farther away from the resulting separating hyperplane.

The **Robust Max-margin** classifier is defined as,

$$\begin{aligned} \mathbf{w}_{RM}^{(\boldsymbol{\epsilon})} := \arg\min_{\mathbf{w}\in\mathbb{R}^p} \quad & \|\mathbf{w}\| \\ \text{s.t.} \quad & y_i \mathbf{x}_i^T \mathbf{w} \geq 1 + \epsilon_i \|\mathbf{w}\| \ , \quad 1 \leq i \leq n. \end{aligned} \tag{10.6}$$

As observed in the constraints of this optimization, the RM classifier enforces data points with higher perturbations to keep a larger distance from the separating hyperplane $\{\mathbf{x} : \mathbf{w}_M^T \mathbf{x} = 0\}$.

When the data is perturbed, we expect the RM classifier to outperform the max-margin classifier. Figure 10.1 depicts a comparison in generalization error between the max-margin and the RM classifier. Although for small perturbations, the two models behave the same way, the RM classifier has a better performance as we increase the norm of perturbations.

While the separability of the data is necessary for the existence of the RM classifier, it is not sufficient. The following lemma provides a sufficient condition for its existence.

Lemma 27. *The RM classifier exists when the data set $\mathcal{D} = \{(\mathbf{x}_i, y_i) : 1 \leq i \leq n\}$ is separable and,*

$$\|\boldsymbol{\epsilon}\|_\infty < \frac{1}{\|\mathbf{w}_M\|_2} \ , \tag{10.7}$$

where $\mathbf{w}_M$ is the max-margin classifier.

Proof. The max-margin classifier, $\mathbf{w}_M$, exists when $\mathcal{D}$ is linearly separable. Also, $\bar{\mathbf{w}} = \frac{1}{1-\|\boldsymbol{\epsilon}\|_\infty \|\mathbf{w}_M\|_2} \mathbf{w}_M$ is a feasible point of the optimization (10.6). Therefore, the RM classifier exists and,

$$\|\mathbf{w}_M\| \leq \|\mathbf{w}_{RM}\| \leq \|\bar{\mathbf{w}}\| \ .$$

□

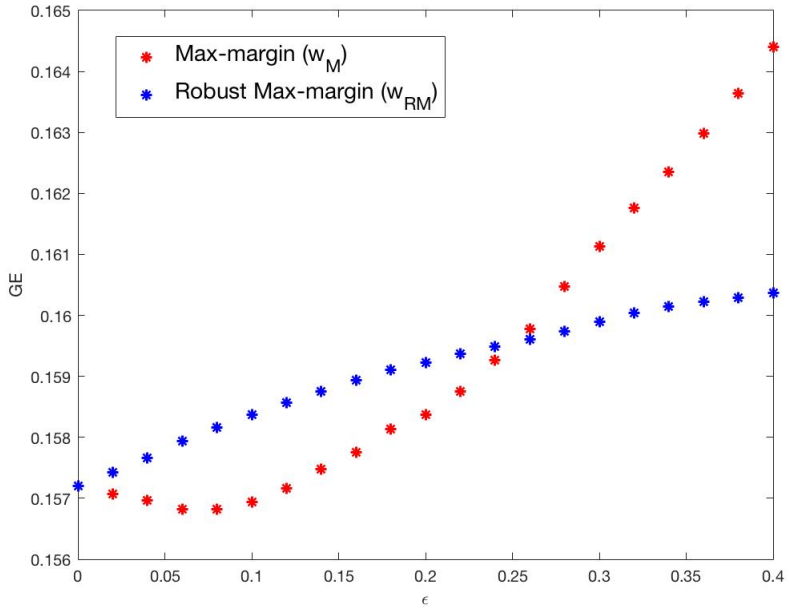


Figure 10.1: A comparison in generalization error (GE) between the max-margin (10.3) and the robust max-margin (10.6). The result is the average over 20 independent trials with $n = 100$ and $p = 40$. The data is generated from a Gaussian distribution and 40% of data points are perturbed with maximum norm of ϵ. For large values of ϵ, the RM classifier has a better generalization error than the max-margin classifier.

When the perturbation sets are the same for different data points, one expects the RM classifier to be the same as the max-margin classifier.

Lemma 28. *If $\boldsymbol{\epsilon} = \epsilon \times \mathbf{1}_n$ for some $\epsilon \geq 0$, the RM classifier exists if and only if $\epsilon < \frac{1}{\|\mathbf{w}_M\|}$. In this case,*

$$\mathbf{w}_{RM} = \frac{\mathbf{w}_M}{1 - \epsilon \|\mathbf{w}_M\|}. \tag{10.8}$$

Proof. Assume $\mathbf{w}_{RM}$ exists, then we have $\bar{\mathbf{w}} = \frac{\mathbf{w}_{RM}}{1+\epsilon\|\mathbf{w}_{RM}\|}$ satisfy the constraints in the optimization (10.3). Since $\mathbf{w}_M$ is the solution to this optimization, we have $\|\mathbf{w}_M\| \leq \|\bar{\mathbf{w}}\|$ which gives $\epsilon \cdot \|\mathbf{w}_M\| < 1$. It is easy to check that $\mathbf{w} = \frac{\mathbf{w}_M}{1-\epsilon\|\mathbf{w}_M\|}$ is the solution to the optimization (10.6), as it satisfies the constraints and $\mathbf{w}_M$ is the optimal value of the optimization program (10.3). □

Convergence of GD Iterates

In this section, we present the main result of the chapter that is the convergence of the gradient descent iterates to the RM classifier. As discussed earlier in Section 10.3, the goal is to solve the

following optimization problem,

$$\min_{\mathbf{w}\in\mathbb{R}^p} \mathcal{L}_{\boldsymbol{\epsilon}}(\mathbf{w}), \tag{10.9}$$

where $\mathcal{L}_{\boldsymbol{\epsilon}}(\cdot)$ is defined in (10.5). Gradient descent (GD) is the common method of choice to find a minimizer of this optimization. Starting from an initialization, $\mathbf{w}_0 \in \mathbb{R}^p$, the GD iterates are generated through the following update rule:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \cdot \nabla \mathcal{L}_{\boldsymbol{\epsilon}}(\mathbf{w}_t), \text{ for } t \in \mathbb{N}, \tag{10.10}$$

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

Our goal is to study the behavior of the GD iterates as t grows large. For our analysis, we need some assumptions to hold for the loss function $\ell(\cdot)$.

Assumption 5. *The function $\ell : \mathbb{R} \to \mathbb{R}_+$ is twice-differentiable, monotonically decreasing, and β-smooth.*

We note that the common choices of the loss function satisfy the conditions in Assumption 5. For instance, the logistic loss defined as $\ell(u) = \log\left(1+\exp(-u)\right)$ satisfies these conditions (with $\beta = 1$.)

We first state the following lemma which provides some insights on the behavior of GD iterates, $\mathbf{w}_t$, as $t \to \infty$.

Lemma 29. *Consider the gradient descent iterates* (10.10) *with step size $\eta < 2 \cdot \beta^{-1} \cdot (\sigma_{\max}(\mathbf{X}) + \|\boldsymbol{\epsilon}\|)^{-2}$, where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^T \in \mathbb{R}^{n\times p}$ is the data matrix, $\mathcal{L}_{\boldsymbol{\epsilon}}$ is defined in* (10.5)*, and $\ell(\cdot)$ satisfies Assumption 5. If the RM classifier exists, then, as $t \to +\infty$, we have,*

i. $\|\mathbf{w}_t\| \to +\infty$,

ii. $\nabla \mathcal{L}_{\boldsymbol{\epsilon}}(\mathbf{w}_t) \to \mathbf{0}_p$ *, and,*

iii. $y_i \mathbf{x}_i^T \mathbf{w}_t - \epsilon_i \|\mathbf{w}_t\| \to +\infty$, *for $i = 1, 2, \ldots, n$.*

The proof of this lemma is provided in Section 10.5.

Lemma 29 provides useful insights on the behavior of the gradient descent iterates. With small enough step size, as t grows, the norm of $\mathbf{w}_t$ becomes unbounded while making $\mathcal{L}(\mathbf{w}_t)$ closer to zero. Since $\mathbf{w}_t$ diverges, we focus our attention on its direction, i.e., the normalized vector $\frac{\mathbf{w}_t}{\|\mathbf{w}_t\|}$. In fact, the classifier defined by $\mathbf{w}_t$, $C_{\mathbf{w}_t}(\cdot)$ only depends on its direction. Therefore, if $\frac{\mathbf{w}_t}{\|\mathbf{w}_t\|}$ converges, we can claim that the classifiers generated by GD iterates converge.

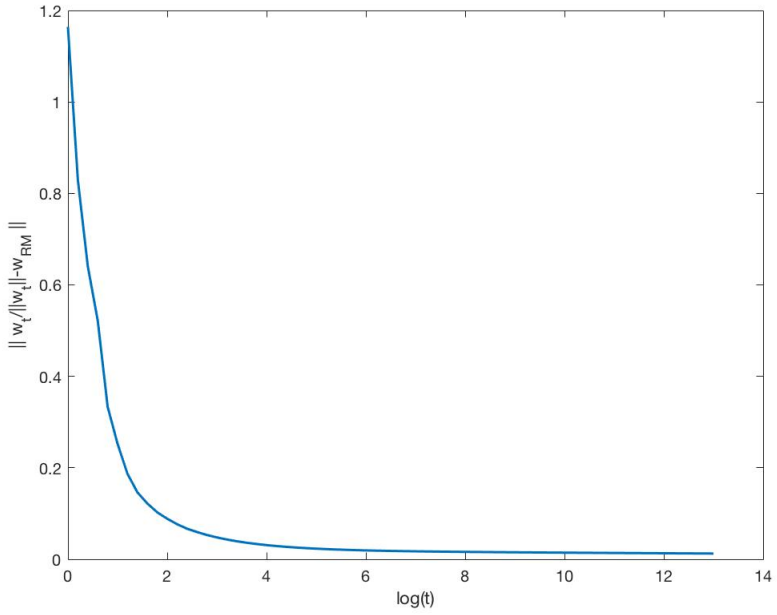


Figure 10.2: Convergence of GD iterates to the RM classifier. For our experiment, we have $n = 30$, $p = 10$, number of iterations is 10^{13}, and $\epsilon_i \sim \text{Unif}(0, \frac{1}{\|\mathbf{w}_M\|})$. The distance between the max-margin and the RM classifier is $\left\| \frac{\mathbf{w}_M}{\|\mathbf{w}_M\|} - \frac{\mathbf{w}_{RM}}{\|\mathbf{w}_{RM}\|} \right\| = 0.2192$.

Our main result in Theorem 14 states that the classifiers generated from the GD iterates converges to the RM classifier defined in 10.6. Before stating this result, we need the following definition which is a modified version of an assumption in [120].

Definition 14. *A function $f(u)$ has a tight exponential tail if there exist positive constants a, c, τ, μ such that for all $u > \tau$:*

$$\begin{cases} f(u) \leq c\big(1 + \exp(-\mu \cdot u)\big) \exp(-a \cdot u), \text{ and,} \\ f(u) \geq c\big(1 - \exp(-\mu \cdot u)\big) \exp(-a \cdot u). \end{cases} \tag{10.11}$$

Theorem 14. *Let Assumption 5 hold and $-\ell'(\cdot)$ have a exponential tail. Consider the gradient descent iterates in* (10.10) *with $\eta < 2 \cdot \beta^{-1} \cdot (\sigma_{\max}(\mathbf{X}) + \|\boldsymbol{\epsilon}\|)^{-2}$. Then, for almost every data set, we have,*

$$\lim_{t \to \infty} \left\| \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|} - \frac{\mathbf{w}_{RM}}{\|\mathbf{w}_{RM}\|} \right\| = 0. \tag{10.12}$$

Therefore, the resulting classifier converges to the RM classifier.

Remark 17. *The assumption on $-\ell'(\cdot)$ having a tight exponential tail holds for common loss functions in binary classification. As an example, the derivative of the logistic function satisfies* (10.11) *with* $a = c = \mu = 1$.

Remark 18. *Theorem 14 states that while $\mathbf{w}_t$ diverges as t grows, its direction converges to the direction of the robust max-margin classifier. We should note that this convergence is quite slow. Figure 10.2 depicts the convergence of the direction of GD iterates to the RM classifier as $t \to \infty$ where it can be observed that the convergence becomes slow as t grows (the horizontal axis has a logarithmic scale). In our proof in Section 10.6, we theoretically establish that the rate of convergence is logarithmic.*

10.5 Proof of Lemma 29

In our proof, we use the following lemma which characterizes the behavior of gradient descent iterates on smooth functions.

Lemma 30 (Lemma 10 in [120])**.** *Let $\mathcal{L}(\mathbf{w})$ be a γ-smooth non-negative objective. If $\eta < \frac{2}{\gamma}$, then, for any starting point $\mathbf{w}(0)$, with the GD sequence*

$$\mathbf{w}(t+1) = \mathbf{w}(t) - \eta \nabla \mathcal{L}(\mathbf{w}(t)),$$

we have that:

$$\sum_{u=0}^{\infty} \|\nabla \mathcal{L}(\mathbf{w}(u))\|^2 < +\infty.$$

We also use the following corollary.

Corollary 8. *For any positive constant $C < \beta\big(\sigma_{max}(\mathbf{X}) + \|\boldsymbol{\epsilon}\|\big)^2$, there exist $R > 0$, such that $\big\|\nabla^2 \mathcal{L}_{\boldsymbol{\epsilon}}(\mathbf{w})\big\| < C$ when $\|\mathbf{w}\| > R$.*

The proof is straightforward, by computing the Hessian of $\mathcal{L}_{\boldsymbol{\epsilon}}(\cdot)$ and using the fact that $\ell(\cdot)$ is twice-differentiable and β-smooth.

Since the function $\ell(\cdot)$ is monotonically decreasing, we can write,

$$\mathcal{L}_{\boldsymbol{\epsilon}}(\mathbf{w}) = \sum_{i=1}^{n} \ell\left(y_i \mathbf{x}_i^T \mathbf{w} - \epsilon_i \|\mathbf{w}\|\right) . \tag{10.13}$$

The gradient of the loss function can be computed as,

$$\nabla \mathcal{L}_{\boldsymbol{\epsilon}}(\mathbf{w}) = \sum_{i=1}^{n} \ell'(y_i \mathbf{x}_i^T \mathbf{w} - \epsilon_i \|\mathbf{w}\|)(y_i \mathbf{x}_i - \epsilon_i \frac{\mathbf{w}}{\|\mathbf{w}\|}). \tag{10.14}$$

Consider the sequence $s_t := \frac{1}{\eta} \mathbf{w}_{RM}^T \mathbf{w}_t$ for $t \in \mathbb{N}$. First, we show that this sequence is increasing.

$$\begin{aligned} s_t - s_{t+1} &= \mathbf{w}_{RM}^T \nabla \mathcal{L}_{\boldsymbol{\epsilon}}(\mathbf{w}_t) \\ &= \sum_{i=1}^{n} \ell'(y_i \mathbf{x}_i^T \mathbf{w}_t - \epsilon_i \|\mathbf{w}_t\|) \mathbf{w}_{RM}^T \big(y_i \mathbf{x}_i - \epsilon_i \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|}\big) \\ &\leq \sum_{i=1}^{n} \ell'(y_i \mathbf{x}_i^T \mathbf{w}_t - \epsilon_i \|\mathbf{w}_t\|) \big(y_i \mathbf{x}_i^T \mathbf{w}_{RM} - \epsilon_i \|\mathbf{w}_{RM}\|\big) \\ &\leq \sum_{i=1}^{n} \ell'(y_i \mathbf{x}_i^T \mathbf{w}_t - \epsilon_i \|\mathbf{w}_t\|) < 0 \,, \end{aligned} \tag{10.15}$$

where for the first inequality, we used the fact that $\ell'(u) < 0$ and Cauchy-Schwartz, and for the second inequality, we used the constraints of the optimization (10.6).

Since $\{s_t\}_{t \geq 0}$ is an increasing sequence in $\mathbb{R}$, it either grows to $+\infty$ or approaches a limit value. We analyze each of these cases separately.

Case 1 : $\lim_{t \to \infty} s_t = L < +\infty$

When the sequence has a limit, we have $\lim_{t \to \infty} s_t - s_{t+1} = 0$. From the last inequality in (10.15), this implies that as $t \to \infty$,

$$\ell'(y_i \mathbf{x}_i^T \mathbf{w}_t - \epsilon_i \|\mathbf{w}_t\|) \to 0, \text{ for } 1 \leq i \leq n. \tag{10.16}$$

Since $\ell'(u)$ is negative for $u \in \mathbb{R}$, we must have

$$y_i \mathbf{x}_i^T \mathbf{w}_t - \epsilon_i \|\mathbf{w}_t\| \to +\infty, \text{ for } 1 \leq i \leq n, \tag{10.17}$$

which is (iii). This also implies that $\|\mathbf{w}_t\| \to \infty$. Finally, from (10.14), we have that $\nabla \mathcal{L}_{\boldsymbol{\epsilon}}(\mathbf{w}_t) \to \mathbf{0}_p$.

Case 2 : $\lim_{t \to \infty} s_t = +\infty$

$\|\mathbf{w}_t\| \geq \frac{\eta s_t}{\|\mathbf{w}_{RM}\|}$ implies that $\lim_{t \to \infty} \|\mathbf{w}_t\| = +\infty$. Using Corollary 8, for any constant $C < \beta\big(\sigma_{max}(\mathbf{X}) + \|\boldsymbol{\epsilon}\|\big)^2$, there exists a non-negative integer t_0 such that the second derivative is bounded

by C for any $t > t_0$. Hence, we can use the result of Lemma 30 with $\eta < 2 \cdot \beta^{-1} \cdot (\sigma_{\max}(\mathbf{X}) + \|\boldsymbol{\epsilon}\|)^{-2}$ which gives $\|\nabla \mathcal{L}_{\boldsymbol{\epsilon}}(\mathbf{w}_t)\| \to 0$ as $t \to +\infty$.

In order to show (iii), we use the last inequality in (10.15), as $t \to \infty$ since $\mathbf{w}_{RM}^T \nabla \mathcal{L}_{\boldsymbol{\epsilon}}(\mathbf{w}_t) \to 0$, we have:

$$\ell'(y_i \mathbf{x}_i^T \mathbf{w}_t - \epsilon_i \|\mathbf{w}_t\|) \to 0, \text{ for } 1 \leq i \leq n, \tag{10.18}$$

which gives the desired result.

10.6 Proof of Theorem 14

For the RM classifier, we define the set of support vectors as:

$$\mathcal{S} = \mathcal{S}_{RM} := \{i \in [n] : y_i \mathbf{x}_i^T \mathbf{w}_{RM} = 1 + \epsilon_i \|\mathbf{w}_{RM}\|\} . \tag{10.19}$$

First, we consider the KKT conditions for the optimization (10.6) which gives:

$$\mathbf{w}_{RM} = \sum_{i \in \mathcal{S}} \alpha_i \big(y_i \mathbf{x}_i - \epsilon_i \hat{\mathbf{w}}\big), \tag{10.20}$$

where $\hat{\mathbf{w}} := \frac{\mathbf{w}_{RM}}{\|\mathbf{w}_{RM}\|}$ and $\alpha_i \geq 0$. It can be shown that when the data points are drawn from a continuous distribution, for almost every data set, the support vectors are linearly independent and α_i's are all positive (see also [73] and Appendix B in [120]). Given the fact that $-\ell'(u)$ has an exponential tail, we assume that $\alpha, \gamma, \tau, \mu$ are positive constants such that:

$$\begin{cases} -\ell'(u) \leq \gamma\big(1 + \exp(-\mu \cdot u)\big) \exp(-\alpha \cdot u), \text{ and,} \\ -\ell'(u) \geq \gamma\big(1 - \exp(-\mu \cdot u)\big) \exp(-\alpha \cdot u), \end{cases} \tag{10.21}$$

for every $u \geq \tau$.

We define a vector $\tilde{\mathbf{w}}$ such that:

$$\exp\left(\tilde{\mathbf{w}}^T (y_i \mathbf{x}_i - \epsilon_i \hat{\mathbf{w}})\right) := \frac{\alpha_i}{\gamma \cdot \eta} , \quad \text{for } i = 1, 2, \ldots, n. \tag{10.22}$$

Recall that the gradient descent iterates are defined as,

$$\mathbf{w}_{t+1} - \mathbf{w}_t = -\eta \nabla \mathcal{L}\big(\mathbf{w}_t\big),\ t \in \mathbb{N}. \tag{10.23}$$

Next, for $t \geq 0$, we define the residual vector $\mathbf{r}_t \in \mathbb{R}^p$.

$$\mathbf{r}_t := \mathbf{w}_t - \frac{1}{\alpha}\log(t)\mathbf{w}_{RM} - \tilde{\mathbf{w}}. \tag{10.24}$$

In our proof, we adopt a similar strategy as [120] and bound the norm of the residual vector $\|\mathbf{r}(t)\|$ by a constant C for every $t \geq 1$. Consider the following equation,

$$\|\mathbf{r}_{t+1}\|^2 - \|\mathbf{r}_t\|^2 = \|\mathbf{r}_{t+1} - \mathbf{r}_t\|^2 + 2\,\mathbf{r}_t^T\big(\mathbf{r}_{t+1} - \mathbf{r}_t\big). \tag{10.25}$$

We bound each of the two terms in the RHS of (10.25).

We start with bounding the first term in the (10.25). We have,

$$\begin{aligned}
\|\mathbf{r}_{t+1} - \mathbf{r}_t\|^2 &= \left\|\mathbf{w}_{t+1} - \mathbf{w}_t - \mathbf{w}_{RM}\big(\log(\frac{t+1}{t})/\alpha\big)\right\|^2 \\
&\leq \eta^2 \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2 + (\alpha t)^{-2}\|\mathbf{w}_{RM}\|^2 \\
&\quad + 2(\eta/\alpha)\log(1 + t^{-1})\mathbf{w}_{RM}{}^T \nabla \mathcal{L}(\mathbf{w}_t) \\
&\leq \eta^2 \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2 + (\alpha t)^{-2}\|\mathbf{w}_{RM}\|^2\ ,
\end{aligned} \tag{10.26}$$

where in the first inequality, we replaced $\mathbf{w}_{t+1} - \mathbf{w}_t$ using the gradient descent iterates (10.23) along with $\log(1 + u) \leq u$, and in the second inequality, we exploit the inequality (10.15) that gives $\hat{\mathbf{w}}^T \nabla \mathcal{L}(\mathbf{w}(t)) < 0$.

Since the norm of $\mathbf{w}_t$ approaches infinity as t grows, when $\eta < 2 \cdot \beta^{-1} \cdot (\sigma_{\max}(\mathbf{X}) + \|\boldsymbol{\epsilon}\|)^{-2}$, we can use the result of Corollary 8 and Lemma 30 to have:

$$\sum_{t=0}^{\infty} \|\nabla \mathcal{L}(\mathbf{w}_t)\| < C_1, \tag{10.27}$$

for some constant $C_1 > 0$. Therefore, we can bound the sum over the first term in (10.25).

$$\sum_{t \geq 1} ||\mathbf{r}_{t+1} - \mathbf{r}_t||^2 \leq \eta^2 C_1 + \alpha^{-2}\|\mathbf{w}_{RM}\|^2 \sum_{t \geq 1} t^{-2} < C_2. \tag{10.28}$$

Next, we will bound the second term in (10.25), i.e., $\mathbf{r}_t^T\left(\mathbf{r}_{t+1}-\mathbf{r}_t\right)$. To do so, we first define the constant θ as follows:

$$\theta := \min_{i\in\mathcal{S}^c} y_i\mathbf{x}_i\mathbf{w}_{RM} - \epsilon_i \left\|\mathbf{w}_{RM}\right\| > 1, \tag{10.29}$$

where $\mathcal{S}^c = [n] - \mathcal{S}$ indicates the indices of non-support vectors. The following lemma provides an upper bound on $\mathbf{r}_t^T\left(\mathbf{r}_{t+1}-\mathbf{r}_t\right)$ for $t \geq 1$.

Lemma 31. *With the assumptions of Theorem 14, consider the gradient descent iterates* (10.23)*, $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$, and the vector $\mathbf{r}_t$ defined in* (10.24)*. Then, for constants $C \geq 0$ and $t_0 \in \mathbb{N}$, we have:*

$$\mathbf{r}_t^T\left(\mathbf{r}_{t+1}-\mathbf{r}_t\right) \leq Ct^{-\min(\theta,1+\frac{\mu}{2\alpha})}\ ,\quad \forall t \geq t_0. \tag{10.30}$$

Using the result of Lemma 31, since $\theta > 1$ and $\mu/\alpha > 0$, we have:

$$\begin{aligned}\sum_{t\geq 0}\mathbf{r}_t^T\left(\mathbf{r}_{t+1}-\mathbf{r}_t\right) &< \sum_{t=1}^{t_0-1}\mathbf{r}_t^T\left(\mathbf{r}_{t+1}-\mathbf{r}_t\right) + C\sum_{t\geq t_0} t^{-\min(\theta,1+\frac{\mu}{2\alpha})}\\ &< C_3.\end{aligned} \tag{10.31}$$

Therefore, from (10.25), (10.28), and (10.31), we have,

$$\left\|\mathbf{r}_k\right\|^2 = \left\|\mathbf{r}_1\right\|^2 + \sum_{t=1}^{k-1}\left\|\mathbf{r}_{t+1}\right\|^2 - \left\|\mathbf{r}_t\right\|^2 < C_4\ ,\quad \forall k \geq 1, \tag{10.32}$$

for a positive constant C_4. Consequently, from (10.24), we have,

$$\left\|\mathbf{w}_t - \frac{1}{\alpha}\log(t)\mathbf{w}_{RM}\right\| \leq C_4 + \left\|\tilde{\mathbf{w}}\right\|\ . \tag{10.33}$$

By some straightforward calculations we can get,

$$\left\|\frac{\mathbf{w}_t}{\left\|\mathbf{w}_t\right\|} - \frac{\mathbf{w}_{RM}}{\left\|\mathbf{w}_{RM}\right\|}\right\|^2 \leq 2\Big[\frac{\alpha(C_4+\left\|\tilde{\mathbf{w}}\right\|)}{\log(t)\left\|\mathbf{w}_{RM}\right\|}\Big]^2, \tag{10.34}$$

which gives the desired result, i.e.,

$$\lim_{t\to\infty}\left\|\frac{\mathbf{w}_t}{\left\|\mathbf{w}_t\right\|} - \frac{\mathbf{w}_{RM}}{\left\|\mathbf{w}_{RM}\right\|}\right\| = 0. \tag{10.35}$$

[1] Ehsan Abbasi, Fariborz Salehi, and Babak Hassibi. "Performance analysis of convex data detection in mimo". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 4554–4558.

[2] Ehsan Abbasi, Fariborz Salehi, and Babak Hassibi. "Universality in learning from linear measurements". In: *arXiv preprint arXiv:1906.08396* (2019).

[3] Ehsan Abbasi, Christos Thrampoulidis, and Babak Hassibi. "General performance metrics for the LASSO". In: *Information Theory Workshop (ITW), 2016 IEEE*. IEEE. 2016, pp. 181–185.

[4] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton, NJ: Princeton University Press, 2008.

[5] J. Kennedy et al. "Super-resolution in PET imaging". In: *Medical Imaging, IEEE Transaction on* 25(2) (2006), pp. 137–147.

[6] M. Dierolf et al. "Ptychographic x-ray computed tomography". In: *Nature* 467 (2010), pp. 436–440.

[7] Dennis Amelunxen et al. "Living on the edge: Phase transitions in convex programs with random data". In: *Information and Inference: A Journal of the IMA* 3.3 (2014), pp. 224–294.

[8] Sohail Bahmani and Justin Romberg. "Efficient compressive phase retrieval with constrained sensing vectors". In: *Advances in Neural Information Processing Systems*. 2015, pp. 523–531.

[9] Sohail Bahmani and Justin Romberg. "Phase retrieval meets statistical learning theory: A flexible convex relaxation". In: *Artificial Intelligence and Statistics*. 2017, pp. 252–260.

[10] Milad Bakhshizadeh, Arian Maleki, and Shirin Jalali. "Compressive phase retrieval of structured signals". In: *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2018, pp. 2291–2295.

[11] Radu Balan, Pete Casazza, and Dan Edidin. "On signal reconstruction without phase". In: *Applied and Computational Harmonic Analysis* 20.3 (2006), pp. 345–356.

[12] Peter L Bartlett et al. "Benign overfitting in linear regression". In: *arXiv preprint arXiv:1906.11300* (2019).

[13] Mohsen Bayati and Andrea Montanari. "The LASSO risk for Gaussian matrices". In: *IEEE Transactions on Information Theory* 58.4 (2012), pp. 1997–2017.

[14] Mikhail Belkin, Daniel J Hsu, and Partha Mitra. "Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate". In: *Advances in Neural Information Processing Systems*. 2018, pp. 2300–2311.

[15] Mikhail Belkin et al. “Reconciling modern machine-learning practice and the classical bias–variance trade-off”. In: *Proceedings of the National Academy of Sciences* 116.32 (2019), pp. 15849–15854.

[16] M. Berger and B. Gostiaux. *Differential geometry: Manifolds, curves, and surfaces*. Graduate texts in mathematics. Springer-Verlag, 1988.

[17] Battista Biggio et al. “Evasion attacks against machine learning at test time”. In: *Joint European conference on machine learning and knowledge discovery in databases*. Springer. 2013, pp. 387–402.

[18] Carl R Boyd, Mary Ann Tolson, and Wayne S Copes. “Evaluating trauma care: the TRISS method. Trauma Score and the Injury Severity Score.” In: *The Journal of trauma* 27.4 (1987), pp. 370–378.

[19] Florentina Bunea et al. “Honest variable selection in linear and logistic regression models via 1 and 1+ 2 penalization”. In: *Electronic Journal of Statistics* 2 (2008), pp. 1153–1194.

[20] T Tony Cai, Xiaodong Li, Zongming Ma, et al. “Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow”. In: *The Annals of Statistics* 44.5 (2016), pp. 2221–2251.

[21] T Tony Cai, Anru Zhang, et al. “ROP: Matrix recovery via rank-one projections”. In: *The Annals of Statistics* 43.1 (2015), pp. 102–138.

[22] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. “Phase retrieval from coded diffraction patterns”. In: *Applied and Computational Harmonic Analysis* 39.2 (2015), pp. 277–299.

[23] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. “Phase retrieval from coded diffraction patterns”. In: *Applied and Computational Harmonic Analysis* 39.2 (2015), pp. 277–299.

[24] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. “Phase retrieval via Wirtinger flow: Theory and algorithms”. In: *IEEE Transactions on Information Theory* 61.4 (2015), pp. 1985–2007.

[25] Emmanuel J Candes, Thomas Strohmer, and Vladislav Voroninski. “Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming”. In: *Communications on Pure and Applied Mathematics* 66.8 (2013), pp. 1241–1274.

[26] Emmanuel J Candes and Terence Tao. “Decoding by linear programming”. In: *IEEE Transactions on Information Theory* 51.12 (2005), pp. 4203–4215.

[27] Emmanuel J Candes et al. “Phase retrieval via matrix completion”. In: *SIAM review* 57.2 (2015), pp. 225–251.

[28] Emmanuel J Candès and Carlos Fernandez-Granda. “Towards a mathematical theory of super-resolution”. In: *Communications on Pure and Applied Mathematics* 67.6 (2014), pp. 906–956.

[29] Emmanuel J Candès and Benjamin Recht. "Exact matrix completion via convex optimization". In: *Foundations of Computational mathematics* 9.6 (2009), pp. 717–772.

[30] Emmanuel J Candès, Justin Romberg, and Terence Tao. "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information". In: *IEEE Transactions on information theory* 52.2 (2006), pp. 489–509.

[31] Emmanuel J Candès and Pragya Sur. "The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression". In: *arXiv preprint arXiv:1804.09753* (2018).

[32] Nicholas Carlini and David Wagner. "Towards evaluating the robustness of neural networks". In: *2017 IEEE Symposium on Security and Privacy (sp)*. IEEE. 2017, pp. 39–57.

[33] Venkat Chandrasekaran et al. "The convex geometry of linear inverse problems". In: *Foundations of Computational Mathematics* 12.6 (2012), pp. 805–849.

[34] Yuxin Chen and Emmanuel Candes. "Solving random quadratic systems of equations is nearly as easy as solving linear systems". In: *Advances in Neural Information Processing Systems*. 2015, pp. 739–747.

[35] Yuxin Chen, Yuejie Chi, and Andrea J Goldsmith. "Exact and stable covariance estimation from quadratic sampling via convex programming". In: *IEEE Transactions on Information Theory* 61.7 (2015), pp. 4034–4059.

[36] Aldo Conca et al. "An algebraic characterization of injectivity in phase retrieval". In: *Applied and Computational Harmonic Analysis* 38.2 (2015), pp. 346–356.

[37] Corinna Cortes and Vladimir Vapnik. "Support-vector networks". In: *Machine learning* 20.3 (1995), pp. 273–297.

[38] Thomas M Cover. "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition". In: *IEEE transactions on electronic computers* 3 (1965), pp. 326–334.

[39] Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. "A model of double descent for high-dimensional binary linear classification". In: *arXiv preprint arXiv:1911.05822* (2019).

[40] Oussama Dhifallah and Yue M Lu. "Fundamental limits of PhaseMax for phase retrieval: A replica analysis". In: *arXiv preprint arXiv:1708.03355* (2017).

[41] Oussama Dhifallah, Christos Thrampoulidis, and Yue M Lu. "Phase retrieval via polytope optimization: Geometry, phase transitions, and new algorithms". In: *arXiv preprint arXiv:1805.09555* (2018).

[42] David Donoho and Andrea Montanari. "High dimensional robust m-estimation: Asymptotic variance via approximate message passing". In: *Probability Theory and Related Fields* 166.3-4 (2016), pp. 935–969.

[43] David L Donoho. "Compressed sensing". In: *IEEE Transactions on information theory* 52.4 (2006), pp. 1289–1306.

[44] David L Donoho, Arian Maleki, and Andrea Montanari. "Message-passing algorithms for compressed sensing". In: *Proceedings of the National Academy of Sciences* 106.45 (2009), pp. 18914–18919.

[45] David L Donoho, Arian Maleki, and Andrea Montanari. "The noise-sensitivity phase transition in compressed sensing". In: *IEEE Transactions on Information Theory* 57.10 (2011), pp. 6920–6941.

[46] David L Donoho and Jared Tanner. "Sparse nonnegative solution of underdetermined linear equations by linear programming". In: *Proceedings of the National Academy of Sciences* 102.27 (2005), pp. 9446–9451.

[47] Ahmed Douik, Fariborz Salehi, and Babak Hassibi. "A Novel Riemannian Optimization Approach and Algorithm for Solving the Phase Retrieval Problem". In: *2019 53rd Asilomar Conference on Signals, Systems, and Computers*. IEEE. 2019, pp. 1962–1966.

[48] Noureddine El Karoui et al. "On robust regression with high-dimensional predictors". In: *Proceedings of the National Academy of Sciences* 110.36 (2013), pp. 14557–14562.

[49] Melikasadat Emami et al. "Generalization error of generalized linear models in high dimensions". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 2892–2901.

[50] Melikasadat Emami et al. "Implicit Bias of Linear RNNs". In: *arXiv preprint arXiv:2101.07833* (2021).

[51] Melikasadat Emami et al. "Low-rank nonlinear decoding of μ-ECoG from the primary auditory cortex". In: *arXiv preprint arXiv:2005.05053* (2020).

[52] C Fienup and J Dainty. "Phase retrieval and image reconstruction for astronomy". In: *Image Recovery: Theory and Application* (1987), pp. 231–275.

[53] James R Fienup. "Phase retrieval algorithms: A comparison". In: *Applied optics* 21.15 (1982), pp. 2758–2769.

[54] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. "Regularization paths for generalized linear models via coordinate descent". In: *Journal of statistical software* 33.1 (2010), p. 1.

[55] Cedric Gerbelot, Alia Abbara, and Florent Krzakala. "Asymptotic Errors for Teacher-Student Convex Generalized Linear Models (or: How to Prove Kabashima's Replica Formula)". In: *arXiv preprint arXiv:2006.06581* (2020).

[56] Ralph W Gerchberg. "A practical algorithm for the determination of phase from image and diffraction plane pictures". In: *Optik* 35 (1972), pp. 237–246.

[57] Michel X Goemans and David P Williamson. "Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming". In: *Journal of the ACM (JACM)* 42.6 (1995), pp. 1115–1145.

[58] Tom Goldstein and Christoph Studer. "PhaseMax: Convex phase retrieval via basis pursuit". In: *IEEE Transactions on Information Theory* (2018).

[59] Hayit Greenspan. “Super-resolution in medical imaging”. In: *The computer journal* 52.1 (2009), pp. 43–63.

[60] Paul Hand and Vladislav Voroninski. “An elementary proof of convex phase retrieval in the natural parameter space via the linear program Phasemax”. In: *arXiv preprint arXiv:1611.03935* (2016).

[61] Paul Hand and Vladislav Voroninski. “Compressed sensing from phaseless gaussian measurements via linear programming in the natural parameter space”. In: *arXiv preprint arXiv:1611.05985* (2016).

[62] Trevor Hastie et al. “Surprises in high-dimensional ridgeless least squares interpolation”. In: *arXiv preprint arXiv:1903.08560* (2019).

[63] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*. Vol. 398. John Wiley & Sons, 2013.

[64] Kishore Jaganathan. “Convex programming-based phase retrieval: Theory and applications”. In: Ph.D. Dissertion, California Institute of Technology. 2016.

[65] Kishore Jaganathan, Yonina Eldar, and Babak Hassibi. “Phase retrieval with masks using convex optimization”. In: *2015 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2015, pp. 1655–1659.

[66] Kishore Jaganathan, Yonina C Eldar, and Babak Hassibi. “Phase retrieval: An overview of recent developments”. In: *arXiv preprint arXiv:1510.07713* (2015).

[67] Kishore Jaganathan, Samet Oymak, and Babak Hassibi. “Recovery of sparse 1-D signals from the magnitudes of their Fourier transform”. In: *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium On*. IEEE. 2012, pp. 1473–1477.

[68] Kishore Jaganathan, Samet Oymak, and Babak Hassibi. “Sparse phase retrieval: Uniqueness guarantees and recovery algorithms”. In: *IEEE Transactions on Signal Processing* 65.9 (2017), pp. 2402–2410.

[69] Kishore Jaganathan et al. “Phaseless super-resolution using masks”. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2016, pp. 4039–4043.

[70] 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*. 2013, pp. 665–674.

[71] Gareth James et al. *An introduction to statistical learning*. Vol. 112. Springer, 2013.

[72] Adel Javanmard and Mahdi Soltanolkotabi. “Precise statistical analysis of classification accuracies for adversarial training”. In: *arXiv preprint arXiv:2010.11213* (2020).

[73] Ziwei Ji and Matus Telgarsky. “Risk and parameter convergence of logistic regression”. In: *arXiv preprint arXiv:1803.07300* (2018).

[74] Jürgen Jost and Jèurgen Jost. *Riemannian geometry and geometric analysis*. Vol. 42005. Springer, 2008.

[75] Abderrahim Jourani, Lionel Thibault, and Dariusz Zagrodny. "Differential properties of the Moreau envelope". In: *Journal of Functional Analysis* 266.3 (2014), pp. 1185–1237.

[76] Sham Kakade et al. "Learning exponential families in high-dimensions: Strong convexity and sparsity". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. 2010, pp. 381–388.

[77] Leonid Genrikhovich Khachiyan. "A polynomial algorithm in linear programming". In: *Doklady Akademii Nauk*. Vol. 244. 5. Russian Academy of Sciences. 1979, pp. 1093–1096.

[78] Gary King and Langche Zeng. "Logistic regression in rare events data". In: *Political analysis* 9.2 (2001), pp. 137–163.

[79] Ganesh Kini and Christos Thrampoulidis. "Analytic Study of Double Descent in Binary Classification: The Impact of Loss". In: *arXiv preprint arXiv:2001.11572* (2020).

[80] Kwangmoo Koh, Seung-Jean Kim, and Stephen Boyd. "An interior-point method for large-scale l1-regularized logistic regression". In: *Journal of Machine learning research* 8.Jul (2007), pp. 1519–1555.

[81] Balaji Krishnapuram et al. "Sparse multinomial logistic regression: Fast algorithms and generalization bounds". In: *IEEE transactions on pattern analysis and machine intelligence* 27.6 (2005), pp. 957–968.

[82] Hyounghan Kwon et al. "Computational complex optical field imaging using a designed metasurface diffuser". In: *Optica* 5.8 (2018), pp. 924–931.

[83] J. Lee. *Introduction to Topological Manifolds*. Graduate Texts in Mathematics. Springer New York, 2010.

[84] Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[85] Xiaodong Li and Vladislav Voroninski. "Sparse signal recovery from quadratic measurements via convex programming". In: *SIAM Journal on Mathematical Analysis* 45.5 (2013), pp. 3019–3033.

[86] Xiaodong Li et al. "Rapid, robust, and reliable blind deconvolution via nonconvex optimization". In: *Applied and Computational Harmonic Analysis* 47.3 (2019), pp. 893–934.

[87] Yuanxin Li, Yue Sun, and Yuejie Chi. "Low-rank positive semidefinite matrix recovery from corrupted rank-one measurements". In: *IEEE Transactions on Signal Processing* 65.2 (2016), pp. 397–408.

[88] Yue M Lu and Gen Li. "Phase transitions of spectral initialization for high-dimensional nonconvex estimation". In: *arXiv preprint arXiv:1702.06435* (2017).

[89] Junjie Ma, Ji Xu, and Arian Maleki. "Optimization-based AMP for phase retrieval: The impact of initialization and ℓ_2-regularization". In: *arXiv preprint arXiv:1801.01170* (2018).

[90] Henrik Madsen and Poul Thyregod. *Introduction to general and generalized linear models*. CRC Press, 2010.

[91] Olvi L Mangasarian and Benjamin Recht. "Probability of unique integer solution to a system of linear equations". In: *European Journal of Operational Research* 214.1 (2011), pp. 27–30.

[92] Song Mei and Andrea Montanari. "The generalization error of random features regression: Precise asymptotics and double descent curve". In: *arXiv preprint arXiv:1908.05355* (2019).

[93] Rick P Millane. "Phase retrieval in crystallography and optics". In: *JOSA A* 7.3 (1990), pp. 394–411.

[94] Marco Mondelli and Andrea Montanari. "Fundamental limits of weak recovery with applications to phase retrieval". In: *arXiv preprint arXiv:1708.05932* (2017).

[95] Andrea Montanari et al. "The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime". In: *arXiv preprint arXiv:1911.01544* (2019).

[96] John Ashworth Nelder and Robert WM Wedderburn. "Generalized linear models". In: *Journal of the Royal Statistical Society: Series A (General)* 135.3 (1972), pp. 370–384.

[97] Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. "Phase retrieval using alternating minimization". In: *Advances in Neural Information Processing Systems*. 2013, pp. 2796–2804.

[98] Samet Oymak and Mahdi Soltanolkotabi. "Overparameterized nonlinear learning: Gradient descent takes the shortest path?" In: *International Conference on Machine Learning*. PMLR. 2019, pp. 4951–4960.

[99] Samet Oymak, Christos Thrampoulidis, and Babak Hassibi. "The squared-error of generalized lasso: A precise analysis". In: *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE. 2013, pp. 1002–1009.

[100] Samet Oymak and Joel A Tropp. "Universality laws for randomized dimension reduction, with applications". In: *Information and Inference: A Journal of the IMA* 7.3 (2017), pp. 337–446.

[101] Samet Oymak et al. "Simultaneously structured models with application to sparse and low-rank matrices". In: *IEEE Transactions on Information Theory* 61.5 (2015), pp. 2886–2908.

[102] P. Pal and P. P. Vaidyanathan. "Nested arrays: a novel approach to array processing with enhanced degrees of freedom". In: *Signal Processing, IEEE Transactions on* 58.8 (2010), pp. 4167–4181.

[103] Ashkan Panahi and Babak Hassibi. "A universal analysis of large-scale regularized least squares solutions". In: *Advances in Neural Information Processing Systems*. 2017, pp. 3381–3390.

[104] Neal Parikh, Stephen Boyd, et al. "Proximal algorithms". In: *Foundations and Trends® in Optimization* 1.3 (2014), pp. 127–239.

[105] P. Petersen. *Riemannian Geometry*. Graduate texts in mathematics. Springer, 1998.

[106] K. G. Puschmann and F. Kneer. "On super-resolution in astronomical imaging". In: *Astronomy and Astrophysics* 436 (2005), pp. 373–378.

[107] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization". In: *SIAM Review* 52.3 (2010), pp. 471–501.

[108] Ralph Tyrell Rockafellar. *Convex analysis*. Princeton university press, 2015.

[109] R. Roy and T. Kailath. "ESPRIT-estimation of signal parameters via rotational invariance techniques". In: *IEEE Transactions on Acoustics, Speech and Signal Processing* (1989), pp. 984–995.

[110] Mark Rudelson and Roman Vershynin. "Sparse reconstruction by convex relaxation: Fourier and Gaussian measurements". In: *Information Sciences and Systems, 2006 40th Annual Conference on*. IEEE. 2006, pp. 207–212.

[111] Walter Rudin et al. *Principles of mathematical analysis*. Vol. 3. McGraw-hill New York, 1976.

[112] Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. "A precise analysis of phasemax in phase retrieval". In: *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2018, pp. 976–980.

[113] Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. "The impact of regularization on high-dimensional logistic regression". In: *Advances in Neural Information Processing Systems*. 2019, pp. 11982–11992.

[114] Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. "The performance analysis of generalized margin maximizers on separable data". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 8417–8426.

[115] Fariborz Salehi, Kishore Jaganathan, and Babak Hassibi. "Multiple illumination phaseless super-resolution (MIPS) with applications to phaseless DoA estimation and diffraction imaging". In: *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE. 2017, pp. 3949–3953.

[116] R. O. Schmidt. "Multiple emitter location and signal parameter estimation". In: *Antennas and Propagation, IEEE Transactions on* (1986), pp. 276–280.

[117] Yoav Shechtman et al. "Phase retrieval with application to optical imaging: A contemporary overview". In: *IEEE Signal Processing Magazine* 32.3 (2015), pp. 87–109.

[118] Shirish Krishnaj Shevade and S Sathiya Keerthi. "A simple and efficient algorithm for gene selection using sparse logistic regression". In: *Bioinformatics* 19.17 (2003), pp. 2246–2253.

[119] Mahdi Soltanolkotabi. "Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization". In: *arXiv preprint arXiv:1702.06175* (2017).

[120] Daniel Soudry et al. "The implicit bias of gradient descent on separable data". In: *The Journal of Machine Learning Research* 19.1 (2018), pp. 2822–2878.

[121] Mihailo Stojnic. "A framework to characterize performance of lasso algorithms". In: *arXiv preprint arXiv:1303.7291* (2013).

[122] Mihailo Stojnic. "Various thresholds for l1-optimization in compressed sensing". In: (2009).

[123] Pragya Sur and Emmanuel J Candès. "A modern maximum-likelihood theory for high-dimensional logistic regression". In: *arXiv preprint arXiv:1803.06964* (2018).

[124] Pragya Sur, Yuxin Chen, and Emmanuel J Candès. "The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square". In: *Probability Theory and Related Fields* (2017), pp. 1–72.

[125] Christian Szegedy et al. "Intriguing properties of neural networks". In: *arXiv preprint arXiv:1312.6199* (2013).

[126] Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis. "Sharp guarantees for solving random equations with one-bit information". In: *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE. 2019, pp. 765–772.

[127] Gongguo Tang et al. "Compressed sensing off the grid". In: *IEEE transactions on information theory* 59.11 (2013), pp. 7465–7490.

[128] Christos Thrampoulidis. "Recovering structured signals in high dimensions via non-smooth convex optimization: Precise performance analysis". PhD thesis. California Institute of Technology, 2016.

[129] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. "Precise error analysis of regularized M-estimators in high dimensions". In: *IEEE Transactions on Information Theory* 64.8 (2018), pp. 5592–5628.

[130] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. "Regularized linear regression: A precise analysis of the estimation error". In: *Conference on Learning Theory*. 2015, pp. 1683–1709.

[131] Christos Thrampoulidis et al. "BER analysis of the box relaxation for BPSK signal recovery". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2016, pp. 3776–3780.

[132] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[133] Joel A Tropp. “Convex recovery of a structured signal from independent random linear measurements”. In: *Sampling Theory, a Renaissance*. Springer, 2015, pp. 67–101.

[134] Joel A Tropp and Anna C Gilbert. “Signal recovery from random measurements via orthogonal matching pursuit”. In: *IEEE Transactions on information theory* 53.12 (2007), pp. 4655–4666.

[135] Jack V Tu. “Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes”. In: *Journal of clinical epidemiology* 49.11 (1996), pp. 1225–1231.

[136] T. E. Tuncer and B. Friedlander. “Classical and modern direction-of-arrival estimation”. In: *Academic Press* (2009).

[137] P. P. Vaidyanathan and P. Pal. “Sparse sensing with co-prime samplers and arrays”. In: *IEEE Transactions on Signal Processing* 59.2 (2011), pp. 573–586.

[138] Sara A Van de Geer et al. “High-dimensional generalized linear models and the lasso”. In: *The Annals of Statistics* 36.2 (2008), pp. 614–645.

[139] Aad W Van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge university press, 2000.

[140] V Vapnik. *Estimation of Dependences Based on Empirical Data Berlin*. 1982.

[141] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge University Press, 2018.

[142] Irène Waldspurger, Alexandre d?Aspremont, and Stéphane Mallat. “Phase recovery, maxcut and complex semidefinite programming”. In: *Mathematical Programming* 149.1-2 (2015), pp. 47–81.

[143] Philipp Walk, Peter Jung, and Babak Hassibi. “Constrained blind deconvolution using wirtinger flow methods”. In: *2017 International Conference on Sampling Theory and Applications (SampTA)*. IEEE. 2017, pp. 322–326.

[144] Adriaan Walther. “The question of phase retrieval in optics”. In: *Journal of Modern Optics* 10.1 (1963), pp. 41–49.

[145] Gang Wang, Georgios B Giannakis, and Yonina C Eldar. “Solving systems of random quadratic equations via truncated amplitude flow”. In: *IEEE Transactions on Information Theory* 64.2 (2018), pp. 773–794.

[146] Gang Wang et al. “Sparse phase retrieval via truncated amplitude flow”. In: *IEEE Transactions on Signal Processing* 66.2 (2016), pp. 479–491.

[147] Chris D White, Sujay Sanghavi, and Rachel Ward. “The local convexity of solving systems of quadratic equations”. In: *arXiv preprint arXiv:1506.07868* (2015).

[148] Shanshan Wu et al. “The sparse recovery autoencoder”. In: (2019).

[149] Ji Xu and Daniel Hsu. “How many variables should be entered in a principal component regression equation?” In: *arXiv preprint arXiv:1906.01139* (2019).

[150] Teng Zhang. “Phase retrieval using alternating minimization in a batch setting”. In: *arXiv preprint arXiv:1706.08167* (2017).

SOME TECHNICAL TOOLS

A.1 Convex Gaussian Min-max Theorem

Several times in this writing, we appeal to the recently developed Convex Gaussian Min-max Theorem (CGMT) [130] for analysis of optimization programs. The CGMT associates with a Primary Optimization (PO) problem an Auxiliary Optimization (AO) problem from which we can investigate various properties of the primary optimization, such as phase transitions. In particular, the (PO) and the (AO) problems are defined respectively as follows:

$$\Phi(\mathbf{G}) := \min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}} \max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}} \mathbf{u}^T\mathbf{G}\mathbf{w} + \psi(\mathbf{u},\mathbf{w}), \tag{PO}$$

$$\phi(\mathbf{g},\mathbf{h}) := \min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}} \max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}} \|\mathbf{w}\|\mathbf{g}^T\mathbf{u} - \|\mathbf{u}\|\mathbf{h}^T\mathbf{w} + \psi(\mathbf{u},\mathbf{w}), \tag{AO}$$

where $\mathbf{G}\in\mathbb{R}^{m\times n}, \mathbf{g}\in\mathbb{R}^m, \mathbf{h}\in\mathbb{R}^n, \mathcal{S}_{\mathbf{w}}\subset\mathbb{R}^n, \mathcal{S}_{\mathbf{u}}\subset\mathbb{R}^m$ and $\psi:\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}$. Denote $\mathbf{w}_\Phi := \mathbf{w}_\Phi(\mathbf{G})$ and $\mathbf{w}_\phi := \mathbf{w}_\phi(\mathbf{g},\mathbf{h})$ for any optimal minimizers in (PO) and (AO), respectively. The following Theorem establishes the connection between the two optimizations in the Gaussian setting.

Theorem 15 (CGMT)**.** *[128] In* (A.1)*, let $\mathcal{S}_{\mathbf{w}}$, $\mathcal{S}_{\mathbf{u}}$, be convex and compact sets, and assume that $\psi(\cdot,\cdot)$ is convex-concave on $\mathcal{S}_{\mathbf{w}}\times\mathcal{S}_{\mathbf{u}}$. Also assume that* $\mathbf{G}$, $\mathbf{g}$, *and* $\mathbf{h}$ *all have entries i.i.d. standard normal. The following statements are true:*

1. *For all $\mu\in\mathbb{R}$, and $t>0$,*

$$\mathbb{P}(|\Phi(\mathbf{G})-\mu|>t) \le 2\mathbb{P}(|\phi(\mathbf{g},\mathbf{h})-\mu|\ge t)\ . \tag{A.2}$$

2. *Let $\mathcal{S}$ be an arbitrary open subset of $\mathcal{S}_{\mathbf{w}}$ and $\mathcal{S}^c := \mathcal{S}_{\mathbf{w}}/\mathcal{S}$. Denote $\Phi_{\mathcal{S}^c}(\mathbf{G})$ and $\phi_{\mathcal{S}^c}(\mathbf{g},\mathbf{h})$ be the optimal costs of the optimizations in* (PO)*, and* (AO)*, respectively, when the minimization over $\mathbf{w}$ is now constrained over $\mathbf{w}\in\mathcal{S}^c$. If there exists constants $\bar{\phi}$, $\bar{\phi}_{\mathcal{S}^c}$, and $\eta>0$ such that,*

   - $\bar{\phi}_{\mathcal{S}^c} \ge \bar{\phi} + 3\eta$,
   - $\phi(\mathbf{g},\mathbf{h}) < \bar{\phi}+\eta$, *with probability at least* $1-p$,
   - $\phi_{\mathcal{S}^c}(\mathbf{g},\mathbf{h}) > \bar{\phi}_{\mathcal{S}^c} - \eta$, *with probability at least* $1-p$,

then, $\mathbb{P}(\mathbf{w}_\Phi(\mathbf{G}) \in \mathcal{S}) \geq 1 - 4p$.

The probabilities are taken with respect to the randomness in $\mathbf{G}$*,* $\mathbf{g}$*, and* $\mathbf{h}$*.*

We also state the following lemma which is a consequence of previous theorem in the asymptotic regime,

Lemma 32 (Asymptotic CGMT)**.** *[128] using the same notations and assumptions as in Theorem 15, suppose that there exist constants* $\bar{\phi} < \bar{\phi}_{\mathcal{S}^c}$ *such that* $\phi(\mathbf{g}, \mathbf{h}) \xrightarrow{p} \bar{\phi}$*, and* $\phi_{\mathcal{S}^c}(\mathbf{g}, \mathbf{h}) \longrightarrow \bar{\phi}_{\mathcal{S}^c}$*. Then,*

$$\lim_{n\to\infty} \mathbb{P}(\mathbf{w}_\Phi(\mathbf{G}) \in \mathcal{S}) = 1 \ . \tag{A.3}$$

We refer the interested reader to [128, 130, 129] for further reading on the subject, its premises, and applications.

A.2 Useful Technical Lemmas

We gathered here some useful lemmas that are used in the proof of our main results.

In the analysis of the auxiliary optimization, we replace several functions with their limits in probability. This can be done through the same tricks used in section A.4 of [129] and Lemma B.1 in the same paper. The following lemma is used in our analysis of (AO) optimization and allows us (when the conditions are satisfied) to replace the objective with the function it converges to in the asymptotic regime. Here, we state this lemma without proof.

Lemma 33 (Min-convergence – Open Sets)**.** *Consider a sequence of proper, convex stochastic functions* $M_n : (0, \infty) \to \mathbb{R}$*, and a deterministic function* $M : (0, \infty) \to \mathbb{R}$*, such that:*

1. $M_n(x) \xrightarrow{P} M(x)$*, for all* $x > 0$*,*
2. *there exists* $z > 0$ *such that* $M(x) > \inf_{x>0} M(x)$ *for all* $x \geq z$*.*

Then, $\inf_{x>0} M_n(x) \xrightarrow{P} \inf_{x>0} M(x)$.

The next lemma provides the partial derivatives of the Moreau envelope function.

Lemma 34. *Let $\Phi : \mathbb{R}^d \to \mathbb{R}$ be a convex function. For $\mathbf{v} \in \mathbb{R}^d$ and $t \in \mathbb{R}_+$, the* Moreau envelope *function is defined as,*

$$M_{\Phi(\cdot)}(\mathbf{v}, t) = \min_{\mathbf{x}\in\mathbb{R}^d} \Phi(\mathbf{x}) + \frac{1}{2t}||\mathbf{x} - \mathbf{v}||^2 \,, \tag{A.4}$$

and the proximal operator *is the solution to this optimization, i.e.,*

$$Prox_{t\Phi(\cdot)}(\mathbf{v}) = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \Phi(\mathbf{x}) + \frac{1}{2t}||\mathbf{x} - \mathbf{v}||^2 \,. \tag{A.5}$$

The derivative of the Moreau envelope function can be computed as follows,

$$\frac{\partial M_{\Phi(\cdot)}}{\partial \mathbf{v}} = \frac{1}{t}(\mathbf{v} - Prox_{t\Phi(\cdot)}(\mathbf{v})) \;\;, \qquad \frac{\partial M_{\Phi(\cdot)}}{\partial t} = -\frac{1}{2t^2}(\mathbf{v} - Prox_{t\Phi(\cdot)}(\mathbf{v}))^2 \,. \tag{A.6}$$

We refer the interested reader to [75] for the proof as well as a detailed study of the properties of the Moreau envelope.

Lemma 35 (Stein's lemma)**.** *[141] For a function $f : \mathbb{R} \to \mathbb{R}$, we have $\mathbb{E}_Z[Zf(Z)] = \mathbb{E}_Z[f'(Z)]$.*

Lemma 36. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be an invariantly separable function such that $f(\mathbf{x}) = \sum_{i=1}^d \tilde{f}(x_i)$ for all $\mathbf{x} \in \mathbb{R}^d$, where $\tilde{f}$ is a real-valued function. Then, we have:*

$$M_{f(\cdot)}(\mathbf{v}, t) = \sum_{i=1}^{d} M_{\tilde{f}(\cdot)}(v_i, t) \;\;, \quad \textit{and} \quad Prox_{tf(\cdot)}(\mathbf{v}) = \begin{bmatrix} Prox_{t\tilde{f}(\cdot)}(v_1) \\ Prox_{t\tilde{f}(\cdot)}(v_2) \\ \vdots \\ Prox_{t\tilde{f}(\cdot)}(v_d) \end{bmatrix} . \tag{A.7}$$

Proof. We can write,

$$\begin{aligned} M_{f(\cdot)}(\mathbf{v}, t) = \min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) + \frac{1}{2t}||\mathbf{x} - \mathbf{v}||^2 &= \min_{\mathbf{x}\in\mathbb{R}^d} \sum_{i=1}^{d} \tilde{f}(x_i) + \frac{(x_i - v_i)^2}{2t} \,, \\ &= \sum_{i=1}^{d} \min_{x_i} \tilde{f}(x_i) + \frac{(x_i - v_i)^2}{2t} \,, \\ &= \sum_{i=1}^{d} M_{\tilde{f}(\cdot)}(v_i, t) \,. \end{aligned} \tag{A.8}$$

□

In the next lemma, we show that the Moreau envelope of a Lipschitz function is itself a Lipschitz function.

Lemma 37. *Let $\Phi : \mathbb{R}^d \to \mathbb{R}$ be an L-Lipschitz function. Then, $M_{\Phi(\cdot)}(\cdot, t)$ is a $2L$-Lipschitz function, i.e., for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$,*

$$|M_{\Phi(\cdot)}(\mathbf{u}, t) - M_{\Phi(\cdot)}(\mathbf{v}, t)| \leq 2L \|\mathbf{u} - \mathbf{v}\| . \tag{A.9}$$

Proof. In order to show this result, we need to find an upper bound on the derivative of the Moreau envelope. For all $\mathbf{v} \in \mathbb{R}^d$, we have,

$$\begin{aligned} L \left\|\mathbf{v} - \text{Prox}_{t\Phi(\cdot)}(\mathbf{v})\right\| &\geq \Phi\big(\mathbf{v}\big) - \Phi\big(\text{Prox}_{t\Phi(\cdot)}(\mathbf{v})\big) \\ &\geq \frac{1}{2t} \left\|\mathbf{v} - \text{Prox}_{t\Phi(\cdot)}(\mathbf{v})\right\|^2 , \end{aligned} \tag{A.10}$$

where the first inequality is due to the L-Lipschitzness of the function $\Phi(\cdot)$, and the second inequality is derived from the fact that $\text{Prox}_{t\Phi(\cdot)}(\mathbf{v})$ is the solution to the optimization (A.4). This gives the following bound on the distance of the proximal operator to the underlying vector.

$$\left\|\mathbf{v} - \text{Prox}_{t\Phi(\cdot)}(\mathbf{v})\right\| \leq 2tL. \tag{A.11}$$

We can now bound the derivative $\frac{\partial M_{\Phi(\cdot)}}{\mathbf{v}}$ as follows,

$$\left\|\frac{\partial M_{\Phi(\cdot)}}{\partial \mathbf{v}}\right\| = \frac{1}{t} \left\|\big(\mathbf{v} - \text{Prox}_{t\Phi(\cdot)}(\mathbf{v})\big)\right\| \leq 2L \ , \ \forall \mathbf{v} \in \mathbb{R}^d. \tag{A.12}$$

This concludes the proof. □