Universality Laws and Performance Analysis of the Generalized Linear Models

Thesis by
Ehsan Abbasi

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

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2020
Defended May 18, 2020

Ehsan Abbasi
ORCID: 0000-0002-0185-7933

ACKNOWLEDGEMENTS

It feels good to have an end to this chapter of my life, but it is the journey that matters in the end. Fortunately, this journey has been incredible for me since its beginning. Of course as life happens, there has been tons of ups and downs, there has been moments on the mountaintops and moments in deeps valleys of despair. But as I look back to my years at Caltech, all I remember is the incredible moments I had among colleagues and friends, whom I consider my second family now.

First of all, I would like to express my deepest gratitude to my advisor, Prof. Babak Hassibi. He is the main reason behind my spectacular stay at Caltech, by creating a trusting, friendly, and unstressful workplace. In fact, he was the best advisor, mentor, teacher, researcher and role model I could ever think of. His unique perspective taught me how to also see the big picture in every problem, how to think creatively and ask fundamental questions, how to present my ideas and results, how to work with others and more importantly, how to do independent research. His words have always been encouraging, his trust made me confident in my abilities and his patience and kindness has always been inspiring. After years of working with him, his brightness, even when encountering a problem for the first time, never seizes to impress me. Academic aspects aside, his unique ways of operating his group promotes a heartwarming sense of trust, togetherness, and friendship among the members. Not only is he the person I look up to in my academic career, he is also the person who has helped me a great deal on a personal level, with invaluable advice, discussions, and debates on a wide variety of topics such as ethics, culture, books, movies, and philosophies. He is in every aspect a great teacher and advisor, and I feel really grateful and fortunate to be one of his students.

I would also like to thank the members of my defense and candidacy committees, Professor Joel Tropp, Professor P. P. Vaidyanathan, Professor Venkat Chandrasekaran, Professor Thomas Vidik, and Professor Shuki Bruck, for their time, interest, and useful inputs. Their knowledge and expertise have been instrumental to my study at Caltech.

One of the greatest things at Caltech was the opportunity to work with some of the smartest people in the world. My deepest appreciation goes to my collaborators, Christos Thrampoulidis and Fariborz Salehi, for their great help as well as their friendship. I truly enjoyed working with my former and current labmates who made my experience at Caltech a sheer pleasure. Thank you Wael, Kishore, Ramya, Matt,

Wei, Anatoly, Hikmet, Sahin, Ahmed, Navid, Taylan, Oren, and Gautam, for maintaining a friendly environment in Moore 155. I would like to specifically thank my first collaborator, my mentor and my friend, Christos Thrampoulidis. I was fortunate and lucky to work under his supervision during my first years at Caltech, that taught me a great deal of my knowledge as well as my academic perspective. Thank you Christos, for being so patient with me, for the great help and true friendship, as well as the invaluable lessons I learned from you through countless technical discussions we had. No matter how far you are, these thoughts will always rekindle a great deal of amazing memories we shared in old town Pasadena, our trips to conferences, our workouts in Caltech's gym and all the stories we shared with each other about our dramas! I will be forever in your debt. I am also greatly thankful to Fariborz, with whom I shared many sleepless deadline nights. Thank you, Fariborz, for keeping me motivated to write down my proofs, for our amazing trips to Illinois, Michigan, our recurrent drives to San Diego, for our amazing academic and philosophical discussions, for being understanding and patient with me in my ups and downs, and for being a great friend as well as a smart perfectionist collaborator.

During my years at Caltech, I was not able to visit home, and unfortunately, due to the travel ban, my family could not visit me here either. So my journey wouldn't have been possible without the countless help and company of many friends I have in Caltech and Los Angeles. One of life's greatest joys is to meet someone who understands you in every way. In that aspect, I was lucky to meet my wife Hanieh during my PhD. Thank you, Hanieh, for being so patient with me during my long working night, for being so supportive of my works, for being my partner in crime and best friend and for being down for every crazy idea I come up with. Especially, the last few month in quarantine would have been intolerable without you being by my side. Thanks to you, I cannot image my life being any better. Among my friends whom helped me the most, I would like to especially thank my closest friend Pooya Vahidi, who became the brother I never had. Thank you Pooya for making my experience at Caltech so amazing, for the amazing surprise trips we planed (or at least wanted to plan), parties we attended together, dramas we went through together, and for making every day of my stay at Caltech full of unforgettable memories. I was greatly blessed to have many other amazing friends in southern California. I am going to miss Parham and our intellectual conversations, Aryan and our shared interest in tech gadgets, Peyman and his funny jokes and pranks, Pourya and his obsession with black holes, Omid and our etical conversations, Shaghayegh's company in every plan we made, Homa's funny moments, Ehsan Emamjome and our

hangouts in LA, Nazanin and Cameron, a couple I have always enjoyed hanging out with, Roohy and our endless nights, Hossein and his amazing plans, and all others, for their friendship and support during my graduate studies.

Last but not least, I would like to thank my parents, Abolfazl Abbasi and Parvin Naghash, as well as my little sister Mahsan Abbasi, for their unconditional love and unceasing support throughout my entire life. One of my greatest lucks was being raised in such an understanding family. Their continuous encouragement and prayer is what has motivated me in my life, and I am forever indebted to them for their sacrifices. You are the best family I could ever dream of. I'm especially grateful that my lovely sister has joined me here in US. Because of her, I feel more complete, as a big part of me is now closer to me. In the end, I would like to take this remarkable moment to say to my my parents: "I love you, Dad and Mom. Thank you for everything you have done for me!

To my beloved parents,
Parvin, and Abolfazl.

ABSTRACT

In the past couple of decades, non-smooth convex optimization has emerged as a powerful tool for the recovery of structured signals (sparse, low rank, etc.) from noisy linear or non-linear measurements in a variety of applications in genomics, signal processing, wireless communications, machine learning, etc.. Taking advantage of the particular structure of the unknown signal of interest is critical since in most of these applications, the dimension p of the signal to be estimated is comparable, or even larger than the number of observations n. With the advent of Compressive Sensing there has been a very large number of theoretical results that study the estimation performance of non-smooth convex optimization in such a *high-dimensional setting*.

A popular approach for estimating an unknown signal $\beta_0 \in \mathbb{R}^p$ in a *generalized linear model*, with observations $\mathbf{y} = g(\mathbf{X}\beta_0) \in \mathbb{R}^n$, is via solving the estimator $\hat{\beta} = \arg\min_\beta \ \mathcal{L}(\mathbf{y}, \mathbf{X}\beta) + \lambda f(\beta)$. Here, $\mathcal{L}(\cdot,\cdot)$ is a loss function which is convex with respect to its second argument, and $f(\cdot)$ is a regularizer that enforces the structure of the unknown β_0. We first analyze the generalization error performance of this estimator, for the case where the entries of $\mathbf{X}$ are drawn *independently from real standard Gaussian* distribution. The *precise* nature of our analysis permits an accurate performance comparison between different instances of these estimators, and allows to optimally tune the hyperparameters based on the model parameters. We apply our result to some of the most popular cases of generalized linear models, such as M-estimators in linear regression, logistic regression and generalized margin maximizers in binary classification problems, and Poisson regression in count data models. The key ingredient of our proof is the *Convex Gaussian Min-max Theorem (CGMT)*, which is a tight version of the Gaussian comparison inequality proved by Gordon in 1988. Unfortunately, having real iid entries in the features matrix $\mathbf{X}$ is crucial in this theorem, and it cannot be naturally extended to other cases.

But for some special cases, we prove some universality properties and indirectly extend these results to more general designs of the features matrix $\mathbf{X}$, where the entries are not necessarily real, independent, or identically distributed. This extension, enables us to analyze problems that CGMT was incapable of, such as models with quadratic measurements, phase-lift in phase retrieval, and data recovery in massive MIMO, and help us settle a few long standing open problems in these areas.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Number *Page*

LIST OF TABLES

Number *Page*

Chapter 1

INTRODUCTION

Data in today's technology and industry work is indispensable. Most organizations now understand that if they gather all the data that is available to them, they can analyze and get significant value from it. As a result, the last decade has seen a sustained exponential growth rate in data stored and used. This has led to an immense interest in the buzz words such as "Big Data" and "Statistical Inference", where the question is how to efficiently deduce the most information about the unknown variables of interest. Classical estimation theory has extensively investigated this question under various models, when the number of unknown variables is small compared to collected data. But in many modern applications (e.g. financial data, machine learning, wireless communications, sensor networks, genome signal processing, image processing, DNA sequencing, etc.), the number of unknown variable of interest has become larger and larger. Therefore, a lot of classical tools in estimation theory fail to address the same questions, with the dimensionality explosion that we experience in today's applications.

More importantly, in many of these applications, the number of unknown variables is even larger than the number of observations (or measurements) we have (e.g. consider the DNA sequencing where the unknown data is the human genome, or in image processing where the unknown is a large scale image, or in finance). Consider the following simple but fundamental example to make our idea concrete. We would like to recover an unknown vector $\beta \in \mathbb{R}^p$ from the system of linear equations below, with n equations,

$$\mathbf{y} = \mathbf{X}\beta \in \mathbb{R}^n \,, \tag{1.1}$$

where $\mathbf{y} \in \mathbb{R}^n$ and $\mathbf{X} \in \mathbb{R}^{n\times p}$ are given. Obviously, when we do not have any other information about the unknown β, it is necessary to have more measurements than unknowns for a consistent recovery (n should be greater than or equal to p). But in many applications, the unknown β is constrained by some structure (e.g. sparsity, where only a limited unknown number entries of β are non-zero, or the case where the entries of β are chosen from a finite alphabet like ± 1.) In these examples, although the unknown data is p-dimensional with $p \geq n$, it lies on a lower dimensional manifold with a lower degree of freedom, which may make recovery of the

unknown data feasible. These examples, give rise to problems like, *how to exploit the given structure to efficiently recover the unknown?*, or *under what conditions are our estimators consistent?* Among different estimation methods, convex estimators are popular as they exhibit numerical stability, and more flexibility. Besides, they are more tractable when it comes to computational analysis.

In the past couple of decades, non-smooth convex optimization has emerged as a powerful tool for the recovery of structured signals (sparse, low rank, etc.) from generalized linear measurements in a variety of applications in genomics, signal processing, wireless communications, machine learning, etc.. How to take advantage of the particular structure of the unknown signal of interest is critical since as explained, in most of these applications, the dimension p of the signal to be estimated is comparable, or even larger than the number of observations n ([40] and references therein). With the advent of Compressive Sensing there has been a very large number of theoretical results that study the estimation performance of non-smooth convex optimization in such a *high-dimensional setting*.

1.1 Generalized Linear Models

Linear models describe a continuous output variable as a function of the predictors, and are widely used in statistical data analysis. But in practice, the underlying models can be much more complicated than a simple linear model. In this thesis, we focus on a special class of models, known as the generalized linear models. In this section, we define the set up for these models, mention some of their applications and state of the art, and finally explain the thesis organization.

Mathematical Formulation

Consider the problem of recovering a p-dimensional signal $\beta_0 \in \mathbb{R}^p$, from n measurements of the form

$$y_i = g_i(\mathbf{x}_i^\mathsf{T}\beta_0), \qquad i = 1, \dots, n \ . \tag{1.2}$$

Here $g_i(.)$ is a known or unknown *link function*, and may include a random component such as noise. For instance, in the case of linear measurements we may have,

$$y_i = \mathbf{x}_i^\mathsf{T}\beta_0 + z_i \ , \tag{1.3}$$

where z_i's are the unknown noise entries. Our goal is to recover the vector β_0, given the measurements y_i's, the feature vectors $\mathbf{x}_i$'s and depending on the problem, some

information about the link function $g_i(.)$ or the structure of the unknown vector β_0. Henceforth, let

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^n \,, \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_m^\top \end{bmatrix} \in \mathbb{R}^{n \times p} \,, \quad g(\mathbf{x}) = \begin{bmatrix} g_1(x_1) \\ \vdots \\ g_m(x_m) \end{bmatrix} \in \mathbb{R}^n \,. \tag{1.4}$$

Here, $\mathbf{y}$ denotes the vector of n measurements, $\mathbf{X}$ is the features matrix, and $g : \mathbb{R}^n \to \mathbb{R}^n$ is the link function for which

$$\mathbf{y} = g(\mathbf{X}\beta_0) \tag{1.5}$$

Our model so far is very general as we did not impose any assumptions on any of the parameters. Thus, to answer "how to recover the unknown vector?", we have to make the model more clear, define the tools we would like to work with, and our performance measures.

1.2 The Link Function in the Generalized Linear Models

The link function $g(\cdot)$, plays a major role in how we approach this problem. We will explain a few important and popular cases of the link function.

Linear Inverse Problems in Compressed Sensing

The idea in compressed sensing is to recover a signal with low-dimensional structures from high dimensional measurements. The structured signal can be a sparse vector [38], a low rank matrix [135], a vector chosen from a finite alphabet[166], etc. There has been tremendous research under the name of compressed sensing in the last decades [6, 40, 42, 61, 73, 129, 165, 179].

The classical setting of linear inverse problems considers recovering an unknown $\beta_0 \in \mathbf{R}^p$, given linear noisey measurements of the form

$$\mathbf{y} = \mathbf{X}\beta_0 + \mathbf{z} \in \mathbb{R}^n \,, \tag{1.6}$$

where the measurements $\mathbf{y} \in \mathbb{R}^n$ and the features matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ are given. Convex estimators are a popular method of recovering the unknown in these problems. Especially, for the case of structured β_0, there are principled ways to recover the unknown using a convex estimator, based on the idea of atomic decomposition [40] and also other methods in [10, 15], as well as non-convex methods such as [122].

Classification Problems

Logistic regression is the most commonly used statistical model for predicting dichotomous outcomes [85]. It has been extensively employed in many areas of engineering and applied sciences, such as in the medical [27, 181] and social sciences [96]. 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 quantitive response. However, in many situations the response variable is qualitative (or categorical) and linear regression is no longer appropriate [89]. This is mainly due to the fact that least-squares 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 [124].

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 this problem, we assume the the measurements are given by the following model,

$$\mathbf{y} = g(\mathbf{x}) = \text{Sign}(\rho(\mathbf{x}) - \epsilon) \ , \quad \rho(x) = \frac{e^x}{e^x + e^{-x}} \ , \quad \epsilon_i \sim \text{Bernouli}(P) \ .$$

In this case, each measurement, y_i, will be $+1$ or -1 with probability $\rho(x_i)$ or $1 - \rho(x_i)$, respectively.

Phase Retrieval Problem

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 physics such as astronomical imaging [69], X-ray crystallography [116], medical imaging [55], and optics [191]. 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.

The goal is to recover an unknown data β_0 from the magnitude only measurements of the form,

$$\mathbf{y} = |\mathbf{X}\beta_0| \in \mathbb{C}^n \ , \tag{1.7}$$

where $\mathbf{y}$ and $\mathbf{X}$ are given and $|\cdot|$ is the element-wise absolute value operator.

Poisson Regression in Count Data Models

Poisson regression assumes that the observations y_i take a Poisson distribution with mean $\mathbf{x}^\mathsf{T}\beta_0$,

$$y_i = \text{Pois}\left(\mathbf{x}_i^\mathsf{T}\beta_0\right) , \qquad i = 1, \dots, n . \tag{1.8}$$

Poisson distribution has applications in many areas such as telecommunications (number of arriving calls in a system), Biology (number of mutations on a DNA), Finance and insurance (number of losses or claims in a period of time), etc..

1.3 High Dimensional Regime

In the classical regime, the common modeling assumption was that the number of unknown variables, p, is fixed, while the number of measurements, n, grow large. This problem is well studied for the cases of linear regression, classification problems and some other cases. But with the rise of big data, the number of unknowns in modern applications of statistical inference could grow as large as the number of observations. This new assumption, requires a difference performance analysis for the previous problems. There has been a lot of effort in the last decade, to answer the same questions in the classical regime, under this new assumption, which we will mention in Section 2.1, for different applications.

Throughout this thesis, we are especially interested in the over-parameterized regime, where the number of measurements, n, is less than the number of unknowns, p. This problem is usually ill-posed unless some prior information about the structure of the unknown data in given. For instance, sometimes the true unknown data, lies on a low-dimensional manifold in its original p-dimensional space. One of the most famous structures is data sparsity. In this case, most of the entries of the unknown data is zero, but of course the indices of the zero entries are unknown. Other popular structures includes, signals that are block sparse [152, 160], signals with entries drawn from a finite alphabet [166, 174], low rank matrices [135], or sometime vectors or matrices that exhibit a few simultaneous structures [128]. Recently there has been a unifying framework that can extend the analysis techniques that were initially developed for the sparse signal recovery, to other structures [18, 56, 63, 165].

Formally, most of the results of this Thesis will be applied to a sequence of problem instance $\{\beta_0, \mathbf{X}, g(\cdot), \mathcal{L}(\cdot,\cdot), f(\cdot)\}_{p\in\mathbb{N}}$, indexed by p, such that the properties we mentioned and our assumptions hold for all members of this sequence. We will not write the subscript n for arguments to avoid overloading notations. Our results are asymptotic and hold as $n \to \infty$.

1.4 Convex Recovery Method and Performance Measure

Among various recovery methods, we focus on convex optimization based estimations. Convex methods are often preferred as they exhibit numerical stability, and more flexibility. Besides, they are more tractable when it comes to computational or performance analysis.

These convex methods estimate the unknown vector β_0, by solving the following convex optimization problem,

$$\hat{\beta} = \arg\min_{\beta} \mathcal{L}(\mathbf{X}\beta, \mathbf{y}) + \lambda f(\beta) \ . \tag{1.9}$$

Here, $\mathcal{L}(\mathbf{X}\beta, \mathbf{y})$ is a loss function that penalizes the residual, and is convex with respect to its first argument. The function $f(\cdot)$ (which we call the regularizer) is a convex function that enforces the structure of the unknown vector β_0, to the final estimation $\hat{\beta}$. The positive parameter λ, is a regularization parameter that balances the cost function and the regularizer. This form includes a lot of famous estimators including ℓ_1-regularized least squares (aka the LASSO), penalized least absolute devations estimator (aka LAD), ridge regression, maximum-likelihood estimators, Support vector machine or perceptron or logistic regression in classification problems, etc.

There are standard solvers for each one of these examples, all of which benefit from the convex nature of this estimator. In this thesis, our focus is on the recovery performance of such estimators, rather than algorithmic issues. So our main question will be, *how well the optimization* (1.9) *can estimate the unknown data* β_0*?*

Performance Measures If we want to measure the recovery performance of the estimator (1.9), we need to define our performance measure first. For now, we keep our performance measure general in the form of

$$\psi(\hat{\beta}, \beta_0) \ , \tag{1.10}$$

where $\psi(\cdot)$ is a function that is supposed to measure the deviance of $\hat{\beta}$ from β_0, in our desired way depending on our application. For instance, in the classification problems, the performance measure could be the generalization error, which is

$$\begin{aligned} \psi(\hat{\beta}, \beta_0) &= \text{Prob}\left(\text{Sign}\left(\rho(\mathbf{x}^{\mathsf{T}}\hat{\beta}) - \epsilon\right) \neq \text{Sign}\left(\rho(\mathbf{x}^{\mathsf{T}}\beta_0) - \epsilon\right)\right), \\ \rho(t) &= \frac{e^t}{e^t + e^{-t}}, \quad \epsilon \sim \text{Unif}(0,1), \quad \mathbf{x} \sim \mathcal{N}(0, \mathbb{I}) \ . \end{aligned} \tag{1.11}$$

We will impose some natural assumptions on the function $\psi(\cdot, \cdot)$ later for our analysis.

1.5 Design of the Features Matrix, X

Performance of the estimator (1.9) depends on all problem parameters, including properties of the features matrix $\mathbf{X}$. As the most basic example, consider the problem of recovering $\beta_0 \in \mathbb{R}^p$ from the system of linear equations

$$\mathbf{y} = \mathbf{X}\beta_0 \in \mathbb{R}^p \ . \tag{1.12}$$

We need the features matrix $\mathbf{X}$ to be full rank for a consistent recovery of β_0. So in a worst case scenario analysis of the recovery performance for this problem, we should make sure $\mathbf{X}$ is full rank. But in practice, if we assume a random ensemble for the entries of $\mathbf{X}$, the probability of $\mathbf{X}$ being singular may be very small. For example, in the case that the entries of $\mathbf{X}$ are independently drawn from a continuous probability distribution, the probability of $\mathbf{X}$ being singular world be zero.
Besides, it has been observed that randomly generated features matrices can yield good estimates in the high dimensional regime [26, 68, 72]. Besides, assuming a random ensemble for the features matrix enables us to analyze the average case performance, or high probability performance in high dimensional regime. In particular, matrices sampled from Gaussian distribution has been traditionally useful in the performance analysis of estimators in compressed sensing [38]. Recently, with the development of two frameworks known as AMP [56] and CGMT [165], researchers have been able to answer most of the previously unknown questions in compressed sensing and classification. For these frameworks, the it is essential to assume that the entries of the features matrix $\mathbf{X}$, are independently drawn from standard real Gaussian distribution. Although their analysis answered a lot of open questions in compressed sensing and classification problems, the Gaussian assumption is very restrictive when it comes to practical problems.

Universality As discussed, assuming a randomly drawn features matrix $\mathbf{X}$ with iid Gaussian entries, has several benefits. First, it enables us to utilize a wide set of tools in probability. And second, we will not have to consider the worst case scenarios in the design of the features matrix. But even more importantly, this can be a good start in the analysis of such complex problems. Once we have a clear understanding of how the estimator behaves with an iid Gaussian features matrix, we can move forward and investigate how general these results are and how can one extend them

to other cases.

As a matter of fact, some of the results that hold under this Gaussian assumption, enjoy a remarkable universality property in high dimensions, that extend the same result to a wide variety of other distributions. Donoho and Tanner in [58] and Bayati et al, [17] were first to observe and show some universality in phase transition of an special case of estimator (1.9). Later, Oymak and Tropp [130] showed this universality for a wider range of problems and distributions. But in most of these cases, having real independent entries in the features matrix was essential. Although these works are of great interest, the independence assumption on the entries of the measurement vectors can be restrictive. In certain applications in communications, phase retrieval, covariance estimation, the entries of the measurement matrix have correlations. In this thesis, we show a much stronger universality result which holds for a broader class of measurement distributions. Here is some of the applications in which at least one of the key assumptions on the features matrix (real and iid entries) does not hold.

Quadratic Measurements

Consider the case, where we wish to recover an unknown matrix $\mathbf{\Sigma}_0$, from measurements of the form,

$$y_i = \mathbf{a}_i^\mathsf{T} \mathbf{\Sigma}_0 \mathbf{a}_i + z_i \ , \quad i = 1, \dots, n \ , \tag{1.13}$$

where the features vectors $\mathbf{a}_i$' and measurements y_i's are given. In this example, the measurements are still linear with respect to the unknown matrix $\mathbf{\Sigma}_0$, but quadratic with respect to the features vectors $\mathbf{a}_i$'s. We define $\beta_0 = \vec{\mathbf{\Sigma}}_0$ and $\mathbf{x}_i = \vec{\mathbf{a}_i \mathbf{a}_i^\mathsf{T}}$, where $\vec{\cdot}$ is the vectorized version of a matrix. Then we have

$$y_i = \mathbf{x}_i^\mathsf{T} \beta_0 + z_i \ , \quad i = 1, \dots, n \ , \tag{1.14}$$

Obviously, even by imposing a generic iid distribution on the entries of $\mathbf{a}_i$'s, the entries of $\mathbf{x}_i$ will be highly dependent. Therefore, one cannot simply apply the classical result on this new problem.

This problem shows up in many applications such as Covariance sketching for data streams [44, 120], non-coherent energy measurements in communications [180], phase retrieval problem, [36, 86, 148, 190] etc.

Data Recovery in Massive MIMO

Here, the goal is to recover a p-dimensional vector $\beta_0 \in \mathbb{C}^p$ where the entries of β_0 are independently drawn from the discrete set $\mathbf{S} \subset \mathbb{C}$ with distribution $\beta_{0,i} \sim p_\beta$. The set $\mathbf{S}$ defines the modulation used for data transmission (e.g. QAM, PSK, etc.). For this purpose, we are given the noisy multiple-input multiple-output (MIMO) relation of the form

$$\mathbf{y} = \mathbf{X}\beta_0 + \mathbf{z} \in \mathbb{C}^n, \tag{1.15}$$

where $\mathbf{X} \in \mathbb{C}^{n\times p}$ is the known MIMO channel matrix with i.i.d. entries drawn from $\mathcal{N}_{\mathbb{C}}(0, \frac{1}{p})$ and $\mathbf{z} \in \mathbb{C}^n$ is the unknown noise vector with i.i.d. random complex Gaussian $\mathcal{N}_{\mathbb{C}}(0, \sigma^2)$ entries. The important question here would be *does the same performance analysis techniques in real case hold for the case of complex features matrix as well?*
Interestingly, we show that the same results and techniques are not necessarily applicable to the case of a complex features matrix, as we will have examples of both scenarios in future chapters.

1.6 Organization of the Thesis

In this thesis, we investigate various scenarios for the generalized linear model in (1.9). In Chapter 2, we impose an iid standard Gaussian distribution on the entries of the features matrix $\mathbf{X}$, and analyze the performance of the general estimator in (1.9). Then we will apply our analysis of some interesting examples of generalized linear models such as M-estimators (Linear Regression models), binary classification problem (such as logistic regression and generalized margin maximizers), and also in data recover in massive MIMO with a real channel. In Chapter 3, we apply our result to the square-root LASSO problem with a general performance function $\psi(\cdot)$.
Later in Chapter 4, we investigate the problem of covariance estimation with quadratic measurements of the form (1.13). We show some universality properties that enables us to use the same methods as in Chapter 2 to analyze the performance of convex estimators in such scenario. In Chapter 5, we propose a fast algorithm for covariance estimation in graphical models with theoretical guarantees.
In Chapter 6, we prove a universality result for the phase transition of the linear inverse problems with a wide range of distribution for the features matrix $\mathbf{X}$. As a result, we show that the phase transition in successful recovery of the unknown data, depends only on the first and second order statistics of the rows of the features ma-

trix $\mathbf{X}$. As an application, we show that the minimum number of random quadratic measurements (also known as rank-one projections) required to recover a low rank positive semi-definite matrix is $3nr$, where n is the dimension of the matrix and r is its rank. 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 $3n$.

Chapter 7, investigates the problem of data recovery in massive MIMO, and shows that under specific conditions on the constellation (among other conditions) and after some modifications of the original problem, we can come up with an equivalent real estimation problem that can be analyzed by the tools introduced in Chapter 2. We use these results in Chapter 8, and propose a two-step algorithm for a near-maximum likelihood data recovery in massive MIMO. As Chapter 9 derives a phase transition in perfect recovery in complex phase-max problem, it concludes that universality does not always holds from problems with complex features matrices to their corresponding real problems. Finally, we gather some of the proofs of previous sections in the Appendix.

Chapter 2

PRECISE PERFORMANCE ANALYSIS OF GENERALIZED LINEAR MODELS

In this chapter[1], we first summarize the problem setup and introduce our assumptions, and then state the main theorem. We would like to recover the p dimensional vector β from measurements of the form

$$\mathbf{y} = g(\mathbf{X}\beta_0) \in \mathbb{R}^n \ . \tag{2.1}$$

We are given the observation vector $\mathbf{y}$ and the features matrix $\mathbf{X}$ and sometimes some information about the *link function* $\mathbf{g}(\cdot)$. We do so by solving the following convex optimization,

$$\hat{\beta} = \arg\min_{\beta} \mathcal{L}(\mathbf{X}\beta, \mathbf{y}) + \lambda f(\beta) \ . \tag{2.2}$$

Here the loss function $\mathcal{L}(\cdot,\cdot)$ is convex with respect to its first argument, and the regularizer $f(\cdot)$ is also convex. In this section, our goal is to analyze performance of the optimization (2.2), in terms of

$$\psi(\hat{\beta}, \beta_0) \ , \tag{2.3}$$

where $\psi(\cdot,\cdot)$ is a function that measures the distance between β_0 and $\hat{\beta}$. As already noted in the introduction, this performance measure depends on all the problem parameters including design properties of the features matrix $\mathbf{X}$, distribution of the noise and the underlying vector β_0, properties of the loss function and the regularizer function, etc.

In this section, we focus on the case that the features matrix $\mathbf{X}$ has *iid standard Gaussian entries*. We will generalize the model in the next chapters.

The performance of this optimization, depends on the Loss function, the regularizer, distribution of the link function $\mathbf{g}(\cdot)$, and distribution (or properties) of the unknown data β_0 through the following Moreau envelope transformation. The Moreau envelopes of the loss function $\mathcal{L}(\cdot,\cdot)$ and the regularizer function $f(\cdot)$ are respectively defined as

$$\begin{aligned} \mathrm{e}_{\mathcal{L}}(\mathbf{x}, \mathbf{y}, \tau) &:= \min_{\mathbf{v}} \ \frac{1}{2\tau} \|\mathbf{v} - \mathbf{x}\|^2 + \mathcal{L}(\mathbf{v}, \mathbf{y}) \ , \\ \mathrm{e}_{f}(\mathbf{x}, \tau) &:= \min_{\mathbf{v}} \ \frac{1}{2\tau} \|\mathbf{v} - \mathbf{x}\|^2 + f(\mathbf{x}) \ . \end{aligned} \tag{2.4}$$

[1] Some of the materials of this chapter are based on the works in [143, 165, 166]

Similarly, the proximal operator is defined as the minimizers of the above optimizations,

$$\begin{aligned}\text{Prox}_{\mathcal{L}}\left(\mathbf{x},\mathbf{y},\tau\right) &:= \arg\min_{\mathbf{v}} \ \frac{1}{2\tau}\left\|\mathbf{v}-\mathbf{x}\right\|^2 + \mathcal{L}(\mathbf{v},\mathbf{y}) \ , \\ \text{Prox}_{f}\left(\mathbf{x},\tau\right) &:= \arg\min_{\mathbf{v}} \ \frac{1}{2\tau}\left\|\mathbf{v}-\mathbf{x}\right\|^2 + f(\mathbf{x}) \ . \end{aligned} \tag{2.5}$$

Note that since the loss function takes two input variable, its corresponding Moreau envelope takes three input, whearas the corresponding Moreau envelope for the regularizer takes only two inputs.
Assumption 1, introduces an essential functional, through which the performance of the convex optimization depends on the loss function, the regularizer, and distribution and properties of the link function and unknown data.

Assumption 1 (Functionals L and F) *We say that Assumption 1 holds for the functions $\mathcal{L}$ and f, and for the distributions $p_{\mathbf{g}(\cdot)}$ and p_{β_0}, if for all $c_1, c_2 \in \mathbb{R}$ and $\tau > 0$, there exist continuous functions $L : \mathbb{R}\times\mathbb{R}\times\mathbb{R}_{>0} \to \mathbb{R}$ and $F : \mathbb{R}\times\mathbb{R}\times\mathbb{R}_{>0} \to \mathbb{R}$ such that*

$$\begin{aligned}&\mathrm{e}_{\mathcal{L}}\left(c_1\,\mathbf{h}_1 + c_2\,\mathbf{X}\beta_0\,,\mathbf{g}(\mathbf{X}\beta_0)\,,\tau\right) \xrightarrow{P} L(c_1,c_2,\tau) \qquad \textit{and} \\ &\mathrm{e}_{f}\left(c_1\,\mathbf{h}_2 + c_2\,\beta_0\,,\tau\right) \xrightarrow{P} F(c_1,c_2,\tau) \ . \end{aligned} \tag{2.6}$$

where, the convergence is in probability over distribution of the function $\mathbf{g}(\cdot)$, distribution (or properties) of the unknown data β_0, the random matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ with iid standard Gaussian entries, and the random vectors $\mathbf{h}_1 \in \mathbb{R}^n$ and $\mathbf{h}_2 \in \mathbb{R}^p$ with iid standard Gaussian random entries.

Assumption 1 holds naturally for a wide range of norms, loss functions, and regularizers, due to the law of large numbers. In Section 2.1, we will derive the funcionals L and F for some interesting examples.
The second assumption we introduce is related to the performance function $\psi(\cdot,\cdot)$, which measures how well the optimization works.

Assumption 2 (Performance Function $\psi(\cdot,\cdot)$) *We say that Assumption 2 holds for the function $\psi(\cdot,\cdot)$ and for the distributions p_{β_0}, if for all $c_1, c_2 \in \mathbb{R}$ and $\tau > 0$, there exist continuous function $\Psi : \mathbb{R}\times\mathbb{R}\times\mathbb{R}_{>0} \to \mathbb{R}$ such that*

$$\psi\left(\text{Prox}_{f}\left(c_1\beta_0 + c_2\mathbf{h}_2\,,\tau\right),\beta_0\right) \xrightarrow{P} \Psi(c_1,c_2,\tau) \ . \tag{2.7}$$

where, the convergence is in probability over the distribution (or properties) of the unknown data β_0, and the random vector $\mathbf{h}_2 \in \mathbb{R}^p$ with iid standard Gaussian random entries.

Assumption 2 is valid for a wide range of function ψ. Simple instances include p-norms, for which the assumption holds by the law of large numbers when the entries β_0 are drawn independently from the same distribution.
Out main theorem, analyzes the convex estimator (2.2), in its most general form. The proof of Theorem 1 is deferred to Section 2.6, as well as explanation about the CGMT framework, which is the main tool that we use to prove this theorem.

Theorem 1 *Let $\hat{\beta}$ be the solution the convex estimator* (2.2)*, which is an estimation of the unknown data β_0, where the entries of the features matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ are drawn independently from standard Gaussian distribution. Also Assumptions 1 and 2 hold with functionals L, F and Ψ, $\delta = n/p$ and $\kappa = \|\beta_0\|/\sqrt{p}$. Consider the following min-max optimization over 6 scalars $(\alpha, \sigma, \tau_1, \tau_2, t, \gamma)$,*

$$\min_{\substack{\alpha\in\mathbb{R}\\ \sigma,\tau_1\geq 0}} \max_{\substack{t,\tau_2\geq 0\\ \gamma\in\mathbb{R}}} \frac{t}{2\tau_1} - \frac{\sigma}{2\tau_2} - \frac{\sigma\tau_2 t^2}{2\delta} + \frac{\sigma\tau_2\gamma^2\kappa^2}{2} + L(\sigma, \alpha, \frac{1}{t\tau_1}) + F(-\frac{\sigma\tau_2 t}{\sqrt{\delta}}, \alpha - \sigma\tau_2\gamma, \sigma\tau_2) \ . \tag{2.8}$$

If this min-max has a unique solution, $(\hat{\alpha}, \hat{\sigma}, \hat{\tau}_1, \hat{\tau}_2, \hat{t}, \hat{\gamma})$, then as p and n grow to infinity with $\delta = p/n$, we have

$$\lim_{p,n\to\infty} \psi(\hat{\beta}, \beta_0) = \Psi(\hat{\alpha} - \hat{\sigma}\hat{\tau}_2\hat{\gamma}, \frac{\hat{\sigma}\hat{\tau}_2\hat{t}}{\sqrt{\delta}}, \hat{\sigma}\hat{\tau}_2) \ . \tag{2.9}$$

Theorem 1 derives a precise analysis for the performance of the convex estimator (2.2). A few remarks are in place. As discussed earlier, the result of this theorem applies to a sequence of problem instances of the convex estimator, with growing dimensions n and p, such that $n/p = \delta$. Then, the convergence in theorem is in probability over the randomness of the features matrix $\mathbf{X}$, distribution of the link function $\mathbf{g}$ and unknown data β_0 (if there's any).
We now investigate two classes of popular models with this theorem. First, we consider the case when the link function $\mathbf{g}$, simply adds a random noise to its input, as in linear models. Second, we analyze the case of binary classification.

2.1 Linear Models and M-Estimators

We consider the standard problem of recovering an unknown signal $\beta_0 \in \mathbb{R}^p$ from a vector $\mathbf{y} \in \mathbb{R}^n$ of n noisy, linear observations given by $\mathbf{y} = \mathbf{X}\beta_0 + \mathbf{z} \in \mathbb{R}^n$.

Here, $\mathbf{X} \in \mathbb{R}^{n\times p}$ is the (known) measurement matrix, and, $\mathbf{z} \in \mathbb{R}^n$ is the noise vector; the latter is generated from some distribution density in $\mathbb{R}^n$, say $p_{\mathbf{z}}$. Our focus is on the *high-dimensional regime* where both the dimensions of the ambient space n and the number of measurements m are large [60, 146]. This is different than the classical one, where p is small and fixed and only n is assumed large. Of special interest is the scenario of *compressed* measurements, in which $p < n$. In principle, such inverse problems are ill-posed, unless the unknown vector is somehow structurally constrained to only have very few degrees of freedom relative to its ambient space. Such signals are called *structured* signals; popular examples of such structure include sparsity, block-sparsity, low-rankness, etc. [10, 40]. We model such structural information on β_0 by assuming that it is sampled from an n-dimensional probability density p_{β_0}.

Regularized M-estimators. The most widely used approach to obtain an estimate $\hat{\mathbf{x}}$ of the unknown β_0 from the vector $\mathbf{y}$ of observations is via solving the *convex* program

$$\hat{\beta} := \arg\min_{\beta} \ \mathcal{L}(\mathbf{y} - \mathbf{X}\beta) + \lambda f(\beta). \tag{2.10}$$

The *loss function* $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}$ measures the deviation of $\mathbf{X}\hat{\beta}$ from the observations $\mathbf{y}$, the *regularizer* $f : \mathbb{R}^p \to \mathbb{R}$ aims to promote the particular structure of β_0, and, the regularizer parameter $\lambda > 0$ balances between the two. Henceforth, both $\mathcal{L}$ and f are assumed to be convex. Also, f will typically be non-smooth. We refer to the minimization problems of the form in (2.10) as *regularized M-estimators*. Different choices of the loss function and of the regularizer give rise to a number of well-known estimators. A few concrete examples might suffice: (i) the *LASSO* [176] corresponds to (2.10) with $\mathcal{L}(\mathbf{v}) = \frac{1}{2}\|\mathbf{v}\|_2^2$ and $f(\mathbf{x}) = \|\mathbf{x}\|_1$. General choices of the regularizer for the same loss function lead to the Generalized LASSO [129, 133] (ii) The *regularized-LAD* [192] minimizes an ℓ_1-loss function. (iii) The (generalized) *square-root LASSO* [22] solves (2.10) for $\mathcal{L}(\mathbf{v}) = \|\mathbf{v}\|_2$. In the first two examples the loss function is *separable* over its entries, i.e. $\mathcal{L}(\mathbf{v}) = \sum_{j=1}^n \ell(\mathbf{v}_j)$ for convex $\ell : \mathbb{R} \to \mathbb{R}$; in contrast, the square-root LASSO does not belong to this category. Accordingly, the regularizer function might be separable (e.g. ℓ_1-norm) or not (e.g. nuclear-norm).

Prior Work

With the advent of Compressed Sensing there is a very large number of theoretical results that have appeared in recent years in place for various types of regularized

M-estimators. The vast majority of those results hold under standard incoherence or restricted eigenvalue conditions on the measurement matrix $\mathbf{X}$ [2], but they are *order-wise* in nature, i.e., they characterize the error performance only up to loose constants. While this line of work includes unifying frameworks for the analysis of general instances of (2.10), the loose constants involved in the error bounds do *not* permit any accurate comparisons among the different instances (e.g. [14, 106, 123, 189] and references therein); therefore, they cannot be used to answer optimality questions of the nature discussed in this Section.

This section derives precise characterizations of the error behavior (ones that do not involve unknown constants). Results of this nature have appeared in the literature under the additional assumption of an iid Gaussian distribution imposed on the entries of the matrix $\mathbf{X}$. The inspiration behind these studies can be traced back to the seminal work of Donoho [57, 59] on the phase-transition of ℓ_1-minimization in the Compressed Sensing problem. This and the extensive follow-up literature mostly focused on the *noiseless* signal recovery problem. More recently, researchers have initiated the study of the exact reconstruction error of instances of (2.10) in the presence of *noise*. Unfortunately, no unifying treatment that holds for general instances has hitherto been available. To the best of our knowledge, our work is the first to obtain *precise* characterizations of the error performance of (2.10) for *general* convex loss functions, convex regularizers, and noise and signal distributions under a Gaussian assumption on the random measurement matrix $\mathbf{X}$. In the rest of this section, we briefly outline the relevant literature.

The first precise results on the performance of non-smooth convex optimization methods appear in the literature in the context of *noiseless* linear inverse problems that arise in Compressed Sensing. Here, the vector of measurements of the unknown structured signal β_0 takes the form $\mathbf{y} = \mathbf{X}\beta_0 \in \mathbb{R}^n$ and recovery is attempted via solving $\min_{\mathbf{y}=\mathbf{X}\beta} f(\beta)$, for an appropriately chosen convex regularizer f. In the absence of noise, the standard measure of performance becomes that of the minimum number of measurements required for *exact recovery* of β_0. By now, there is an elegant and complete theory that precisely characterizes this number when $\mathbf{X}$ has entries iid Gaussian. The theory was built in a series of recent papers [6, 17, 40, 57, 129, 153, 156]. Our work extends the analysis to the *noisy* setting. In the presence of noise, the analysis is inherently more challenging since: (a)

[2] Such conditions have been shown to be satisfied by a wide class of randomly designed measurement matrices, (e.g. [50, 68, 72] and references therein). A more recent line of works obtains similar order-wise bounds under even weaker assumptions on the randomness properties of $\mathbf{X}$ [101, 150, 178].

one needs to characterize the precise value of the estimation error, rather than just discriminating between exact recovery or not; (b) the performance depends not only on the number of measurements but also on the noise and signal statistics. Also, it naturally includes the results of the noiseless case as special instances. However, many of the ideas, analytical tools and concepts developed in the works [6, 40, 153, 156] have proved to be useful in extending the results to the noisy setting.

In the noisy setting, the first precise results analyzed the error performance of regularized least-squared (a.k.a. generalized-LASSO) under an iid gaussianity assumption on the noise distribution [18, 63, 129, 154, 168, 172, 173]. It has been only very recently, that El Karoui [65, 95], and, Donoho and Montanari [56, 64] were able to rigorously[3] predict the error performance of M-estimators under more general assumptions on the loss function and on the noise distribution. However, the papers by Donoho and Montanari assume *no* regularization and El Karoui considers the special case of ridge regularization. Finally, the very recent paper [29] builds upon [56] and extends the study to the case of ℓ_1-regularization. In short, our work achieves by several means a more complete and transparent treatment of the subject, overcoming the limitations of previous endeavors as follows: (i) We consider arbitrary convex regularizers, (ii) We identify minimal and generic assumptions under which the general result holds, (iii) We remove any smoothness and strong-convexity assumptions on the loss function, which are required in all previous works. Also, the loss function (and regularizer) need not be separable (e.g., we allow $\mathcal{L}(\mathbf{v}) = \|\mathbf{v}\|_2$ or $\|\mathbf{v}\|_\infty$), and, the distributions need not be iid. (iv) We remove boundedness assumptions on the moments of the noise distribution. Notably, our proof technique is fundamentally different than that of [65] and [56], and, it appears to be more direct and insightful in several ways.

Applying Theorem 1 to M-Estimators

In order to apply Theorem 1, we first need to translate Assumption 1 and 2 to this special case. Note that the Moreau envelope for this subtractive case of loss function in (2.10) becomes

$$\begin{aligned} \mathrm{e}_{\mathcal{L}}\left(\mathbf{x},\mathbf{y},\tau\right) &= \min_{\mathbf{v}} \ \frac{1}{2\tau}\left\|\mathbf{v}-\mathbf{x}\right\|^2 + \mathcal{L}(\mathbf{v},\mathbf{y}) = \min_{\mathbf{v}} \ \frac{1}{2\tau}\left\|\mathbf{v}-\mathbf{x}\right\|^2 + \bar{\mathcal{L}}(\mathbf{v}-\mathbf{y}) \\ &= \min_{\mathbf{v}} \ \frac{1}{2\tau}\left\|\mathbf{v}-(\mathbf{x}-\mathbf{y})\right\|^2 + \bar{\mathcal{L}}(\mathbf{v}) \end{aligned} \tag{2.11}$$

[3] The study of high-dimensional M-estimators has been previously considered in [19, 66]. However, those results are only based on heuristic arguments and simulations.

Recall from Assumption 1, that the link function and the loss function affect the estimation performance through the link function L that depends on the Moreau envelope as follows,

$$\begin{aligned} L(c_1, c_2, \tau) &= \lim_{n,p\to\infty} \mathrm{e}_{\mathcal{L}} \left(c_1\, \mathbf{h}_1 + c_2\, \mathbf{X}\beta_0\, , \mathbf{g}(\mathbf{X}\beta_0)\, , \tau\right) \\ &= \lim_{n,p\to\infty} \min_{\mathbf{v}} \frac{1}{2\tau}\, \|\mathbf{v} - (c_1\, \mathbf{h}_1 + (c_2 - 1)\, \mathbf{X}\beta_0 - \mathbf{z})\|^2 + \bar{\mathcal{L}}(\mathbf{v}) \quad (2.12) \end{aligned}$$

Note that $\mathbf{X}$ is a random matrix with iid standard Gaussian entries. Thus, $\mathbf{X}\beta_0$ can be replaces with $\tilde{\mathbf{h}}\, \|\beta_0\|$, where $\tilde{\mathbf{h}}$ is a random vector with iid standard Gaussian entries. Finally, $c_1\, \mathbf{h}_1 + (c_2 - 1)\mathbf{X}\beta_0$ can be replaces by $\sqrt{c_1^2 + (c_2 - 1)^2\|\beta_0\|^2}\mathbf{h}$, where $\mathbf{h}$ is a random standard Gaussian vector. Therefore, the functional $L(c_1, c_2, \tau)$ is simply a function of τ and $\sqrt{c_1^2 + (c_2 - 1)^2\|\beta_0\|^2}$. If we use this new function in the equations of Theorem 1, we can derive the result of [165], after a few changes of variables.

We specialize the general result of Theorem 1 to the popular case where the loss function $\mathcal{L}$ and the regularizer f are both separable, and the noise vector and signal β_0 both have entries iid. To make things concrete, assume[4]

$$\mathcal{L}(\mathbf{v}) = \sum_{j=1}^{n} \ell(\mathbf{v}_j) \qquad \text{and} \qquad z_j \overset{\text{iid}}{\sim} p_Z,\ j = 1, \ldots, n.$$

$$f(\mathbf{x}) = \sum_{i=1}^{p} f(x_i) \qquad \text{and} \qquad \beta_{0i} \overset{\text{iid}}{\sim} p_x,\ i = 1, \ldots, n.$$

Henceforth, both ℓ and f are proper closed convex functions. Also, it is further assumed that

$$\ell(0) = 0 = \min_{v} \ell(v) \quad \text{and} \quad f(0) = 0. \qquad (2.13)$$

Satisfying Assumption 1

To apply Theorem 1, we first need to verify that Assumption 1 holds for both the loss function and the noise distribution, and, for the regularizer and the signal distribution.

[4]Note the slight abuse of notation here in using f to denote both the vector-valued and scalar regularizer function.

Loss function and noise distribution

In the separable case Assumption 1 essentially translate to the following requirement on ℓ and p_Z:

$$\mathbb{E}\left[|\ell'_+(cG+Z)|^2\right] < \infty, \quad \text{for all } c \in \mathbb{R}. \tag{2.14}$$

where the expectation is over $Z \sim p_Z$ and $G \sim \mathcal{N}(0,1)$. This is shown in Lemma 1 below.

Lemma 1 (Expected Moreau envelope–Loss fcn) *If ℓ and p_Z satisfy* (2.14)*, then, Assumption 1 hold with*

$$L(c,\tau) = \mathbb{E}\left[\mathrm{e}_\ell\left(cG+Z,\mathbf{y},\tau\right) - \ell(Z)\right]. \tag{2.15}$$

This lemmas is simply a result of the law of large numbers.

Regularizer and Signal Distribution

Not surprisingly, the required condition on f and p_x becomes

$$\mathbb{E}\left[|f'_+(cH+\beta_0)|^2\right] < \infty, \quad \text{for all } c \in \mathbb{R}. \tag{2.16}$$

where the expectation is over $\beta_0 \sim p_x$ and $H \sim \mathcal{N}(0,1)$. Additionally, the following mild assumptions are required:

$$\exists\, \beta_+ > 0,\ \beta_- < 0 \ \text{ such that } \ 0 \le f(x_\pm) < \infty \qquad \text{and} \qquad \mathbb{E}\beta_0^2 < \infty. \tag{2.17}$$

Lemma 2 (Expected Moreau Envelope–Regularizer fcn) *If f and p_x satisfy* (2.16) *and* (2.17)*, then, Assumption 1 hold with*

$$F(c,\tau) = \mathbb{E}\left[\mathrm{e}_f\left(cH+\beta_0,\tau\right) - f(X_0)\right]. \tag{2.18}$$

The Expected Moreau Envelope

If conditions (2.14), (2.16) and (2.17) are satisfied, then Theorem 1 is applicable with L and F given as in (2.15) and (2.18), respectively. We call those functions, the *Expected Moreau Envelopes*. It is apparent from Theorem 1 that they play a key role in determining the error performance of the corresponding M-estimators. Moreover, they possess two key features, namely, *smoothness* and *strict convexity*; we elaborate on these here.

Lemma 3 (Smoothness) *Suppose ℓ is a closed proper convex function and p_Z a noise density such that* (2.14) *holds. Then, the function $L(c,\tau) := \mathbb{E}\left[\mathrm{e}_\ell\left(cG+Z,\tau\right)-\ell(Z)\right]$ is differentiable in $\mathbb{R}\times\mathbb{R}_{>0}$ with*

$$\frac{\partial L}{\partial c} = \mathbb{E}\left[\mathrm{e}'_\ell\left(cG+Z;\tau\right)G\right] \quad \text{and} \quad \frac{\partial L}{\partial \tau} = -\frac{1}{2}\mathbb{E}\left[\left(\mathrm{e}'_\ell\left(cG+Z;\tau\right)\right)^2\right].$$

Note that L is smooth, regardless of any non-smoothness of ℓ. This is a well-known fact about Moreau envelope approximations, and also, one of the primal reasons behind the important role those functions play in convex analysis [136]. The property is naturally inherited to the *Expected* Moreau envelopes as revealed by the lemma above.

Lemma 4 (Strict Convexity) *Suppose ℓ is a closed proper convex function and p_Z a noise density such that* (2.14) *holds and the following are satisfied:*

1. *Either there exists $x \in \mathbb{R}$ at which ℓ is* not *differentiable, or, there exists interval $\mathcal{I} \subset \mathbb{R}$ where ℓ is differentiable with a strictly increasing derivative,*
2. $\mathrm{Var}(Z) \neq 0$ [5]*, and, at each $z \in \mathbb{R}$, $p_Z(z)$ is either a Dirac delta function or it is continuous.*

Then, $L(c,\tau) := \mathbb{E}\left[\mathrm{e}_\ell\left(cG+Z,\tau\right)-\ell(Z)\right]$ is jointly strictly convex in $\mathbb{R}_{>0}\times\mathbb{R}_{>0}$.

Remark 1 *The function L is strictly convex, without requiring any strong or strict convexity assumption on ℓ. Interestingly, this property is not in general true for Moreau envelope approximations, but, it turns out to be the case for the* Expected *Moreau envelope L. The fact that the latter further involves taking an expectation over $cG+Z$, with G having a nonzero density on the entire real line, turns out to be critical.*

We are now ready to state our main result of this section which characterizes the squared error of separable M-estimators. This is essentially a corollary of Theorem 1.

[5] We require that there exist at least two values of $z \in \mathbb{R}$ for which $p_Z(z) > 0$. In particular, there is *no* requirement that $\mathrm{Var}(Z)$ be defined, e.g. Cauchy distribution is allowed.

Theorem 2 *Suppose ℓ and p_Z satisfy* (2.14)*, and, the two conditions of Lemma 4. Further assume that f, p_x satisfy* (2.16) *and* (2.17)*. Let $\hat{\beta}$ be any minimizer of the separable M-estimator, and consider the problem in* (2.8) *with L and F given as in* (2.15) *and* (2.18)*, respectively. If the solution to the* (2.8) *is unique and bounded, then it holds in probability that*

$$\lim_{n\to\infty} \frac{1}{p}\|\hat{\beta} - \beta_0\|_2^2 = \alpha_\star^2 \,.$$

where $\alpha_\star$ is the solution to the system of equations in (2.19)*, in 4 unknowns $\alpha, \gamma, \nu, \kappa$.*

As a system of nonlinear equations

Theorem 2 predicts the error of the M-estimator as the optimizer $\alpha_\star$ to a convex-concave optimization problem with four optimization variables. Equivalently, $\alpha_\star$ can be expressed via the first-order optimality conditions (stationary equations) corresponding to this optimization. Recall from Lemma 3 that L and F are differentiable (irrespective of smoothness of ℓ and f). The solution to the (2.8) can be derived by taking derivative of the objective function of (2.8). This results in the following system of non-linear equations.

$$\begin{cases} \alpha^2 = \mathbb{E}\left[\left(\frac{\lambda}{\nu}\cdot \mathrm{e}'_f\left(\frac{\gamma}{\nu}H + X_0; \frac{\lambda}{\nu}\right) - \frac{\gamma}{\nu}H\right)^2\right], \\ \gamma^2 = \delta\cdot\mathbb{E}\left[\left(\mathrm{e}'_\ell(\alpha G + Z, \kappa)\right)^2\right], \\ \nu\alpha = \delta\cdot\mathbb{E}\left[\mathrm{e}'_\ell(\alpha G + Z, \kappa)\cdot G\right], \\ \kappa\gamma = \frac{\gamma}{\nu} - \frac{\lambda}{\nu}\cdot\mathbb{E}\left[\mathrm{e}'_f\left(\frac{\gamma}{\nu}H + X_0; \frac{\lambda}{\nu}\right)\cdot H\right]. \end{cases} \tag{2.19}$$

Here, e'_f and e'_ℓ, denote the first derivatives of the Moureau envelopes with respect to their first argument.

Remark 2 *The system of equations in* (2.19) *can be easily reformulated in terms of the proximal operator of f and ℓ, using*

$$\mathrm{e}'_\ell(\chi, \tau) = \frac{1}{\tau}(\chi - \mathrm{prox}_\ell(\chi; \tau)),$$

and similar for f (see Lemma 40(iii)). In the case of additional smoothness assumptions on the loss function and/or the regularizer, further reformulations are possible. For example, if ℓ is two times differentiable, then using Stein's formula for Normal random variables we can make the following substitution in (2.19)*:*

$$\mathbb{E}\left[\mathrm{e}'_\ell(\alpha G + Z, \kappa)\cdot G\right] = \alpha\cdot\mathbb{E}\left[\mathrm{e}''_\ell(\alpha G + Z, \kappa)\right], \tag{2.20}$$

where the double-prime superscript denotes the second derivative with respect to the first argument. Such reformulations, are often convenient for analysis purposes; see for example Remark 6.

Remark 3 *The system of equations in* (2.19) *comprises of four nonlinear equations in four unknowns. Setting* $\mathbf{t} = (\alpha, \beta, \nu, \kappa)$ *for the vector of unknowns, the system of equations in* (2.19) *can be written as* $\mathbf{t} = S(\mathbf{t})$*, for appropriately defined* $S : \mathbb{R}^4 \to \mathbb{R}^4$*. We have empirically observed that a simple recursion* $\mathbf{t}_{k+1} = S(\mathbf{t}_k), k = 0, 1, \ldots$ *converges to a solution* $\mathbf{t}_*$ *satisfying* $\mathbf{t}_* = S(\mathbf{t}_*)$*. This observation is particularly useful since it allows for efficient numerical experimentations, cf. Section 2.1. It is certainly an interesting and practically useful subject of future work to identify analytic conditions under which such simple recursive schemes provide efficient means of solving* (2.19)*.*

Remark 4 *The results of this section extend naturally, and without any extra effort, to the case of "block-seperable" loss functions and/or regularizers. A popular example that falls in this category is* $\ell_{1,2}$*-regularization, which is typically used for the recovery of block-sparse signals. In such a case* $f(\mathbf{x}) = \sum_{i=1}^{b} \|[\mathbf{x}]_i\|_2$*, where* $[\mathbf{x}]_i = [\mathbf{x}_{(i-1)t+1}, \mathbf{x}_{(i-1)t+2}, \ldots, \mathbf{x}_{(i-1)t+t}]$*,* $i = 1, \ldots, b$ *is the* i^{th} *block of* $\mathbf{x}$*. Here,* b *is the number of blocks and* t *is the length of each block. In the proportional high-dimensional regime, one would assume* b *growing linearly with* p *with a constant ratio of* $1/t$*.*

Next, we explore some popular examples, where we can apply Theorem 2.

No Regularization

Consider an M-estimator without regularization, i.e.,

$$\hat{\beta} := \arg\min_{\beta} \sum_{j=1}^{n} \ell(\mathbf{y}_j - \mathbf{x}_j^T \beta_j). \tag{2.21}$$

For simplicity, we consider $\mathbf{z}_j \overset{\text{iid}}{\sim} p_Z$ and a separable loss function. Assuming that ℓ and p_Z satisfy the assumptions of Theorem 2, and, noting that $f = 0 \implies F(c, \tau) = 0$, the squared error of (2.21) is predicted by the minimizer α_* of the following (SPO) problem

$$\inf_{\substack{\alpha \geq 0 \\ \tau_g > 0}} \sup_{\gamma \geq 0} \frac{\gamma \tau_g}{2} + \delta L(\alpha, \frac{\tau_g}{\gamma}) - \alpha\gamma, \tag{2.22}$$

where we have performed the (straightforward) optimization over τ_h: $\inf_{\tau_h>0} \frac{\tau_h}{2} + \frac{\gamma^2}{2\tau_h} = \gamma$. We may equivalently express α_* as the solution to the first-order optimality conditions of (2.22). In particular, the stationary equations (see (2.19)) simplify in this case to the following system of two equations in two unknowns:

$$\begin{cases} \alpha^2 = \delta\kappa^2 \mathbb{E}\left[\left(\mathrm{e}'_\ell(\alpha G + Z, \kappa)\right)^2\right], \\ \alpha = \delta\kappa \cdot \mathbb{E}\left[\mathrm{e}'_\ell(\alpha G + Z, \kappa) \cdot G\right]. \end{cases} \tag{2.23}$$

Starting from (2.23), some interesting conclusions can be drawn regarding the performance of M-estimators without regularization, which we gather in the following remarks.

Remark 5 *It follows from* (2.23) *that in the absence of regularization, it is required that the number of measurements n is at least as large as the dimension of the ambient space p ($\delta \geq 1$), in order for the recovery to be* stable*, i.e. the error be finite. To see this, assume stable recovery, then there exists (α_*, κ_*) satisfying* (2.23)*. Starting from the second equation, applying the Cauchy-Schwarz inequality and substituting back the first equation we find:*

$$\alpha_* = \delta\kappa_* \cdot \mathbb{E}\left[\mathrm{e}'_\ell(\alpha_* G + Z, \kappa_*) \cdot G\right] \leq \delta\kappa_* \cdot \sqrt{\mathbb{E}\left[\left(\mathrm{e}'_\ell(\alpha_* G + Z, \kappa_*)\right)^2)\right]} = \delta\kappa_* \frac{\alpha_*}{\sqrt{\delta}\kappa_*}$$
$$\Rightarrow \delta \geq 1. \tag{2.24}$$

Remark 6 *Assume e_ℓ is two times differentiable (e.g., this is the case if ℓ is two times differentiable). Then, applying Stein's formula* (2.20)*, a simple rearrangement of* (2.23) *shows that*

$$\alpha_*^2 = \frac{1}{\delta} \frac{\mathbb{E}\left[\left(\mathrm{e}'_\ell(\alpha_* G + Z, \kappa_*)\right)^2\right]}{\left(\mathbb{E}\left[\mathrm{e}''_\ell(\alpha_* G + Z, \kappa_*)\right]\right)^2}. \tag{2.25}$$

The formula above coincides with the corresponding expression in [56], but the latter requires additional smoothness and strong-convexity assumptions on ℓ, which are not necessary for (2.23) *to hold. The proof of [56] is based on the AMP framework [62].*

Remark 7 *The simplest instance of the general M-estimator is the Least-squares, i.e. $\hat{\beta} := \min_\beta \|\mathbf{y} - \mathbf{X}\beta\|_2^2$. Of course, in this case, $\hat{\beta}$ has a closed form expression*

which can be directly used to predict the error behavior [170]. However, for illustration purposes, we show how the same result can be also obtained from (2.23)*. This is also one of the few cases where α_* can be expressed in closed form. Assume $\delta > 1$ and $\mathbf{z}_j \overset{iid}{\sim} p_Z$ with bounded second moment, i.e. $0 < \mathbb{E}Z^2 = \sigma^2 < \infty$. Then, it can be readily checked that all assumptions hold for $\frac{1}{2}(\cdot)^2, p_z$. Also, $e'_{\frac{1}{2}(\cdot)^2}(\chi;\tau) = \frac{\chi}{1+\tau}$ and $e''_{\frac{1}{2}(\cdot)^2}(\chi;\tau) = \frac{1}{1+\tau}$. Solving for the second equation in* (2.23) *gives $\kappa_* = \frac{1}{\delta-1}$. Substituting this into the first, we recover the well-known formula*

$$\alpha_*^2 = \sigma^2 \frac{1}{\delta - 1}. \tag{2.26}$$

Ridge Regularization

A popular regularizer in the machine learning and statistics literature is the ridge regularizer (also known as Tikhonov regularizer), i.e.

$$\hat{\beta} := \arg\min_{\beta} \sum_{j=1}^{n} \ell(\mathbf{y}_j - \mathbf{x}_j^T \beta_j) + \lambda \frac{\|\beta\|_2^2}{2}. \tag{2.27}$$

We specialize Theorem 1 to that case. For simplicity, we assume a separable loss function, and, $\mathbf{z}_j \overset{\text{iid}}{\sim} p_Z$ and $\beta_{0_i} \overset{\text{iid}}{\sim} p_X$.

We will apply Theorem 2. Suppose that ℓ satisfies the assumptions. Also, assume $\mathbb{E}\beta_0^2 = \sigma_\beta^2 < \infty$. Then, for $f = \frac{1}{2}(\cdot)^2$, it is easily verified that $\mathbb{E}[(f'(cH + \beta_0))^2] = \mathbb{E}[(cH + \beta_0)^2] < \infty$. Hence, the squared-error of (2.27) is predicted by α_*, the unique solution to (2.19) with

$$F(c,\tau) = \frac{c^2 + \sigma_\beta^2}{2(\tau + 1)} - \sigma_\beta^2.$$

The first-order optimality conditions (see (2.19)) of this problem simplify after some algebra to the following two equations in two unknowns:

$$\begin{cases} \alpha^2 = \delta\kappa^2 \cdot \mathbb{E}\left[\mathrm{e}'_\ell(\alpha G + Z, \kappa)^2\right] + \lambda^2\kappa^2\sigma_\beta^2, \\ \alpha\,(1 - \lambda\kappa) = \delta\kappa \cdot \mathbb{E}\left[\mathrm{e}'_\ell(\alpha G + Z, \kappa) \cdot G\right]. \end{cases} \tag{2.28}$$

Remark 8 *Assume* $\mathrm{prox}_\ell(x;\tau)$ *is two times differentiable with respect to c (e.g., this is the case if ℓ is two times differentiable), and write* $\mathrm{prox}'_\ell(x,\tau)$ *for the derivative with respect to x. Applying* (2.20)*, a simple rearrangement of* (2.28) *yields the following equivalent system of equations*

$$\begin{cases} \delta - 1 + \kappa\lambda = \delta \cdot \mathbb{E}\left[\mathrm{prox}'_\ell(\alpha G + Z;\kappa)\right], \\ \alpha^2 = \delta\mathbb{E}\left[\left(\alpha G + Z - \mathrm{prox}_\ell(\alpha G + Z;\kappa)\right)^2\right] + \lambda^2\kappa^2\sigma_\beta^2. \end{cases} \tag{2.29}$$

The formula above coincides with the corresponding expression in [95, Thm. 2.1][6]. The result in [95] requires additional smoothness assumptions on ℓ. Our result holds under relaxed assumptions and has been derived as a corollary of Theorem 1. On the other hand, [95, Thm. 2.1] is shown to be true for design matrices $\mathbf{X}$ with iid entries beyond Gaussian, e.g. sub-Gaussian.

Remark 9 *Consider a least-squares loss function where $\ell(x) = \frac{1}{2}x^2$ and a noise distribution of variance $\mathbb{E}Z^2 = \sigma_z^2 < \infty$. Then $\mathrm{prox}_\ell(x;\tau) = \frac{x}{1+\tau}$ and $\mathrm{prox}'_\ell(x;\tau) = \frac{1}{1+\tau}$. Substituting in* (2.29) *gives*

$$\begin{cases} 1-\kappa\lambda = \dfrac{\delta\kappa}{1+\kappa}, \\ \alpha^2\left(1-\delta\cdot\dfrac{\kappa^2}{(1+\kappa)^2}\right) = \delta\cdot\dfrac{\kappa^2}{(1+\kappa)^2}\sigma_z^2 + \lambda^2\kappa^2\sigma_\beta^2. \end{cases} \tag{2.30}$$

Now, we can solve these to get the following closed form expression for α^*:

$$\alpha^2 = \left(\delta\cdot\frac{\kappa^2}{(1+\kappa)^2}\cdot\sigma_z^2 + \lambda^2\sigma_\beta^2\kappa^2\right)\cdot\left(1-\delta\cdot\frac{\kappa^2}{(1+\kappa)^2}\right)^{-1}, \tag{2.31}$$

where

$$\kappa = \frac{1-\delta-\lambda+\sqrt{(1-\delta-\lambda)^2+4\lambda}}{2\lambda}. \tag{2.32}$$

Observe that letting $\lambda \to 0$ (which would correspond to ordinary least-squares) and assuming $\delta > 1$, κ in (2.32) *approaches $1/(\delta-1)$ and the optimal α^2 in* (2.31) *becomes $\sigma_z^2/(\delta-1)$, which agrees with* (2.26)*, as expected.*

Remark 10 *Let a Gaussian input distribution $\beta_{0,i} \overset{iid}{\sim} \mathcal{N}(0,1)$ and any noise distribution of power $\mathbb{E}Z^2 = \sigma_z^2 < \infty$. We show that a ridge-regularized M-estimator with a least-squares loss function and optimally tuned λ achieves asymptotically the Minimum Mean-Squared Error (MMSE) of estimating β_0 from $\mathbf{y} = \mathbf{X}\beta_0 + \mathbf{z}$.*

First, we use the results of Remark 9 to calculate the achieved error of the M-estimator optimized over the values of the regularizer parameter:

$$o_* := \inf_{\lambda>0}\lim_{p\to\infty}\frac{1}{p}\|\hat{\beta}-\beta_0\|_2^2 = \inf_{\lambda>0}\left\{\alpha^2(\kappa(\lambda),\lambda) \text{ as in } (2.31) \mid \kappa(\lambda) \text{ satisfies } (2.32)\right\}. \tag{2.33}$$

[6] In comparing (2.29) to [95, Eqn. (4)], due to some differences in normalizations the following "dictionary" needs to be used to match the results: $\alpha \leftrightarrow r_\rho(\kappa)$, $\kappa \leftrightarrow c_\rho(\kappa)$, $\delta^{-1}\lambda \leftrightarrow \tau$ and $\delta^{-1} \leftrightarrow \kappa$.

The optimization over λ is possible as follows. From (2.30)*, we find*

$$\delta\left(\frac{\kappa}{\kappa+1}\right)^2 = \frac{(1-\kappa\lambda)^2}{\delta}. \tag{2.34}$$

Substituting this in (2.31)*, and denoting $\beta = \kappa\lambda$, gives*

$$\alpha^2 = \frac{\delta\beta^2 + \sigma^2(1-\beta)^2}{\delta - (1-\beta)^2}. \tag{2.35}$$

Minimizing α^2 over $\lambda > 0$ in (2.33) *is equivalent to minimizing the fraction above over $0 < \beta < 1$, since there always exist κ, λ satisfying $\beta = \kappa\lambda$ and* (2.34)*. Thus, performing the optimization over $0 < \beta < 1$ in* (2.35) *we find*

$$o_* = \frac{1}{2}\left(1 - \sigma^2 - \delta + \sqrt{(1-\delta)^2 + 2\sigma^2(\delta+1) + \sigma^4}\right). \tag{2.36}$$

To complete the proof of the claim, we combine this with [196, Thm. 8, Eqn. (56)], where it is shown that the MMSE is given by the same expression as in the right-hand side above.

Cone-constrained M-estimators

Constrained M-estimators solve

$$\hat{\beta} = \arg\min_{\beta\in C} \sum_{j=1}^{n} \ell(\mathbf{y}_j - \mathbf{x}_j^T\beta), \tag{2.37}$$

for some set $\beta \in C$. The role of the regularizer in (2.10) is played here by the constraint $\beta \in C$. In particular, it is common that C takes the form $C = \{\beta \mid g(\beta) \leq g(\beta_0)\}$, i.e., it is chosen as the set of descent directions of some convex function g. In this setting, β_0 is assumed to be a structured signal (e.g. sparse, low-rank) and g is chosen to promote the particular structure (e.g. ℓ_1-norm, nuclear-norm) [40, 73, 129, 133]. Of course, such a formulation assumes prior knowledge of the value of g at β_0. Also, in this case, there exists by Lagrangian duality a value of λ for which the regularized M-estimator with $f(x) = g(x)$ is equivalent to (2.37).

A relaxation that is often undertaken to facilitate the analysis of (2.37) involves substituting C by its conic hull, which is also known as the tangent cone of g at β_0 (e.g. [40]). We call the resulting program, a *cone-constrained M-estimator*. For the special case of an ℓ_2-loss function, the squared error performance of constrained M-estimators has been previously considered in [129, 154] (also, see Remark 12

below). The analysis was performed in the high-SNR regime, where noise variance approaches zero. In this regime it was shown that the conic-relaxation above is *exact*. In this section, we analyze the error performance of cone-constrained M-estimators with general loss functions and derive some interesting conclusions. Before proceeding further, observe that (2.37) is an instance of (2.10) with a *non-separable* regularizer; hence, it also serves to showcase the applicability of Theorem 1 to such settings.

Consider solving (2.37) with

$$C = \mathcal{K} + \beta_0 := \{\lambda \mathbf{h} \mid \lambda \geq 0, g(\beta_0 + \mathbf{h}) \leq g(\beta_0)\} + \beta_0$$

and g a proper, closed, convex function. Here, $\mathcal{K}$ is the tangent cone of g at β_0, and β_0 is assumed fixed. The constrained minimization above can be written in the general form of regularized M-estimators in (2.10)LASSO by choosing the regularizer to be the indicator function for the cone, i.e. $f(\beta) = \delta_{\{\beta \in C\}}$. Let $\mathrm{Dist}_C(\mathbf{v})$, denote the distance of a vector $\mathbf{v}$ to a set C. We have,

$$\mathrm{e}_{\delta_{\{\beta \in C\}}}(c\mathbf{h} + \beta_0; \tau) = \frac{1}{2\tau} \min_{\mathbf{v} \in C - \beta_0} \|c\mathbf{h} - \mathbf{v}\|_2^2 = \frac{1}{2\tau} \mathrm{Dist}_{\mathcal{K}}^2(c\mathbf{h}) = \frac{c^2}{2\tau} \mathrm{Dist}_{\mathcal{K}}^2(\mathbf{h}).$$

In the last equality above we have used the homogeneity of the cone $\mathcal{K}$. Let $\mathcal{K}^\circ$ denote the polar cone of $\mathcal{K}$, and,

$$D_{\mathcal{K}} := \mathbb{E}\left[\mathrm{Dist}_{\mathcal{K}^\circ}^2(\mathbf{h})\right] = \mathbb{E}\left[\|\mathbf{h}\|_2^2 - \mathrm{Dist}_{\mathcal{K}}^2(\mathbf{h})\right].$$

This quantity is known as the *statistical dimension* [6] of the cone $\mathcal{K}$, or, as the *Gaussian distance squared* [129]. It can be though of as a measure of the size of the cone, and also, it is very closely related to the *gaussian width* of $\mathcal{K}$[6]. We assume that

$$\frac{D_{\mathcal{K}}}{p} \to \overline{D}_{\mathcal{K}} \in (0, 1). \tag{2.38}$$

This translates to an assumption on the degrees of freedom of the structured signal β_0 being proportional to its dimension. For example, for a k-sparse β_0 and $g(\beta) = \|\beta\|_1$, (2.38) is satisfied for $k = \rho p$, $\rho \in (0, 1)$.

With (2.38), Assumption 1(a) holds with $F(c, \tau) = \frac{c^2}{2\tau}(1 - \overline{D}_{\mathcal{K}})$. For this, it is straightforward to check that Assumption 1(b) is also satisfied. Overall, if ℓ, p_Z satisfy the conditions of Theorem 2 and g, β_0 are such that (2.38) holds, then

Theorem 1 applies, and the squared error of the cone-constrained M-estimator in (2.37) is predicted by the unique minimizer α_* of the (SPO) problem below:

$$\inf_{\substack{\alpha\geq 0\\ \tau_g>0}} \sup_{\gamma\geq 0} \frac{\gamma\tau_g}{2} + \delta\cdot\mathbb{E}\left[\mathrm{e}_\ell\left(\alpha G + Z;\tau_g/\gamma\right) - \ell(Z)\right] - \alpha\gamma\sqrt{\overline{D}_{\mathcal{K}}}. \tag{2.39}$$

Compared to (2.8) (which involves six optimization variables), we have performed the (straightforward) optimization over τ_h: $\inf_{\tau_h>0}\frac{\tau_h}{2}+\frac{\gamma^2\overline{D}_{\mathcal{K}}^2}{2\tau_h} = \gamma\overline{D}_{\mathcal{K}}$.

Remark 11 *Starting from* (2.39) *we can conclude on the minimum number of measurements required for stable recovery. We show that the normalized number of measurements δ need to be at least as large as $\overline{D}_{\mathcal{K}}$, in order for the error to be finite. This is to be compared with the case where no regularization is used that required $\delta\geq 1 > \overline{D}_{\mathcal{K}}$ (see Remark 5). To prove the claim, assume finite error, then the value where it converges is predicted by* (2.39)*. Standard first-order optimality conditions give*[7]

$$\gamma - \frac{\delta}{\gamma}\mathbb{E}\left[\left(\mathrm{e}_\ell'(\alpha G+Z;\tau_g/\gamma)\right)^2\right] \geq 0, \tag{2.40a}$$

$$\delta\mathbb{E}[\mathrm{e}_\ell'(\alpha G+Z;\tau_g/\gamma)\cdot G] - \gamma\sqrt{\overline{D}_{\mathcal{K}}} \geq 0, \tag{2.40b}$$

$$\frac{\tau}{2} + \frac{\delta\tau}{2\gamma^2}\mathbb{E}\left[\left(\mathrm{e}_\ell'(\alpha G+Z;\tau_g/\gamma)\right)^2\right] - \alpha\sqrt{\overline{D}_{\mathcal{K}}} \leq 0. \tag{2.40c}$$

Starting from the second equation, applying the Cauchy-Schwarz inequality and substituting back the first equation we conclude as follows:

$$\gamma\sqrt{\overline{D}_{\mathcal{K}}} \leq \delta\mathbb{E}[\mathrm{e}_\ell(\alpha G+Z;\tau_g/\gamma)\cdot G] \leq \delta\sqrt{\mathbb{E}[\left(\mathrm{e}_\ell'(\alpha G+Z;\tau_g/\gamma)\right)^2]} \leq \delta\frac{\gamma}{\sqrt{\delta}} \Rightarrow \delta\geq\overline{D}_{\mathcal{K}}. \tag{2.41}$$

Remark 12 *Consider a least-squares loss function and a noise distribution of variance $\mathbb{E}Z^2=\sigma^2<\infty$. Then, the solution to* (2.39) *admits an insightful closed form expression. First, in* (2.39) *perform the optimization over τ_g. Equating* (2.40a) *to 0, gives $\tau_g = \sqrt{\delta}\sqrt{\alpha^2+\sigma^2}-\gamma$. Substituting this in* (2.39)*, we are left to solve for*

$$\inf_{\alpha\geq 0}\sup_{\gamma\geq 0}\gamma\left(\sqrt{\delta}\sqrt{\alpha^2+\sigma^2}-\alpha\sqrt{\overline{D}_{\mathcal{K}}}\right)-\frac{\gamma^2}{2}.$$

[7] The three equations in (2.40) correspond to differentiation of the objective of (2.39) with respect to τ,α and γ, respectively. If any of the variables is zero at the optimal, then, the corresponding equation holding with an inequality is necessary and sufficient. On the other hand, if the optimal is strictly positive, then the equation should hold with equality.

It can be easily checked that if $\delta > \overline{D}_{\mathcal{K}}$, then the optimal α_* is

$$\alpha_*^2 = \sigma^2 \frac{\overline{D}_{\mathcal{K}}}{\delta - \overline{D}_{\mathcal{K}}}. \tag{2.42}$$

It is insightful to compare this with (2.26)*, the corresponding error formula for least-suares: the only difference is that* 1 *is substituted with the statistical dimension $\overline{D}_{\mathcal{K}}$. Also, verifying the conclusion of the previous remark, we now require $\delta > \overline{D}_{\mathcal{K}}$ instead of $\delta > 1$, implying that recovery is in general possible with less measurements than the dimension of the signal.*

The result in (2.42) *was first proved for ℓ_1-regularization in [154], and, was later generalized in [129, 169] (also, [170]). In contrast to the lengthy treatments in those references, the result was derived here as a simple corollary of Theorem 1.*

Remark 13 *In* (2.40b) *apply Stein's inequality and combine it with* (2.40a) *to yield*

$$\alpha^2 \geq \frac{\overline{D}_{\mathcal{K}}}{\delta} \frac{\gamma^2/\delta}{\mathbb{E}\left[\mathrm{e}_\ell''(\alpha G + Z; \tau_g/\gamma)\right]} \geq \frac{\overline{D}_{\mathcal{K}}}{\delta} \frac{\mathbb{E}\left[\left(\mathrm{e}_\ell'(\alpha G + Z; \tau_g/\gamma)\right)^2\right]}{\mathbb{E}\left[\mathrm{e}_\ell''(\alpha G + Z; \tau_g/\gamma)\right]}. \tag{2.43}$$

For the first inequality above, we have assumed that at the optimal, $\mathbb{E}\left[\mathrm{e}_\ell''(\alpha G + Z; \tau_g/\gamma)\right] < \infty$. When this holds, (see Remark 14 for an instance where this is not the case) we can use the above to lower bound the error performance in terms of the Fisher information of the noise. Based on a result of [109], Donoho and Montanari prove in [56, Lem. 3.4,3.5] that the right-hand side in (2.43) *is further lower bounded by $I(Z)/(1+\alpha^2 I(Z))$, where $I(Z) = \mathbb{E}\left(\frac{\partial}{\partial z} \log p_Z(z)\right)^2$ denotes the Fisher information of the random variable Z, which is assumed to have a differentiable density. Using this and solving for α^2, we conclude with*

$$\alpha^2 \geq \frac{\overline{D}_{\mathcal{K}}}{\delta - \overline{D}_{\mathcal{K}}} \frac{1}{I(Z)}. \tag{2.44}$$

For Gaussian noise of variance σ^2, we have $1/I(Z) = \sigma^2$. In this case the lower bound in (2.44) *coincides with the error formula of the least-squares loss function, thus proving optimality of the latter.*

Remark 14 *The lower bound in* (2.44) *only holds if the optimal α_* in* (2.39) *is strictly positive. This is not always the case: under circumstances, it is possible to choose the loss function such that the resulting cone-constrained M-estimator is*

consistent, i.e. $\alpha_* = 0$*. Theorem 1 is the starting point to identifying such interesting scenarios.*

Here, we illustrate this through an example: we assume a sparse gaussian-noise model and use a Least Absolute Deviations (LAD) loss function. More precisely, $p_Z(Z) = \bar{s}\delta_0(Z) + (1-\bar{s})\frac{1}{\sqrt{2\pi}}\exp(-Z^2/2), \bar{s} \in (0,1)$ *and* $\ell(v) = |v|$*. In Section .3 we prove that when* $\bar{s}, \delta$ *and* $\overline{D}_{\mathcal{K}}$ *are such that*

$$\delta \geq \overline{D}_{\mathcal{K}} + \min_{\kappa>0} \left\{ \bar{s}(1+\kappa^2) + (\delta - \bar{s})\sqrt{\frac{2}{\pi}} \int_{\kappa}^{\infty} (G-\kappa)^2 \exp(-G^2/2) \mathrm{d}G \right\}, \quad (2.45)$$

then the first-order optimality conditions in (2.40) *are satisfied for* $\alpha \to 0, \tau_g \to 0$ *and some* $\gamma > 0$*. Thus, when the number of measurements is large enough such that* (2.45) *holds, then* $\alpha_* = 0$*, and,* β_0 *is perfectly recovered*[8]*.*

Generalized LASSO

The generalized LASSO solves

$$\hat{\beta} := \arg\min_{\beta} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda f(\beta). \quad (2.46)$$

For simplicity, suppose that f is separable and satisfies the assumptions of Theorem 2. Also, assume $\mathbf{z}_j \overset{\text{iid}}{\sim} p_Z$ such that $0 < \mathbb{E}Z^2 =: \sigma^2 < \infty$. Then, for $\ell = \frac{1}{2}(\cdot)^2$, it is easily verified that $\mathbb{E}[(\ell'(cG+Z))^2] = \mathbb{E}[(cG+Z)^2] < \infty$. Hence, the squared-error of (2.46) is predicted by α_*, the unique solution to the (2.19) with $L(c,\tau) = \frac{c^2+\sigma^2}{2(\tau+1)} - \sigma^2$.

Equivalently, the error is predicted by the solution to the stationary equations in (2.19) with $e'_{\frac{1}{2}(\cdot)^2}(\chi;\tau) = \frac{\chi}{1+\tau}$. The second and third equations in (2.19) give

$$\gamma^2(1+\kappa)^2 = \delta(\alpha^2+\sigma^2),$$
$$\nu(1+\kappa) = \delta.$$

[8] In the context that it appears here, the perfect recovery condition in (2.45) has been shown previously in [167]. The problem is very closely related to the demixing problem in which one aims to extract two (or more) constituents from a mixture of structured vectors [112]. In that context, recovery conditions like the one in (2.45) have been generalized to other kinds of structures beyond sparsity [73, 112, 113]. Our purpose here has been to illustrate how Theorem 1 can be used to derive such results. Besides, the generality of the paper's setup offers the potential to extend such consistency-type results beyond cone-constrained M-estimators and beyond fixed signals β_0. This is an interesting direction for future research.

Solving these for κ and ν, and substituting them in the remaining two equations results in the following system of two nonlinear equations in two unknowns

$$\begin{cases} \delta \frac{\alpha^2}{\alpha^2+\sigma^2} = \mathbb{E}\left[\left(\frac{\lambda}{\gamma} e'_f\left(\frac{\sqrt{\alpha^2+\sigma^2}}{\sqrt{\delta}} H + \beta_0, \lambda \frac{\sqrt{\alpha^2+\sigma^2}}{\gamma\sqrt{\delta}}\right) - H\right)^2\right], \\ \gamma(1-\delta) + \gamma^2 \frac{\sqrt{\delta}}{\sqrt{\alpha^2+\sigma^2}} = \lambda \mathbb{E}\left[e'_f\left(\frac{\sqrt{\alpha^2+\sigma^2}}{\sqrt{\delta}} H + \beta_0, \lambda \frac{\sqrt{\alpha^2+\sigma^2}}{\gamma\sqrt{\delta}}\right) \cdot H\right]. \end{cases} \tag{2.47}$$

For the special case of ℓ_1-regularization, the result above was proved by Bayati and Montanari [18] using the AMP framework. In the generality presented here, the result appears to be novel.

Remark 15 *An interesting observation from* (2.47) *is that the generalized LASSO cannot achieve perfect recovery, irrespective of the choice of the regularizer function. To see this, the first equation in* (2.47) *for* $\alpha = 0$ *gives* $\mathbb{E}\left[\left(\frac{\lambda}{\gamma} e'_f(\frac{\sigma}{\sqrt{\delta}} H + \beta_0, \frac{\lambda\sigma}{\gamma\sqrt{\delta}}) - H\right)^2\right] = 0$. *Then, it must hold, almost surely, that the argument under the expectation sign be equal to zero. Evaluating the derivative of the envelope function, this becomes equivalent to* $\beta_0 = \operatorname{prox}_f\left(\frac{\sigma}{\sqrt{\delta}} H + \beta_0; \frac{\lambda\sigma}{\gamma\sqrt{\delta}}\right)$. *This, when combined with the optimality conditions for the Moreau envelope gives that almost surely* $\frac{\sigma}{\sqrt{\delta}} H \in \partial f(\beta_0)$. *Thus, we have reached a contradiction because* H *can take any real value as a Gaussian random variable.*

Square-root LASSO

The (Generalized) Square-root LASSO (also known as ℓ_2-LASSO [129]) solves[9]

$$\hat{\beta} := \arg\min_{\beta} \sqrt{p} \|\mathbf{y} - \mathbf{X}\beta\|_2 + \lambda f(\beta). \tag{2.48}$$

In contrast to the other examples in this section, the square-root LASSO is an instance of (2.10) with a *non*-separable loss function. Observe the normalization of the loss function with a $\sqrt{p}$-factor. This is to make the loss function of the same order as the regularizer.

One can show that when $\mathcal{L}(\mathbf{v}) = \sqrt{p}\|\mathbf{v}\|_2$ and $\mathbf{z} \sim p_{\mathbf{z}}$ with $\mathbb{E}\left[\|\mathbf{z}\|_2^2/n\right] = \sigma^2 \in (0, \infty)$, then Assumption 1(a) holds with

$$L(\alpha, \tau) = \begin{cases} \frac{1}{\sqrt{\delta}}(\sqrt{\alpha^2+\sigma^2} - \sigma) - \frac{\tau}{2\delta} & \text{, if } \sqrt{\delta}\sqrt{\alpha^2+\sigma^2} \geq \tau, \\ \frac{1}{2\tau}(\alpha^2+\sigma^2) - \frac{\sigma}{\sqrt{\delta}} & \text{, otherwise.} \end{cases} \tag{2.49}$$

[9] We refer the interested reader to [22, 129, 169] for a discussion on the similarities and differences between (2.48) and the Generalized LASSO in (2.46).

Also, Assumption 1(b) is trivially satisfied. Thus, considering any regularizer that satisfies Assumptions 1(b), Theorem 1 applies, and predicts the squared error of (2.48) as the unique minimizer α_* to the following optimization:

$$\inf_{\alpha\geq 0}\ \sup_{\substack{\gamma\geq 0\\ \tau_h>0}}\ -\frac{\alpha\tau_h}{2}-\frac{\alpha\gamma^2}{2\tau_h}+\lambda\cdot F\left(\frac{\alpha\gamma}{\tau_h},\frac{\alpha\lambda}{\tau_h}\right)+\begin{cases}\gamma\sqrt{\delta}\sqrt{\alpha^2+\sigma^2} & ,\text{if } \gamma\leq 1\\ \sqrt{\delta}\sqrt{\alpha^2+\sigma^2} & ,\text{otherwise}\end{cases}. \tag{2.50}$$

To arrive to (2.50) starting from (2.8), we have replaced L with (2.49) and have performed the minimization over τ_g as shown below:

$$\inf_{\tau_g\geq 0}\begin{cases}\frac{\gamma\tau_g}{2}-\frac{\tau_g}{2\gamma}+\sqrt{\delta}\sqrt{\alpha^2+\sigma^2} & ,\text{if } \delta(\alpha^2+\sigma^2)\geq\frac{\tau_g^2}{\gamma^2}\\ \frac{\gamma\delta}{2\tau_g}(\alpha^2+\sigma^2)+\frac{\gamma\tau_g}{2} & ,\text{otherwise.}\end{cases}=\begin{cases}\beta\sqrt{\delta}\sqrt{\alpha^2+\sigma^2} & ,\text{if } \gamma\leq 1\\ \sqrt{\delta}\sqrt{\alpha^2+\sigma^2} & ,\text{otherwise}\end{cases}. \tag{2.51}$$

The optimization in (2.51) can be simplified one step further. One can easily show that $-\frac{\alpha\gamma^2}{2\tau_h}+\lambda F\left(\frac{\alpha\gamma}{\tau_h},\frac{\alpha\lambda}{\tau_h}\right)$ is a non-increasing function of γ for $\gamma>0$. Therefore, the (SPO) becomes equivalent to the following

$$\inf_{\alpha\geq 0}\ \sup_{\substack{0\leq\gamma\leq 1\\ \tau_h\geq 0}}\ \gamma\sqrt{\delta}\sqrt{\alpha^2+\sigma^2}-\frac{\alpha\tau_h}{2}-\frac{\alpha\gamma^2}{2\tau_h}+\lambda\cdot F\left(\frac{\alpha\gamma}{\tau_h},\frac{\alpha\lambda}{\tau_h}\right). \tag{2.52}$$

The fact that the optimization in (2.52) predicts the squared error of (2.48), has been recently shown by the authors in [164]. That work only considers the square-root LASSO[10], while here, we have (re)-derived the result as a corollary of the general Theorem 1.

Heavy-tails

In this section, we investigate instances where the noise distribution has unbounded moments. In the presence of (say) heavy-tailed noise, it is a common practice to use a loss function that grows to infinity no faster than linearly. This is also suggested by Assumption 1(a) (cf. (2.16) for the separable case), as has already been discussed.

For a mere illustration, we assume $\mathbf{z}\overset{\text{iid}}{\sim}\text{Cauchy}(0,1)$ and consider two examples of loss functions for which we show that Theorem 1 is applicable.

[10]Note however, that [164] considers a more general measurement model than the one of the current paper, one that allows for nonlinearities.

LAD

As a first example, consider the regularized-LAD estimator:

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_1 + \lambda f(\beta). \tag{2.53}$$

The loss function is separable, with $\ell(v) = |v|$. Easily, for all $c \in \mathbb{R}$

$$\mathbb{E}\left[|\ell'_+(cG + Z)|^2\right] = \mathbb{E}\left[|\text{sign}(cG + Z)|^2\right] = 1 < \infty,$$

satisfying the assumption in (2.14). Also, $\mathbb{E}Z^2$ is undefined, but, $\sup_v \frac{\ell(v)}{|v|} = 1 < \infty$, thus, (2.16) holds. Finally, $|\cdot|$ is not differentiable at zero satisfying the conditions of Lemma 4. With these, Theorem 2 is applicable.

Huber-loss

The Huber-loss function with parameter $\rho > 0$ is defined as

$$h_\rho(v) = \begin{cases} \frac{v^2}{2} & , |v| \le \rho, \\ \rho|v| - \frac{\rho^2}{2} & , \text{otherwise.} \end{cases} \tag{2.54}$$

Consider a regularized M-estimator with $\ell(v) = h_\rho(v)$. We show here that this choice satisfies the Assumptions of Theorem 2. Indeed, for all $c \in \mathbb{R}$

$$\mathbb{E}\left[|\ell'_+(cG + Z)|^2\right] \le \mathbb{E}\left[|cG + Z| \,\middle|\, |cG + Z| \le \rho\right] + \mathbb{E}\left[\rho \,\middle|\, |cG + Z| > \rho\right] < \infty, \tag{2.55}$$

satisfying the assumption in (2.14). Also, $\sup_v \frac{\ell(v)}{|v|} = \rho < \infty$, thus, (2.16) holds. Finally, h_ρ is differentiable with a strictly increasing derivative in the interval $[-\rho, \rho]$. With these, Theorem 2 is applicable. Figure 2.3 illustrates the validity of the prediction via numerical simulations.

Numerical Simulations

We have performed a few numerical simulations on specific instances of M-estimators that were previously discussed in Section 2.1. Their purpose is to illustrate the validity of the prediction of Theorem 1, and of the remarks that followed as a consequence of it.

Figure 2.1 . We consider the regularized LAD estimator of (2.53) under an iid sparse-Gaussian noise model. The unknown signal is also considered sparse, which leads to the natural choice of ℓ_1 regularization, i.e. $f(\beta) = \|\beta\|_1$. Apart from the very close agreement of the theoretical prediction of Theorem 1 to the simulated data, the following facts are worth observing.

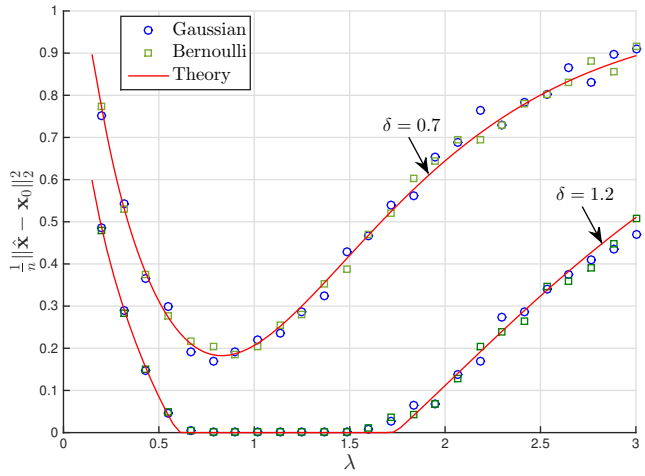


Figure 2.1: Squared error of the l_1-Regularized LAD with Gaussian ($\circ$) and Bernoulli ($\square$) measurements as a function of the regularizer parameter λ for two different values of the normalized number of measurements, namely $\delta = 0.7$ and $\delta = 1.2$. Also, $\beta_{0,i} \overset{\text{iid}}{\sim} p_x(x) = 0.9\delta_0(x) + 0.1\phi(x)/\sqrt{0.1}$ and $\mathbf{z}_j \overset{\text{iid}}{\sim} p_z(z) = 0.7\delta_0(z) + 0.3\phi(z)$ for $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$. For the simulations, we used $p = 768$ and the data were averaged over 5 independent realizations.

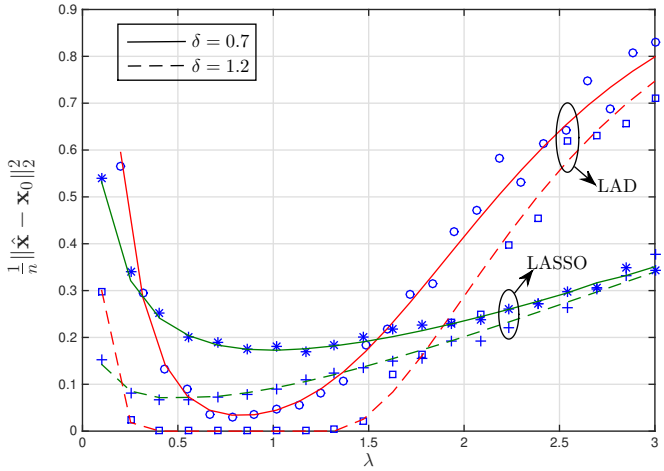


Figure 2.2: Comparing the squared error of the ℓ_1-Regularized LAD with the corresponding error of the LASSO. Both are plotted as functions of the regularizer parameter λ, for two different values of the normalized measurements, namely $\delta = 0.7$ and $\delta = 1.2$. The noise and signal are iid sparse-Gaussian as follows: $\beta_{0,i} \overset{\text{iid}}{\sim} p_x(x) = 0.9\delta_0(x) + 0.1\phi(x)/\sqrt{0.1}$ and $\mathbf{z}_j \sim p_z(z) = 0.9\delta_0(z) + 0.1\phi(z)$ with $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$. For the simulations, we used $p = 768$ and the data were averaged over 5 independent realizations.

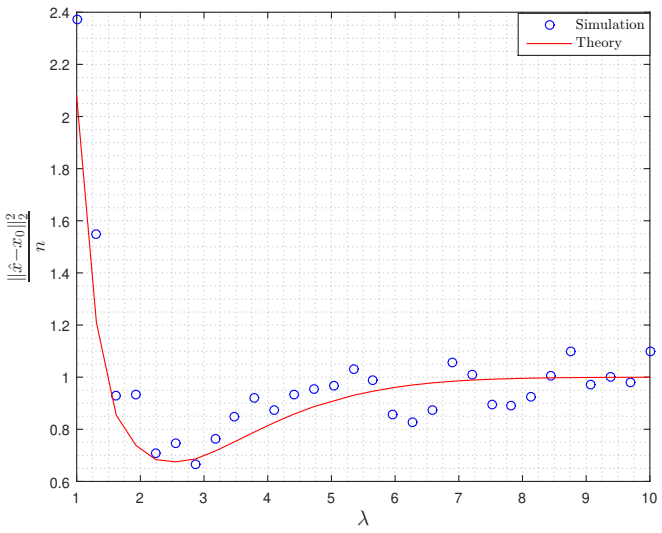


Figure 2.3: Squared error of the ℓ_1-Regularized M-Estimator with Huber-loss as a function of the regularizer parameter λ. Here, $\delta = 0.7$, $\beta_0 \overset{\text{iid}}{\sim} p_x(x) = 0.9\delta_0(x) + 0.1\phi(x)/\sqrt{0.1}$ and $p_z(z) = 0.9\delta(z) + 0.1\eta(z)$ with $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ and $\eta(z) = \frac{1}{\pi(1+z^2)}$. For the simulations, we used $p = 1024$ and the data are averaged over 5 independent realizations.

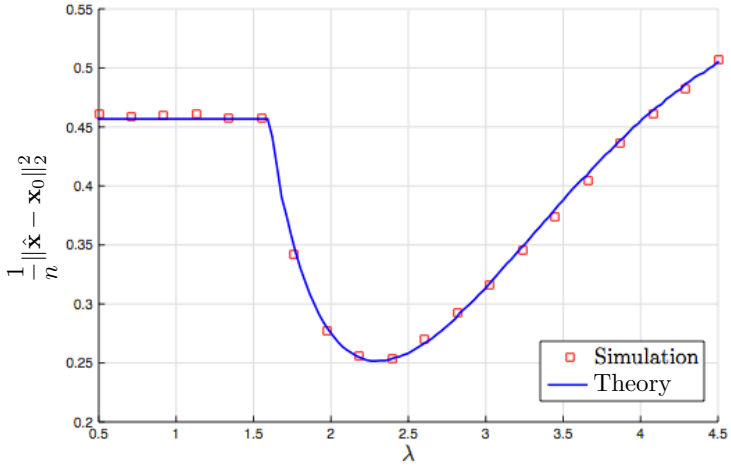


Figure 2.4: Squared error of the $\ell_{1,2}$-Regularized Lasso for group sparse signal composed of 512 blocks of size 3 each, as a function of the regularizer parameter λ. Here, $\delta = 0.75$, each block is zero with probability 0.95, otherwise its entries are i.i.d. $\mathcal{N}(0,1)$ and $\mathbf{z}_j \overset{\text{iid}}{\sim} p_z(z) = 0.3\phi(z)$ with $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$. The simulations are averaged over 10 independent realizations.

- When the number of measurements n gets large enough, then, for an appropriate range of values of the regularizer parameter, the estimator is consistent, i.e. the unknown signal β_0 is perfectly recovered. This is relevant to Remark 14 where we proved this to be the case for the closely related cone-constrained LAD estimator. For that, we were able to quantify how large n should be as a function of the sparsities of the noise and of the signal, see (2.45).

- The prediction of Theorem 1 remains accurate when the measurement matrix has entries iid Bernoulli ($\{\pm 1\}$). This suggests that the error behavior (at least of this specific instant of M-estimator) undergoes some universality properties.

Figure 2.2 . The model for both the noise and for the unknown signal here is the same as in Figure 2.1, i.e. both are iid sparse. We use ℓ_1-regularization, and, two different loss functions, namely, a least-absolute-deviations one and a least-squares one, corresponding to a LAD and a LASSO estimator, respectively. The figure aims to compare the performance of the two. Intuition suggests that the LAD is more appropriate for a sparse noise model, since ℓ_1 promotes sparsity. This is indeed the case, in the sense that for good choices of the regularizer parameter λ, the LAD outperforms by far the LASSO. However, it is worth observing that for a different and relatively big range of values of λ, the LASSO performs better. This indicates the importance of the correct tuning of the regularizer parameter, to which the predictions of Theorem 1 can offer valuable guidelines and insights.

Figure 2.3 . For this figure, we have assumed an ℓ_1-regularized estimator with Huber-loss $\ell(v) = h_1(v)$ (see (2.54)). The noise is iid Cauchy$(0, 1)$. In Section 2.1 it was shown that all the assumptions of Theorem 2 are satisfied in this setting. The figure, validates the prediction. To obtain the prediction we numerically solved the corresponding system of nonlinear equations (see (2.19)) using the efficient iterative scheme described in Remark 3.

Figure 2.4 . We include this as an example of an M-estimator with non-separable loss function. For the plot, we use the square-root LASSO with $\ell_{1,2}$-regularization. The analytical prediction was derived solving (2.52).

2.2 BER Analysis of the Box Relaxation for BPSK Signal Recovery

The problem of recovering an unknown BPSK vector from a set of noise corrupted linearly related measurements arises in numerous applications, such as Massive MIMO [41, 121, 126, 193]. As a result, a large host of exact and heuristic optimiza-

tion algorithms have been proposed. Exact algorithms, such as sphere decoding and its variants, become computationally prohibitive as the problem dimension grows. Heuristic algorithms such as zero-forcing, MMSE, decision-feedback, etc., [71, 81, 83] have inferior performances that are often difficult to precisely characterize. One popular heuristic is the so called "Box Relaxation" which replaces the discrete set $\{\pm 1\}^n$ with the convex set $[-1, 1]^n$ [108, 162, 198]. This allows one to recover the signal via convex optimization followed by hard thresholding. Despite its popularity, very little is known about the performance of this method. In this section, we exactly characterize its bit-wise error probability in the regime of large dimensions and under Gaussian assumptions.

Setup

Our goal is to recover an n-dimensional BPSK vector $\mathbf{x}_0 \in \{\pm 1\}^n$ from the noisy multiple-input multiple output (MIMO) relation $\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{z} \in \mathbb{R}^m$, where $\mathbf{A} \in \mathbb{R}^{m \times n}$ is the MIMO channel matrix (assumed to be known) and $\mathbf{z} \in \mathbb{R}^m$ is the noise vector. We assume that $\mathbf{A}$ has entries iid $\mathcal{N}(0, 1/n)$ and $\mathbf{z}$ has entries iid $\mathcal{N}(0, \sigma^2)$. The normalization is such that the reciprocal of the noise variance σ^2 is equal to the Signal-to-Noise Ratio, i.e. SNR $= 1/\sigma^2$.

Our goal is to recover an n-dimensional BPSK vector $\mathbf{x}_0 \in \{\pm 1\}^n$[11] from the noisy multiple-input multiple output (MIMO) relation $\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{z} \in \mathbb{R}^m$, where $\mathbf{A} \in \mathbb{R}^{m \times n}$ is the MIMO channel matrix (assumed to be known) and $\mathbf{z} \in \mathbb{R}^m$ is the noise vector. We assume that $\mathbf{A}$ has entries iid $\mathcal{N}(0, 1/n)$ and $\mathbf{z}$ has entries iid $\mathcal{N}(0, \sigma^2)$. The normalization is such that the reciprocal of the noise variance σ^2 is equal to the Signal-to-Noise Ratio, i.e. SNR $= 1/\sigma^2$.

The Maximum-Likelihood (ML) decoder. The ML decoder which maximizes the probability of error (assuming the $\mathbf{x}_{0,i}$ are equally likely) is given by $\min_{\mathbf{x} \in \{\pm 1\}^n} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2$. Solving the above, is often computationally intractable, especially when n is large, and therefore a variety of heuristics have been proposed (zero-forcing, mmse, decision-feedback, etc.) [185].

Box Relaxation Optimization. The heuristic we shall use, we refer to it as Box Relaxation Optimization (BRO). It consists of two steps. The first one involves solving a convex relaxation of the ML algorithm, where $\mathbf{x} \in \{\pm 1\}^n$ is relaxed to $\mathbf{x} \in [-1, 1]^n$. The output of the optimization is hard-thresholded in the second step

[11] For this sectoin, we are going to use the conventional notations in communications for consistency with related works. In these works, the dimension of the unknown signal is n (instead of p), while the number of measuements is m (instead of n)

to produce the final binary estimate. Formally, the algorithm outputs an estimate $\mathbf{x}^*$ of $\mathbf{x}_0$ given as

$$\begin{aligned} \hat{\mathbf{x}} &= \arg \min_{-1 \le \mathbf{x}_i \le 1} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2, \\ \mathbf{x}^* &= \operatorname{sign}(\hat{\mathbf{x}}), \end{aligned} \tag{2.56}$$

where the sign function returns the sign of its input and acts element-wise on input vectors.

Bit error probability. We evaluate the performance of the detection algorithm by the bit error probability P_e, defined as the expectation of the Bit Error Rate BER . Formally,

$$BER \ := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{\mathbf{x}_i^* \neq \mathbf{x}_{0,i}\}}, \tag{2.57a}$$

$$P_e := \mathbb{E}\left[BER\right] = \frac{1}{n} \sum_{i=1}^{n} \Pr\left(\mathbf{x}_i^* \neq \mathbf{x}_{0,i}\right). \tag{2.57b}$$

Our main result analyzes the P_e of the BRO in (2.56). We assume a large-system limit where $m, n \to \infty$ at a proportional rate δ. The SNR is assumed constant; in particular, it does not scale with n. Let $Q(\cdot)$ denote the Q-function associated with the standard normal density $p(h) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-h^2/2}$.

Theorem 3 (P_e of the BRO) *Let P_e denote the bit error probability of the detection scheme in* (2.56) *for some fixed but unknown BPSK signal $\mathbf{x}_0 \in \{\pm 1\}^n$. For constant* SNR *and $\frac{m}{n} \to \delta \in (\frac{1}{2}, \infty)$, it holds:*

$$\lim_{n \to \infty} P_e = Q(1/\tau_*),$$

where τ_* is the unique solution to

$$\min_{\tau > 0} \ \frac{\tau}{2}\left(\delta - \frac{1}{2}\right) + \frac{1/\mathrm{SNR}}{2\tau} - \frac{\tau}{2} \int_{\frac{2}{\tau}}^{\infty} \left(h + \frac{2}{\tau}\right)^2 p(h)\, dh. \tag{2.58}$$

Theorem 3 derives a *precise* formula for the bit error probability of the (BRO). The formula involves solving a *convex* and deterministic minimization problem in (2.58).

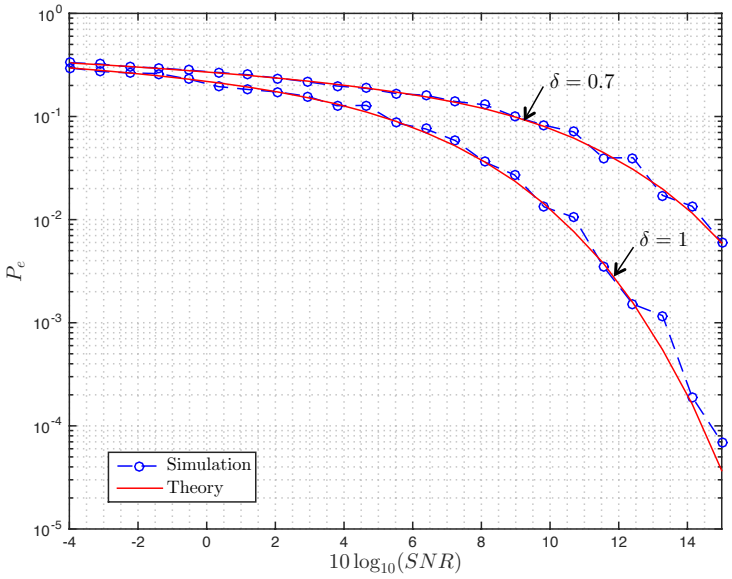


Figure 2.5: BER Performance of the Boxed Relaxation: P_e as a function of SNR for different values of the ration $\delta = \lceil m/n \rceil$. The theoretical prediction follows from Theorem 3. For the simulations, we used $n = 512$. The data are averages over 20 independent realizations of the channel matrix and of the noise vector for each value of the SNR.

Computing τ_*. It can be shown that the objective function of (2.58) is strictly convex when $\delta > \frac{1}{2}$. When $\delta < \frac{1}{2}$, it is well known that even the noiseless box relaxation fails [40]. (In fact, $\delta = \frac{1}{2}$ is the recovery threshold for this convex relaxation.) Thus, (2.58) has a unique solution τ_*. Observe that the problem parameters δ and SNR appear explicitly in (2.58); naturally then τ_* is indeed a function of those. The minimization in (2.58) can be efficiently solved numerically. In addition, owing to the strict convexity of the objective function, τ_* can be equivalently expressed as the unique solution to the corresponding first order optimality conditions.

Numerical illustration. Figure 2.5 illustrates the accuracy of the prediction of Theorem 3. Note that although the theorem requires $n \to \infty$, the prediction is already accurate for n ranging on a few hundreds.

P_e at high-SNR. It can be shown that when SNR $\gg 1$, then $\tau_* = 1/\sqrt{(\delta - 1/2)SNR}$. This can be intuitively understood as follows: at high-SNR, we expect τ_* to be going to zero (correspondingly P_e to be small). When this is the case, the last term in (2.58) is negligible; then, τ_* is the solution to $\min_{\tau>0} \frac{\tau}{2}\left(\delta - \frac{1}{2}\right) + \frac{1/\text{SNR}}{2\tau}$ which gives the derided result. Hence, for SNR $\gg 1$,

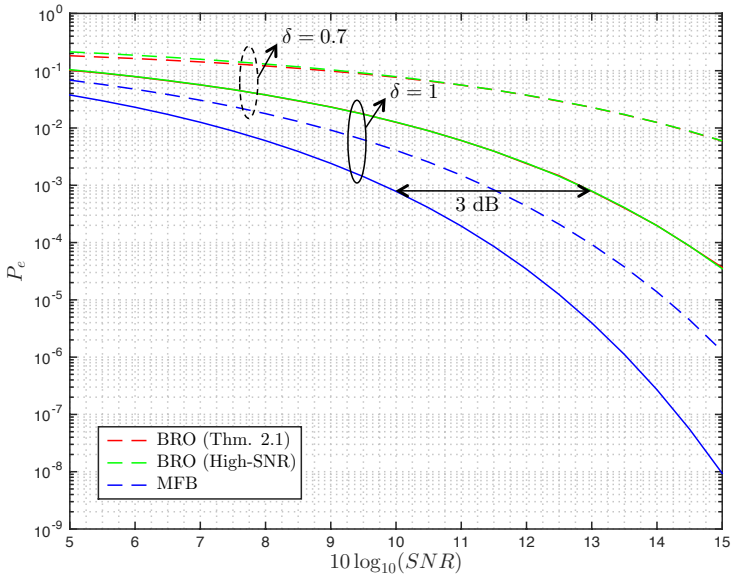


Figure 2.6: Bit error probability of the Box Relaxation Optimization (BRO) in (2.56) in comparison to the Matched Filter Bound (MFB) for $\delta = 0.7$ (dashed lines) and $\delta = 1$ (solid lines). The red curves follow the formula of Thm. 3, the green ones correspond to (2.59), and, P_e^{MFB} of (2.60) is in blue.

$$\lim_{n\to\infty} P_e \approx Q(\sqrt{(\delta - 1/2)\cdot \mathrm{SNR}}). \tag{2.59}$$

In Figure 2.6 we have plotted this high-SNR expression for the $\log_{10}(P_e)$ vs its exact value as predicted by Theorem 3. It is interesting to observe that the former is actually a very good approximation to the latter even for small practical values of SNR. The range of SNR values for which the approximation is valid becomes larger with increasing δ. Heuristically, for $\delta > 0.7$ the expression in (2.59) is a good proxy for the true probability of error at practical SNR values.

Comparison to the matched filter bound. Theorem 3 gives us a handle on the P_e of BRO in (2.56) and therefore allows to evaluate its practical performance. Here, we compare the performance to an idealistic case, where all $n-1$, but 1, bits of $\mathbf{x}_0$ are known to us. As is customary in the field, we refer to the bit error probability of this case as the *matched filter bound* MFB and denote it by P_e^{MFB}. The MFB corresponds to the probability of error in detecting (say) $\mathbf{x}_{0,n} \in \{\pm 1\}$ from: $\tilde{\mathbf{y}} = \mathbf{x}_{0,n}\mathbf{a}_n + \mathbf{z}$, where $\tilde{\mathbf{y}} = \mathbf{y} - \sum_{i=1}^{n-1} \mathbf{x}_{0,i}\mathbf{a}_i$ is assumed known, and, $\mathbf{a}_i$ denotes the i^{th} column of $\mathbf{A}$.

The ML estimate is just the sign of the projection of the vector $\tilde{\mathbf{y}}$ to the direction of $\mathbf{a}_n$. Without loss of generality assume that $\mathbf{x}_{0,n} = 1$. Then, the output of the matched

filter becomes $\text{sign}(\tilde{X})$, where $\tilde{X} = \|\mathbf{a}_n\|^2 + \sigma^2 \nu$, where $\nu \sim \mathcal{N}(0,1)$. When $n \to \infty$, $\|\mathbf{a}_n\|^2 \xrightarrow{P} \delta$[12] . Hence, with probability one,

$$\lim_{n\to\infty} P_e^{MFB} = \lim_{n\to\infty} \mathbb{P}(\tilde{X} < 0) = Q(\sqrt{\delta \cdot \text{SNR}}). \tag{2.60}$$

A direct comparison of (2.60) to (2.59) shows that at high-SNR, the performance of the BRO is $10 \log_{10} \frac{\delta}{\delta - 1/2}$ dB off that of the MFB. In particular, in the square case ($\delta = 1$), where the number of receive and transmit antennas are the same, the BRO is 3 dB off the MFB. When the number of receive antennas is much larger, i.e. when $\delta \to \infty$, then the performance of the BRO approaches the MFB.

Next, we would like to apply our main Theorem 1 to binary classification problems.

2.3 Binary Classification

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 covariance matrix approaching the inverse of the Fisher information [103, 183]. However, in most modern applications in data science, the datasets often have a huge number of features, and therefore, the assumption $\frac{n}{p} \gg 1$ is not valid. Sur and Candes [37, 157, 159] have recently studied the performance of the maximum likeliood 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 likelikehood 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 [30, 94, 182]). 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 orderwise 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.)

[12] We use $\xrightarrow{P}$ to denote convergence in probability with $n \to \infty$.

In this section, 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 [56, 66, 165], we precisely characterize the assymptotic 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 [157]. When the regularizer is quadratic in parameters, the 6 equations also simplifies to 3. When the regularizer is the ℓ_1 norm, which corresponds to the popular sparse logistic regression [98, 99], 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 our knowledge, this is the first work that precisely characterizes the performance of the regularized logistic regression in high dimensions.

Mathematical Setup

Consider the scenario when the observations y_i are binary in the following form,

$$y_i = \begin{cases} +1 & \text{w.p. } \rho(\mathbf{x}_i^\mathsf{T}\beta_0) \\ -1 & \text{w.p. } 1-\rho(\mathbf{x}_i^\mathsf{T}\beta_0) \end{cases} \tag{2.61}$$

where $\rho : \mathbb{R} \to [0,1]$. In practice, the function ρ is often unknown, but choices like Hyperbolic tangent or tangent inverse functions are made for recovery of the separating hyper plane β_0. In this case, the link function $\mathbf{g}$ is

$$g(\mathbf{x}) = \text{Sign}(\rho(\mathbf{x}) - \epsilon) \ , \quad , \quad \epsilon_i \sim \text{Unif}(0,1) \ .$$

where ϵ has a uniform distribution between 0 and 1. It's not hard to check that the output of the link function will be 1 with probability $\rho(\mathbf{x})$, and -1 with probability $1-\rho(\mathbf{x})$.

The performance of the convex estimator (2.2) has been analyzed for a wide variety of loss functions and regularizers.

Schur et al. [158] analyzed this optimization for the case of logistic loss function,

where there is no regularization function f. The logistic loss function is defined as

$$\mathcal{L}(\mathbf{X}\beta, \mathbf{y}) = \sum_{i=1}^{n} -log\left(e^{\mathbf{x}_i\beta} + e^{-\mathbf{x}_i\beta}\right) + y_i\ \mathbf{x}_i\beta\ . \tag{2.62}$$

It's not hard to see that this loss function is the maximum likelihood estimator of β_0 in model (2.61), where the function ρ is the Hyperbolic tangent function which is

$$\rho(x) = \frac{e^x}{e^x + e^{-x}}\ . \tag{2.63}$$

Salehi et al. [143], analyzed the same problem for the regularized case in their work using CGMT framework.

One interesting result that is not derived in these series of works is the performance analysis of Support Vector Machine and the Perceptron method. In support vector machine, the loss function for classification is define as

$$\mathcal{L}_{SVM}(\mathbf{X}\beta, \mathbf{y}) = \sum_{i=1}^{n} \max\left(0, 1 - y_i\ \mathbf{x}_i^{\mathsf{T}}\beta\right)\ . \tag{2.64}$$

For the Perceptron method, the loss function is defined as

$$\mathcal{L}_{Perceptron}(\mathbf{X}\beta, \mathbf{y}) = \sum_{i=1}^{n} \max\left(0, -y_i\ \mathbf{x}_i^{\mathsf{T}}\beta\right)\ . \tag{2.65}$$

Performance Analysis of Logistic Regression

Assume we have n samples from a logistic model with parameter $\beta^* \in \mathbb{R}^p$. Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ denote the set of samples (a.k.a. the training data), where for $i = 1, 2, \ldots, n$, $\mathbf{x}_i \in \mathbb{R}^p$ is the feature vector and the label $y_i \in \{0, 1\}$ is a Bernouli random variable with,

$$\mathbb{P}[y_i = 1|\mathbf{x}_i] = \rho'(\mathbf{x}_i^T\beta^*)\ ,\ \text{ for }\ i = 1, 2, \ldots, n\ , \tag{2.66}$$

where $\rho'(t) := \frac{e^t}{1+e^t}$ is the standard logistic function. The goal is to compute an estimate for β^* from the training data $\mathcal{D}$. The maximum likelihood estimator, $\hat{\beta}_{ML}$, is defined as,

$$\begin{aligned}\hat{\beta}_{ML} = \arg\max_{\beta\in\mathbb{R}^p} \prod_{i=1}^{n} \mathbb{P}_\beta(y_i|\mathbf{x}_i) &= \arg\max_{\beta\in\mathbb{R}^p} \prod_{i=1}^{n} \frac{e^{y_i(\mathbf{x}_i^T\beta)}}{1 + e^{\mathbf{x}_i^T\beta}} \\ &= \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} \rho(\mathbf{x}_i^T\beta) - y_i(\mathbf{x}_i^T\beta)\ .\end{aligned} \tag{2.67}$$

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 β.

In many interesting settings the underlying parameter possesses cerain 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,

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \cdot \Big[\sum_{i=1}^{n} \rho(\mathbf{x}_i^T \beta) - y_i (\mathbf{x}_i^T \beta) \Big] + \frac{\lambda}{p} f(\beta) \ . \tag{2.68}$$

Here, $\lambda \in \mathbb{R}_+$ is the regularization parameter that must be tuned properly. In this section, we study the linear asymptotic regime in which the problem dimensions p, n grow to infinity at a proportional rate, $\delta := \frac{n}{p} > 0$. Our main result characterizes the performance of $\hat{\beta}$ in terms of the ratio, δ, and the signal strength, $\kappa = \frac{||\beta^*||}{\sqrt{p}}$. 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, $\{\mathbf{x}_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \frac{1}{p}\mathbf{I}_p)$. We further assume that the entries of β^* 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.

As we will see in Theorem 4, 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 $(\alpha, \sigma, \gamma, \theta, \tau, r)$.

$$
\begin{cases}
\kappa^2\alpha = \mathbb{E}\big[\beta \operatorname{Prox}_{\lambda\sigma\tau\tilde{f}(\cdot)}\big(\sigma\tau(\theta\beta + \frac{r}{\sqrt{\delta}}Z)\big)\big] , \\
\gamma = \frac{1}{r\sqrt{\delta}} \mathbb{E}\big[Z \operatorname{Prox}_{\lambda\sigma\tau\tilde{f}(\cdot)}\big(\sigma\tau(\theta\beta + \frac{r}{\sqrt{\delta}}Z)\big)\big] , \\
\kappa^2\alpha^2 + \sigma^2 = \mathbb{E}\left[\operatorname{Prox}_{\lambda\sigma\tau\tilde{f}(\cdot)}\big(\sigma\tau(\theta\beta + \frac{r}{\sqrt{\delta}}Z)\big)^2\right] , \\
\gamma^2 = \frac{2}{r^2} \mathbb{E}\big[\rho'(-\kappa Z_1)\big(\kappa\alpha Z_1 + \sigma Z_2 - \operatorname{Prox}_{\gamma\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2)\big)^2\big] , \\
\theta\gamma = -2\,\mathbb{E}\big[\rho''(-\kappa Z_1)\operatorname{Prox}_{\gamma\rho(\cdot)}\big(\kappa\alpha Z_1 + \sigma Z_2\big)\big] , \\
1 - \frac{\gamma}{\sigma\tau} = \mathbb{E}\big[\frac{2\rho'(-\kappa Z_1)}{1 + \gamma\rho''\big(\operatorname{Prox}_{\gamma\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2)\big)}\big] .
\end{cases}
\tag{2.69}
$$

Here Z, Z_1, Z_2 are standard normal variables, and $\beta \sim \Pi$, where Π denotes the distribution on the entries of β^*. The following remarks provide some insights on solving the nonlinear system.

Remark 16 (Proximal Operators) *It is worth noting that the equations in* (2.69) *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 17 (Numerical Evaluation) *Define $\mathbf{v} := [\alpha, \sigma, \gamma, \theta, \tau, r]^T$ as the vector of unknonws. The nonlinear system* (2.69) *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 $\mathbf{v}^* = S(\mathbf{v}^*)$.*

We are now able to present our main result. Theorem 4 below describes the average behavior of the entries of $\hat{\beta}$, the solution of the RLR. The derived expression is in terms of the solution of the nonlinear system (2.69), 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 $\hat{\beta}$ converge as follows,

$$
\hat{\beta}_j \xrightarrow{d} \Gamma(\beta_j^*, Z) , \quad \text{for } j = 1, 2, \ldots, p , \tag{2.70}
$$

where Z is a standard normal random variable, and $\Gamma : \mathbb{R}^2 \to \mathbb{R}$ is defined as,

$$\Gamma(c, d) := \text{Prox}_{\lambda\bar{\sigma}\bar{\tau}\tilde{f}(\cdot)}\big(\bar{\sigma}\bar{\tau}(\bar{\theta}c + \frac{\bar{r}}{\sqrt{\delta}}d)\big) \ . \tag{2.71}$$

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 4 *Consider the optimization program* (2.68)*, where for* $i = 1, 2, \ldots, n$*,* $\mathbf{x}_i$ *has the multivariate Gaussian distribution* $\mathcal{N}(0, \frac{1}{p}\mathbf{I}_p)$*, and* $y_i = Ber(\mathbf{x}_i^T\beta^*)$*, and the entries of* β^* *are drawn independently from a distribution* Π*. Assume the parameters* δ*,* κ*, and* λ *are such that the nonlinear system* (2.69) *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*[13] *function* $\Psi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ *, we have,*

$$\frac{1}{p}\sum_{j=1}^{p} \Psi(\hat{\beta}_j, \beta_j^*) \xrightarrow{P} \mathbb{E}\big[\Psi\big(\Gamma(\beta, Z), \beta\big)\big] \ , \tag{2.72}$$

where $Z \sim \mathcal{N}(0, 1)$*,* $\beta \sim \Pi$ *is independent of* Z*, and the function* $\Gamma(\cdot, \cdot)$ *is defined in* (2.71)*.*

Correlation and variance of the RLR estimate

As the first application of Theorem 4 we compute common descriptive statistics of the estimate $\hat{\beta}$. In the following corollaries, we establish that the parametrs $\bar{\alpha}$, and $\bar{\sigma}$ in (2.69) correspond to the correlation and the mean-squared error of the resulting estimate.

Corollary 1 *As* $p \to \infty$*,* $\frac{1}{||\beta^*||^2}\,\hat{\beta}^T\beta^* \xrightarrow{P} \bar{\alpha}$ *.*

Proof 1 *Recall that* $||\beta^*||^2 = p\kappa^2$*. Applying Theorem 4 with* $\Psi(u, v) = uv$ *gives,*

$$\frac{1}{||\beta^*||^2}\hat{\beta}^T\beta^* = \frac{1}{\kappa^2 p}\sum_{j=1}^{p}\hat{\beta}_j\beta_j^* \xrightarrow{P} \frac{1}{\kappa^2}\mathbb{E}\big[\beta\ \textit{Prox}_{\lambda\bar{\sigma}\bar{\tau}\tilde{f}(\cdot)}\big(\bar{\sigma}\bar{\tau}(\bar{\theta}\beta + \frac{\bar{r}}{\sqrt{\delta}}Z)\big)\big] = \bar{\alpha} \ , \tag{2.73}$$

where the last equality is derived from the first equation in the nonlinear system (2.69)*, along with the fact that* $(\bar{\alpha}, \bar{\sigma}, \bar{\gamma}, \bar{\theta}, \bar{\tau}, \bar{r})$ *is a solution to this system.*

[13] A function $\Phi : \mathbb{R}^d \to \mathbb{R}$ is said to be *locally-Lipschitz* if,

$$\forall M > 0,\ \exists L_M \geq 0,\ \text{such that}\ \ \forall \mathbf{x}, \mathbf{y} \in \big[-M, +M\big]^d \ : \quad |\Phi(\mathbf{x}) - \Phi(\mathbf{y})| \leq L_M||\mathbf{x} - \mathbf{y}|| \ .$$

Corollary 1 states that upon centering $\hat{\beta}$ around $\bar{\alpha}\beta^*$, it becomes decorrelated from β^*. Therefore, we define a new estimate $\tilde{\beta} := \frac{\hat{\beta}}{\bar{\alpha}}$ and compute its mean-squared error in the following corollary.

Corollary 2 *As* $p \to \infty$, $\frac{1}{p}||\tilde{\beta} - \beta^*||^2 \xrightarrow{P} \frac{\bar{\sigma}^2}{\bar{\alpha}^2}$.

Proof 2 *We appeal to Theorem 4 with* $\Psi(u, v) = (u - \bar{\alpha}v)^2$,

$$\frac{1}{p}||\tilde{\beta}-\beta^*||^2 = \frac{1}{\bar{\alpha}^2}\big(\frac{1}{p}||\hat{\beta}-\bar{\alpha}\beta^*||^2\big) \xrightarrow{P} \frac{1}{\bar{\alpha}^2}\mathbb{E}\Big[\big(Prox_{\lambda\bar{\sigma}\bar{\tau}\tilde{f}(\cdot)}\big(\bar{\sigma}\bar{\tau}(\bar{\theta}\beta+\frac{\bar{r}}{\sqrt{\delta}}Z)\big)-\bar{\alpha}\beta\big)^2\Big] = \frac{\bar{\sigma}^2}{\bar{\alpha}^2}\,, \tag{2.74}$$

where the last equality is derived from the third equation in the nonlinear system (2.69) *together with the result of Corollary 1.*

In the next two sections, we investigate other properties of the estimate $\hat{\beta}$ under ℓ_1 and ℓ_2 regularization.

RLR with ℓ_2^2-regularization

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.,

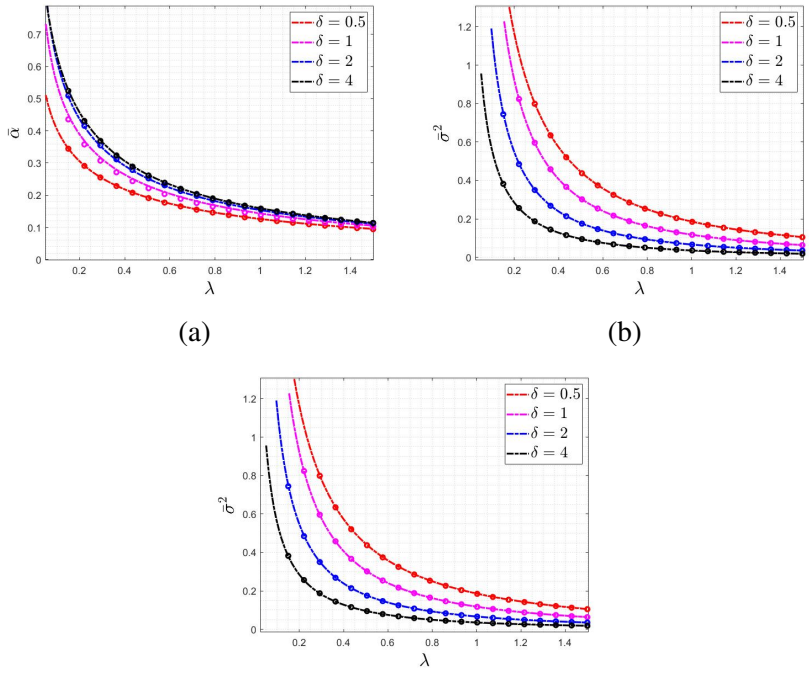
$$\hat{\beta} = \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\cdot\Big[\sum_{i=1}^{n}\rho(\mathbf{x}_i^T\beta) - y_i(\mathbf{x}_i^T\beta)\Big] + \frac{\lambda}{2p}\sum_{i=1}^{p}\beta_i^2\,. \tag{2.75}$$

To analyze (2.75), we use the result of Theorem 4. It can be shown that in the nonlinear system (2.69), $\bar{\theta}$, $\bar{\tau}$, $\bar{r}$ can be derived explicitely 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,

$$\text{Prox}_{t\tilde{f}(\cdot)}(x) = \arg\min_{y\in\mathbb{R}} \frac{1}{2t}(y-x)^2 + \frac{1}{2}y^2 = \frac{x}{1+t}\,. \tag{2.76}$$

This indicates that the proximal operator in this case is just a simple rescaling. Substituting (2.76) in the nonlinear system (2.69), we can rewrite the first three equations as follows,

$$\begin{cases} \theta = \dfrac{\alpha}{\gamma\delta}\,, \\ \tau = \dfrac{\delta\gamma}{\sigma\big(1-\lambda\delta\gamma\big)}\,, \\ r = \dfrac{\sigma}{\gamma\sqrt{\delta}}\,. \end{cases} \tag{2.77}$$



(a) (b)

(c)

Figure 2.7: The performance of the regularized logistic regression under ℓ_2^2 penalty (a) the correlation factor $\bar{\alpha}$ (b) the variance $\bar{\sigma}^2$, and (c) the mean-squared error $\frac{1}{p}||\hat{\beta} - \beta^*||^2$. The dashed lines depict the theoretical result derived from Theorem 5, and the dots are the result of empirical simulations. The empirical results is the average over 100 independent trials with $p = 250$ and $\kappa = 1$.

Therefore we can state the following Theorem for ℓ_2^2-regularization:

Theorem 5 *Consider the optimization* (2.75) *with parameters* κ*,* δ*, and* γ*, and the same assumptions as in Theorem 4. As* $p \to \infty$*, for any locally-Lipschitz function* $\Psi(\cdot,\cdot)$*, the following convergence holds,*

$$\frac{1}{p}\sum_{j=1}^{p} \Psi(\hat{\beta}_j - \bar{\alpha}\beta_j^*, \beta_j^*) \xrightarrow{P} \mathbb{E}\big[\Psi\big(\bar{\sigma}Z, \beta\big)\big] \,, \tag{2.78}$$

where Z *is standard normal,* $\beta \sim \Pi$*, and* $\bar{\alpha}$*,*$\bar{\sigma}$ *are the unique solutions to the following nonlinear system of equations,*

$$\begin{cases} \dfrac{\sigma^2}{2\delta} = \mathbb{E}\big[\rho'(-\kappa Z_1)\big(\kappa\alpha Z_1 + \sigma Z_2 - Prox_{\gamma\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2)\big)^2\big] \,, \\ -\dfrac{\alpha}{2\delta} = \mathbb{E}\big[\rho''(-\kappa Z_1) Prox_{\gamma\rho(\cdot)}\big(\kappa\alpha Z_1 + \sigma Z_2\big)\big] \,, \\ 1 - \dfrac{1}{\delta} + \lambda\gamma = \mathbb{E}\Big[\dfrac{2\rho'(-\kappa Z_1)}{1 + \gamma\rho''\big(Prox_{\gamma\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2)\big)}\Big] \,. \end{cases} \tag{2.79}$$

The proof is deferred to the Appendix. Theorem 5 states that upon centering the estimate $\hat{\beta}$, it becomes decorrelated from β^* and the distribution of the entries approach a zero-mean Gaussian distribution with variance $\bar{\sigma}^2$.

Figure 2.7 depicts the performance of the regularized estimate for different values of λ. As observed in the figure, increasing the value of λ reduces the correlation factor $\bar{\alpha}$ (Figure 2.7a) and the variance $\bar{\sigma}^2$ (Figure 2.7b). Figure 2.7c shows the mean-squared-error of the estimate as a function of λ . It indicates that for different values of δ there exist an optimal value λ_{opt} that achieves the minimum mean-squared error.

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 [89]. 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 4, 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. In Section 2.3, 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,

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \ \frac{1}{n} \cdot \Big[\sum_{i=1}^{n} \rho(\mathbf{x}_i^T \beta) - y_i (\mathbf{x}_i^T \beta) \Big] + \frac{\lambda}{p} ||\beta||_1 \ . \tag{2.80}$$

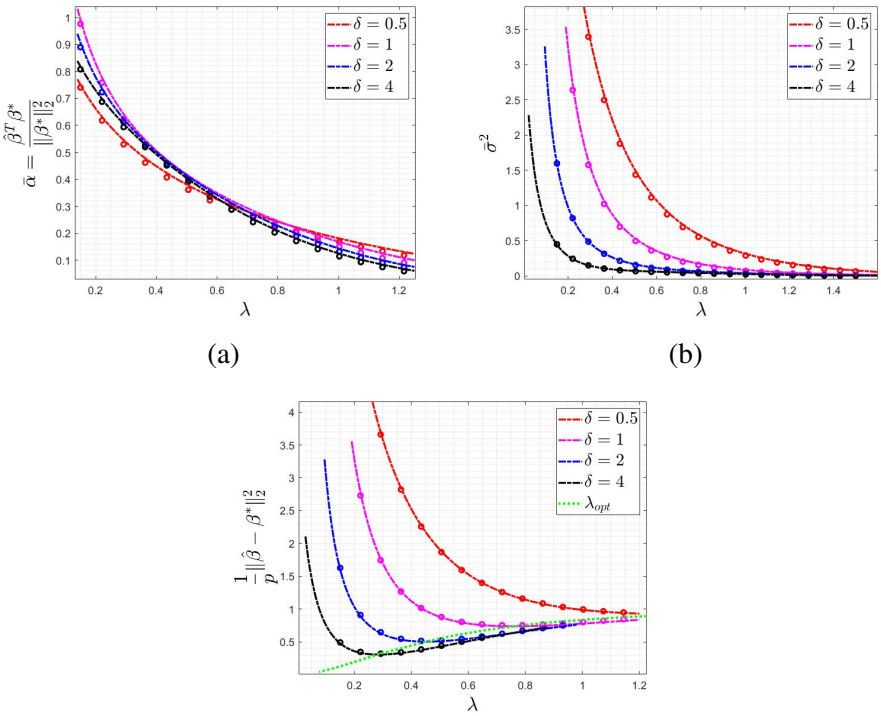
In Section 2.3, we explicitly describe the expectations in the nonlinear system (2.69) using two q-functions[14]. In Section 2.3, we analyze the support recovery in the resulting estimate and show that the two q-functions represent the probability of on and off support recovery.

Convergence behavior of sparse LR

For our analysis in this section, we assume each entry β_i^*, for $i = 1, \ldots, p$, is sampled i.i.d. from a distribution,

$$\Pi(\beta) = (1 - s) \cdot \delta_0(\beta) + s \cdot \Big(\frac{\phi(\frac{\beta}{\frac{\kappa}{\sqrt{s}}})}{\frac{\kappa}{\sqrt{s}}} \Big), \tag{2.81}$$

[14]The q-function is the tail distribution of the standard normal r.v. defined as, $Q(t) := \int_t^\infty \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx$.



(a) (b)

(c)

Figure 2.8: The performance of the regularized logistic regression under ℓ_1 penalty (a) the correlation factor $\bar{\alpha}$ (b) the variance $\bar{\sigma}^2$, and (c) the mean-squared error $\frac{1}{p}||\hat{\beta} - \beta^*||^2$. The dashed lines are the theoretical result derived from Theorem 4, and the dots are the result of empirical simulations. For the numerical simulations, the result is the average over 100 independent trials with $p = 250$ and $\kappa = 1$.

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 β^* 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 (2.69). To streamline the representation, we define the following two proxies,

$$t_1 = \frac{\lambda}{\sqrt{\frac{r^2}{\delta} + \frac{\theta^2\kappa^2}{s}}}, \quad t_2 = \frac{\lambda}{\frac{r}{\sqrt{\delta}}}. \tag{2.82}$$

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. We can rewrite the first three equations in (2.69) as follows,

$$\begin{cases} \dfrac{\alpha}{2\sigma\tau} = \theta \cdot Q(t_1) \ , \\ \dfrac{\delta\gamma}{2\sigma\tau} = s \cdot Q\big(t_1\big) + (1-s) \cdot Q\big(t_2\big) \ , \\ \dfrac{\kappa^2\alpha^2 + \sigma^2}{2\sigma^2\tau^2} = \dfrac{\delta\gamma\lambda^2}{2\sigma\tau} + \dfrac{\gamma r^2}{2\sigma\tau} + \kappa^2\theta^2 \cdot Q\big(t_1\big) - \lambda^2(s \cdot \dfrac{\phi(t_1)}{t1} + (1-s) \cdot \dfrac{\phi(t_2)}{t_2}) \ . \end{cases} \tag{2.83}$$

Appending the three equations in (2.83) to the last three equations in (2.69) gives the nonlinear system for sparse LR. Upon solving these equations, we can use the result of Theorem 4 to compute various performance measures on the estimate $\hat{\beta}$. Figure 2.8 shows the performance of our estimate as a function of λ. It can be seen that the bound derived from our theoretical result matches the empirical simulations. Also, it can be inferrred from Figure 2.8c that the optimal value of λ (λ_{opt} that achieves the minimum mean-squared error) is a decreasing function of δ.

Support recovery

In this section, we study the support recovery in sparse LR. As mentioned earlier, sparse LR is often used when the underlying paramter has few non-zero entries. We define the support of β^* as $\Omega := \{j | 1 \leq j \leq p, \beta^*_j \neq 0\}$. Here, we would like to compute the probability of success in recovery of the support of β^*.

Let $\hat{\beta}$ denote the solution of the optimization (2.80). 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 $\hat{\beta}$,

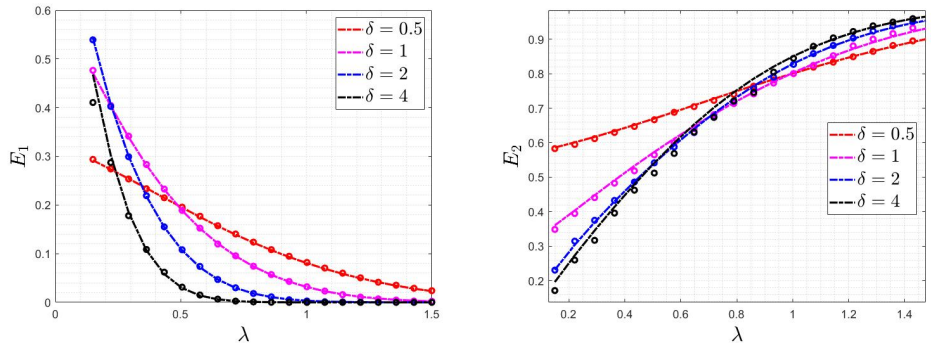
$$\hat{\Omega} = \{j | 1 \leq j \leq p, |\hat{\beta}_j| > \epsilon\} \tag{2.84}$$

In order to evaluate the success in support recovery, we define the following two error measures,

$$E_1(\epsilon) = \text{Prob}\{j \in \hat{\Omega} | j \notin \Omega\} \ \ , \ \ E_2(\epsilon) = \text{Prob}\{j \notin \hat{\Omega} | j \in \Omega\} \ . \tag{2.85}$$

In our estimation, E_1 represents the probability of false alarm, and E_2 is the probability of misdetection of an entry of the support. The following lemma indicates the asymptotic behavior of both errors as ϵ approaches zero .

Lemma 5 (Support Recovery) *Let $\hat{\beta}$ be the solution to the optimization* (2.80)*, and the entries of β^* have distribution* Π *defined in* (2.81)*. Assume λ is chosen such that the nonlinear system* (2.69) *has a unique solution* $(\bar{\alpha}, \bar{\sigma}, \bar{\gamma}, \bar{\theta}, \bar{\tau}, \bar{r})$*. As $p \to \infty$ we*



(a) (b)

Figure 2.9: The support recovery in the regularized logistic regression with ℓ_1 penalty for (a) E_1: the probability of false detection, (b) E_2: the probability of missing an entry of the support. The dashed lines are the theoretical results derived from Lemma 5, and the dots are the result of empirical simulations. For the numerical simulations, the result is the average over 100 independent trials with $p = 250$ and $\kappa = 1$ and $\epsilon = 0.001$.

have,

$$\begin{aligned} &\lim_{\epsilon \downarrow 0} E_1(\epsilon) \xrightarrow{P} 2\,Q\big(\bar{t}_1\big) \;\; \textit{where,} \;\; \bar{t}_1 = \frac{\lambda}{\frac{\bar{r}}{\sqrt{\delta}}}, \;\; \textit{and,} \\ &\lim_{\epsilon \downarrow 0} E_2(\epsilon) \xrightarrow{P} 1 - 2\,Q\big(\bar{t}_2\big) \;\; \textit{where,} \;\; \bar{t}_2 = \frac{\lambda}{\sqrt{\frac{\bar{r}^2}{\delta} + \frac{\bar{\theta}^2\kappa^2}{s}}} \,. \end{aligned} \tag{2.86}$$

2.4 Generalized Margin Maximizers

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 diffrent classifiers for binary classification dates back to the seminal work of Vapnik in the 1980's [184]. 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 datapoints 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 [46].

In this section, we challenge the conventional wisdom by showing that when the underlying parameter has certain structure one can come up with classifiers that outperform the max-margin classifier. We introduce the **G**eneralized **M**argin **M**aximizer (**GMM**) which takes into account the structure of the underlying parameter as well as the minimimum distance of the datapoints to the separating hyperplane. We provide sharp asymptotic results on various performance measures (such as the generalization error) of GMM 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 [16, 21, 84, 114, 197]. Most of these studies focus on the least-squares/ridge regression, where the loss function is the squared ℓ_2-norm, and derive sharp asymptotics on the performance of the estimator. In particular, in [84, 97] the authors have shown that the minimum-norm least square solution demonstrates the so-called "double-descent" behavior [20].

A more recent line of research studies the generalization performance of gradient descent (GD) for binary classification. It has been shown [151]) that for a separable dataset, GD (when applied on the logistic loss) converges in 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 [51, 118].

State of the Art

We analyze the performance of GMM 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 $\delta := \frac{p}{n} > 0$. First, we provide the phase transition condition for the separability of data (i.e., derive the exact value of δ^* such that the data is separable for all $\delta > \delta^*$[15].) Consequently, we analyze the performance in the interpolating regime ($\delta > \delta^*$). To the best of our knowledge, this is the first theoretical result that provides sharp asymptotics on the performance of GMM classifiers on separable data. For our analysis, we exploit the **C**onvex **G**aussian **M**in-max **T**heorem (**CGMT**) [155, 171] which is a strengthened version of a classical Gaussian comparison inequality due

[15]Concurrent to the submission of this paper, a similar phase transition has been demonstrated in [97] for a somewhat different model.

to Gordon [79]. This framework replaces the original optimization with another optimization problem that has a similar performance, yet is much simpler to analyze as it becomes nearly separable. Previously, the CGMT has been successfully applied to derive the precise performance in a number of applications such as regularized M-estimators [165], analysis of the generalized lasso [117, 171], data detection in massive MIMO [1, 9, 175], and PhaseMax in phase retrieval [54, 141, 142].
More recently, this framework has been employed in a series of works by multiple groups of researchers to characterize the performance of the logistic loss minimizer in binary classification [143, 161]. 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 [51, 118].

Mathematical 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 dataset $\mathcal{D}$ is generated from a logistic-type model with the underlying parameter $\mathbf{w}^\star \in \mathbb{R}^p$. This means that

$$y_i \sim \textsc{Rad}(\rho(\mathbf{x}_i^T \mathbf{w}^\star)) \,, \quad i = 1, \dots, n \,, \tag{2.87}$$

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 [157] 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 [143], 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 learn-

ing models, in this section, 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.,

$$\exists \mathbf{w}_0 \;\text{ s.t. }\; \textsc{Sign}(\mathbf{w}_0^T \mathbf{x}_i) = y_i, \;\; \text{for } i = 1, 2, \ldots, n. \tag{2.88}$$

Let $\mathcal{W}$ denote the set of all the parameters that interpolate the data.

$$\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^p : \textsc{Sign}(\mathbf{w}^T \mathbf{x}_i) = y_i \;, \;\; \text{for } 1 \leq i \leq n.\}. \tag{2.89}$$

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 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 *Generalized Margin Maximizer* (GMM) as the solution to the following optimization:

$$\begin{aligned} \min_{\mathbf{w} \in \mathbb{R}^p} \quad & \psi(\mathbf{w}) \\ \text{s.t. } & y_i(\mathbf{x}_i^T \mathbf{w}) \geq 1, \;\; \text{for } 1 \leq i \leq n. \end{aligned} \tag{2.90}$$

It is worth noting that the condition on the separability of the dataset is crucial for the optimization program (2.90) to have a feasible point.

Remark 18 *It can be shown that when $\psi(\cdot)$ is absolutely scalable*[16]*, the GMM can be found by solving the following equivalent optimization program,*

$$\max_{\mathbf{w} \in \mathbb{R}^d} \frac{\psi(\mathbf{w})}{\min\limits_{1 \leq i \leq n} y_i(\mathbf{x}_i^T \mathbf{w})} = \max_{\mathbf{w} \in \mathbb{R}^d} \frac{\|\mathbf{w}\|}{\min\limits_{1 \leq i \leq n} y_i(\mathbf{x}_i^T \mathbf{w})} \times \frac{\psi(\mathbf{w})}{\|\mathbf{w}\|}. \tag{2.91}$$

The first multiplicative term on the right indicates the margin associated with the separator $\mathbf{w}$*, and the second term,* $\frac{\psi(\mathbf{w})}{\|\mathbf{w}\|}$ *takes into account the structure of the*

[16] 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.

model. Hence, we refer to the objective function in the optimization (2.91) *as the generalized margin, and the solution to this optimization is called the generalized margin maximizer (GMM).*

In this section, 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 (2.90), $\hat{\mathbf{w}}$, in terms of the ratio, δ, and the signal strength, $\kappa := \frac{\|\mathbf{w}^\star\|}{\sqrt{p}}$. We assume that the datapoints, $\{\mathbf{x}_i\}_{i=1}^n$, 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.

Main Results

In this section, we present the main results of the paper, that is the characterization of the performance of the generalized margin maximizers. Our results are represented in terms of a summary functional, $c_t(\cdot,\cdot)$, which incorporates the information about the underlying model.

Definition 1 *For the parameter $t > 0$, the function $c_t : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}_+$ is defined as,*

$$c_t(s,r) = \mathbb{E}\big[(1 - tsZ_1Y - rZ_2)_+^2\big], \tag{2.92}$$

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

Asymptotic phase transition

Here, we provide the necessary and sufficient condition for the separability of the data.

Theorem 6 (Phase transition) *Consider the generalized max margin optimization defined in Section 2.4. As $n, p \to \infty$ at a fixed overparameterization ratio $\delta := \frac{p}{n} \in (0,\infty)$, this optimization program (almost surely) has a solution (or equivalenty, the set $\mathcal{W}$ is nonempty) if and only if,*

$$\delta > \delta^* = \delta^*(\kappa) := \inf_{s,r\geq 0} \frac{c_\kappa(s,r)}{r^2}\,. \tag{2.93}$$

Remark 19 *Theorem 6 indicates the necessary and sufficient condition for the existense of GMM. It is worth mentioning that this condition, which is simply the*

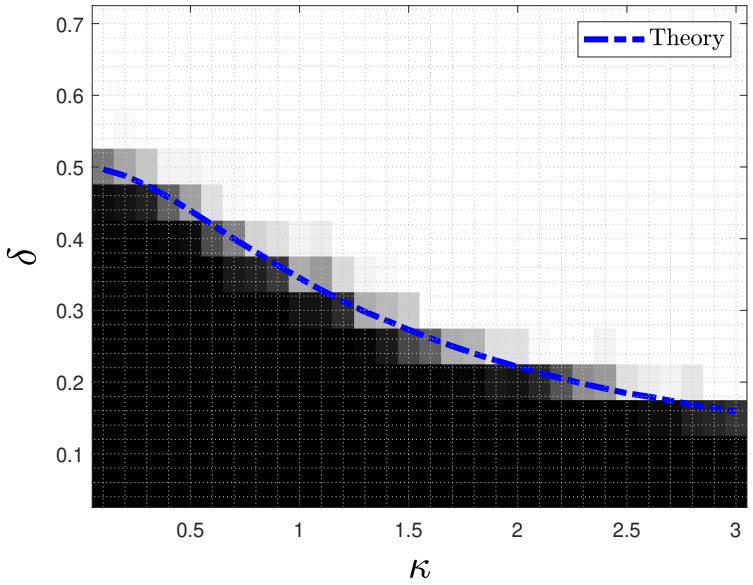


Figure 2.10: The phase transition, δ^*, for the separability of the dataset, where the feature vector, $\mathbf{x}_i$ is drawn from the Gaussian distribution, $\mathcal{N}(\mathbf{0}, \frac{1}{p}\mathbf{I}_p)$, and the labels are $y_i \sim \textsc{Rad}\big(\rho(\mathbf{x}_i^T\mathbf{w}^\star)\big)$, for $\rho(z) = \frac{e^t}{e^t+e^{-t}}$. The empirical result is the average over 20 trials with $p = 150$, and the theoretical results are from Theorem 6.

condition on separability of the dataset $\mathcal{D}$, does not depend on the choice of the potential function $\psi(\cdot)$.

Remark 20 *The phase transition* (2.93)*, is valid for any link function $\rho(\cdot)$. This generalizes the former results in [37]. Note that the summary functional, $c_\kappa(\cdot,\cdot)$, contains the choice of the link function and can be computed numerically.*

The following lemma explains the behavior of δ^* as κ varies.

Lemma 6 *δ^* is a decreasing function of κ, with $\delta^*(0) = \frac{1}{2}$ and $\lim_{\kappa\to+\infty}\delta^*(\kappa) = 0$.*

The result of Lemma 6 can be intuitively verified. Recall that $\kappa = \frac{\|\mathbf{w}^\star\|}{\sqrt{p}}$ and $y_i \sim \textsc{Rad}(\rho(\mathbf{x}_i^T\mathbf{w}^\star))$. Therefore, $\kappa \to \infty$ translates to having $y_i = \textsc{Sign}(\mathbf{x}_i^T\mathbf{w}^\star)$. 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 $\mathbf{x}_i$. [47] 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 2.10 provides a comparison between the theoretical result in Theorem 6, and the empirical results derived from numerical simulations for $p = 150$ and 20 trials. As seen in this plot, the theory matches well with the empirical simulations.

A nonlinear system of equations

Our main result in Section 2.4 precisely characterizes the performance of GMM in terms of a system of 5 nonlinear equations with 5 unknowns, $(\alpha, \sigma, \beta, \gamma, \tau)$, defined as follows,

$$\begin{cases} \frac{1}{p}\mathbb{E}\big[\mathbf{w}^{\star T}\mathrm{Prox}_{\sigma\tau\psi(\cdot)}\big((\alpha-\sigma\tau\gamma)\mathbf{w}^\star+\beta\sigma\tau\sqrt{\delta}\mathbf{h}\big)\big] = \alpha\kappa^2, \\ \frac{1}{p}\mathbb{E}\big[\mathbf{h}^{T}\mathrm{Prox}_{\sigma\tau\psi(\cdot)}\big((\alpha-\sigma\tau\gamma)\mathbf{w}^\star+\beta\sigma\tau\sqrt{\delta}\mathbf{h}\big)\big] = \sqrt{\frac{c_\kappa(\alpha,\sigma)}{\delta}}, \\ \frac{1}{p}\mathbb{E}\left\|\mathrm{Prox}_{\sigma\tau\psi(\cdot)}\big((\alpha-\sigma\tau\gamma)\mathbf{w}^\star+\beta\sigma\tau\sqrt{\delta}\mathbf{h}\big)\right\|^2 = \alpha^2\kappa^2+\sigma^2, \\ \frac{\partial c_\kappa(\alpha,\sigma)}{\partial\alpha} = \frac{2\kappa^2\gamma}{\beta}\sqrt{c_\kappa(\alpha,\sigma)}, \\ \frac{\partial c_\kappa(\alpha,\sigma)}{\partial\sigma} = \frac{2\sqrt{c_\kappa(\alpha,\sigma)}}{\beta\tau}. \end{cases} \tag{2.94}$$

Remark 21 *The first three equations in the nonlinear system* (2.94) *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 [132].*

Asymptotic performance of GMM

We are now ready to present the main result of the paper. Theorem 7 characterizes the asymptotic behavior of GMM, that is the solution to the optimization program (2.90). It connects the performance of GMM to the solution of the nonlinear system of equations (2.94), and informally states that,

$$\hat{\mathbf{w}} \xrightarrow{D} \Gamma(\mathbf{w}^\star, \mathbf{h}), \text{ as } p \to \infty, \tag{2.95}$$

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,

$$\Gamma(\mathbf{v}_1, \mathbf{v}_2) = \mathrm{Prox}_{\bar{\sigma}\bar{\tau}\psi(\cdot)}\big((\bar{\alpha}-\bar{\sigma}\bar{\tau}\bar{\gamma})\mathbf{v}_1+\bar{\beta}\bar{\sigma}\bar{\tau}\sqrt{\delta}\mathbf{v}_2\big), \tag{2.96}$$

where $(\bar{\alpha}, \bar{\sigma}, \bar{\beta}, \bar{\gamma}, \bar{\tau})$ is the solution to the nonlinear system (2.94).

Theorem 7 *Let* $\hat{\mathbf{w}}$ *be the solution of the GMM optimization* (2.90)*, 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* $y_i \sim \textsc{Rad}(\rho(\mathbf{x}_i^T \mathbf{w}^\star))$*, 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* $\delta = \frac{p}{n} > \delta^*(\kappa)$*, the nonlinear system* (2.94) *has a unique solution* $(\bar{\alpha}, \bar{\sigma}, \bar{\beta}, \bar{\gamma}, \bar{\tau})$*. Furthermore, for any locally-Lipschitz function* $F : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$*, we have,*

$$F(\hat{\mathbf{w}}, \mathbf{w}^\star) \xrightarrow{P} \mathbb{E}[F(\Gamma(\mathbf{w}, \mathbf{h}), \mathbf{w})], \tag{2.97}$$

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* (2.96)*.*

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 (2.94) corresponds to its optimality condition.

Remark 22 *The result in Theorem 7 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 7 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 GMM classifier. We do so, by appealing to the result of Theorem 7.

Definition 2 *The generalization error for a binary classifier with parameter* $\hat{\mathbf{w}}$ *is defined as,*

$$GE_{\hat{\mathbf{w}}} = \mathbb{P}_{\mathbf{x}}\{\textsc{Sign}(\mathbf{x}^T \hat{\mathbf{w}}) \neq \textsc{Sign}(\mathbf{x}^T \mathbf{w}^\star)\}, \tag{2.98}$$

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 lemma provides sharp asymptotics on the generalization error of the GMM classifier.

Lemma 7 (Generalization Error) *Let $\hat{\mathbf{w}}$ be the GMM classifier defined in Section 2.4. 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,*

$$GE_{\hat{\mathbf{w}}} \xrightarrow{P} \frac{1}{\pi} acos\Big(\frac{\kappa\bar{\alpha}}{\sqrt{\kappa^2\bar{\alpha}^2 + \bar{\sigma}^2}}\Big), \tag{2.99}$$

where $\bar{\alpha}$ and $\bar{\sigma}$ are derived by solving the nonlinear system (2.94)*.*

Proof 3 *We first note that when the data is normally distributed, the generalization error for $\hat{\mathbf{w}}$ is defined as,*

$$GE_{\hat{\mathbf{w}}} = \frac{1}{\pi} acos\Big(\frac{\hat{\mathbf{w}}^T\mathbf{w}^\star}{\|\mathbf{w}^\star\|\,\|\hat{\mathbf{w}}\|}\Big). \tag{2.100}$$

We appeal to the result of Theorem 7 with two different functions. Using $F_1(\mathbf{u}, \mathbf{v}) = \frac{1}{p}\mathbf{v}^T\mathbf{u}$ in (2.97) *will give,*

$$\frac{1}{p}\hat{\mathbf{w}}^T\mathbf{w}^\star \xrightarrow{P} \frac{1}{p}\mathbb{E}\Big[\mathbf{w}^{\star T} Prox_{\bar{\sigma}\bar{\tau}\psi(\cdot)}\big((\bar{\alpha} - \bar{\sigma}\bar{\tau}\bar{\gamma})\mathbf{w}^\star + \bar{\beta}\bar{\sigma}\bar{\tau}\sqrt{\delta}\mathbf{h}\big)\Big]. \tag{2.101}$$

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 (2.94)*,which gives the following,*

$$\frac{1}{p}\hat{\mathbf{w}}^T\mathbf{w}^\star \xrightarrow{P} \kappa^2\bar{\alpha}. \tag{2.102}$$

Similarly, using the result of Theorem 7 for the measure function $F_2(\mathbf{u}, \mathbf{v}) = \frac{1}{p}\|\mathbf{u}\|^2$, along with the third equation in (2.94) *gives,*

$$\frac{1}{\sqrt{p}}\|\hat{\mathbf{w}}\| \xrightarrow{P} \sqrt{\kappa^2\bar{\alpha}^2 + \bar{\sigma}^2}\,. \tag{2.103}$$

The proof is the consequence of (2.100)*,* (2.102)*, and* (2.103)*, along with the continuity of the function $acos(\cdot)$.*

GMM 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 GMM classifier for some common structures and the corresponding choices of the potential function.

Max-margin classifier (ℓ_2-GMM)

The ℓ_2-norm regularization is commonly used in machine learning applications to stabilize the model. Here, we study the performance of the GMM classifier when $\psi(\cdot) = \frac{1}{2}\|\cdot\|_2^2$, i.e., the solution to the following optimization program,

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^p} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 \\ \text{s.t. } & y_i(\mathbf{x}_i^T\mathbf{w}) \geq 1, \quad \text{for } 1 \leq i \leq n. \end{aligned} \tag{2.104}$$

The optimization program (2.104) is called the hard-margin SVM and the corresponding solution is the max-margin classifier, as it maximizes the minimum distance (margin) of the datapoints from the separating hyperplane. 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-GMM (2.104), has been earlier analyzed in [51] and [118]. The form we present below in (2.106), differes in appearance to the results of [51], but can be shown to be equivalent.

When $\psi(\cdot) = \frac{1}{2}\|\cdot\|_2^2$, the proximal operator has the following closed-form,

$$\text{Prox}_{\frac{t}{2}\|\cdot\|^2}(\mathbf{u}) = \frac{1}{1+t}\mathbf{u}. \tag{2.105}$$

By replacing the proximal operator in the nonlinear system (2.94), 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,

$$\begin{cases} \sqrt{c_\kappa(\alpha,\sigma)} = \sigma\sqrt{\delta}, \\ \dfrac{\partial c_\kappa(\alpha,\sigma)}{\partial\alpha} = \dfrac{-2\kappa^2\alpha\tau\sigma\delta}{1+\sigma\tau}, \\ \dfrac{\partial c_\kappa(\alpha,\sigma)}{\partial\sigma} = \dfrac{2\sigma\delta}{1+\sigma\tau}. \end{cases} \tag{2.106}$$

Sparse classifier (ℓ_1-GMM)

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 [89]. Adding an ℓ_1 penalty is the most popular approach for feature selection.

As a natural consequence of our main result in Theorem 7, here we analyze the asymptotic performance of GMM 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.

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^p} \quad & \|\mathbf{w}\|_1 \\ \text{s.t. } & y_i(\mathbf{x}_i^T\mathbf{w}) \geq 1, \quad \text{for } 1 \leq i \leq n. \end{aligned} \tag{2.107}$$

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.,

$$\text{Prox}_{t\|\cdot\|_1}(\mathbf{u}) = \eta(\mathbf{u}, t), \tag{2.108}$$

where $\eta(x,t) := \frac{x}{|x|}(|x| - t)_+$ is the soft-thresholding operator. 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,

$$\Pi_s(w) = (1-s)\cdot\delta_0(w) + s\cdot\Big(\frac{\phi(\frac{w}{\frac{\kappa}{\sqrt{s}}})}{\frac{\kappa}{\sqrt{s}}}\Big), \tag{2.109}$$

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 (2.94), and present them in terms of q-functions. To streamline our representation, we introduce the following proxies,

$$t_1 = \frac{\sigma\tau}{\sqrt{\frac{\kappa^2}{s}(\alpha - \sigma\tau\gamma)^2 + \beta^2\sigma^2\tau^2\delta}}, \quad t_2 = \frac{1}{\beta\sqrt{\delta}}. \tag{2.110}$$

We also define the function $\chi : \mathbb{R} \to \mathbb{R}_+$ as,

$$\begin{aligned} \chi(t) &= \mathbb{E}\big[(Z-t)_+^2\big], \quad Z \sim \mathcal{N}(0,1) \\ &= Q(t)(1+t^2) - t\phi(t), \end{aligned} \tag{2.111}$$

Where $Q(t) := \int_t^\infty \phi(x)dx$ denotes the tail distribution of standard normal random variable. We are now able to simplify the first three equations in (2.94) and derive the following nonlinear system,

$$\begin{cases} Q(t_1) = \frac{\alpha}{2(\alpha-\sigma\tau\gamma)}, \\ s\cdot Q(t_1) + (1-s)\cdot Q(t_2) = \frac{\sqrt{c_\kappa(\alpha,\sigma)}}{2\beta\sigma\tau\delta}, \\ \frac{s}{t_1^2}\cdot\chi(t_1) + \frac{(1-s)}{t_2^2}\cdot\chi(t_2) = \frac{\kappa^2\alpha^2}{2\sigma^2\tau^2} + \frac{1}{2\tau^2}, \\ \frac{\partial c_\kappa(\alpha,\sigma)}{\partial\alpha} = \frac{2\kappa^2\gamma}{\beta}\sqrt{c_\kappa(\alpha,\sigma)}, \\ \frac{\partial c_\kappa(\alpha,\sigma)}{\partial\sigma} = \frac{2\sqrt{c_\kappa(\alpha,\sigma)}}{\beta\tau}. \end{cases} \tag{2.112}$$

The nonlinear system (2.112) can be solved via numerical methods. For our numerical simulations in Section 2.4 we exploit accelerated fixed-point methods to solve the nonlinear system. Using the result of Lemma 7, 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. $\Omega = \{j : \mathbf{w}^\star_j \neq 0\}$.) For a pre-defined threshold ϵ, we form the following estimate of the support,

$$\hat{\Omega}_\epsilon = \{j : 1 \leq j \leq p, |\hat{\mathbf{w}}_j| > \epsilon\}. \tag{2.113}$$

The following lemma establishes the success in the support recovery:

Lemma 8 (Support Recovery) *For a sparsity factor $s \in (0, 1]$, let the entries of $\mathbf{w}^\star$ have distribution Π_s defined in* (2.109)*, and $\hat{\mathbf{w}}$ be the solution to the optimization* (2.107)*. Then, as $p \to \infty$, we have,*

$$\begin{aligned} \lim_{\epsilon \downarrow 0} P_1(\epsilon) &:= \mathbb{P}\left\{j \notin \hat{\Omega}_\epsilon | j \in \Omega\right\} \xrightarrow{P} 1 - 2Q(\bar{t}_1) \\ \lim_{\epsilon \downarrow 0} P_2(\epsilon) &:= \mathbb{P}\left\{j \in \hat{\Omega}_\epsilon | j \notin \Omega\right\} \xrightarrow{P} 2Q(\bar{t}_2)\ , \end{aligned} \tag{2.114}$$

where $\bar{t}_1$ and $\bar{t}_2$ are defined as in (2.110)*, with variables derived from solving the nonlinear system* (2.112)*.*

Binary classifier (ℓ_∞-GMM)

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 $\mathbf{w}^\star \in \{\pm 1\}^p$ [40]. This problem arises in integer programming and has some connections to the Knapsack problem [110]. Here, we consider analyzing the performance of the solution of the following optimization program,

$$\begin{aligned} \min_{\mathbf{w} \in \mathbb{R}^p} \quad & \|\mathbf{w}\|_\infty \\ \text{s.t. } & y_i(\mathbf{x}_i^T \mathbf{w}) \geq 1, \ \text{ for } 1 \leq i \leq n. \end{aligned} \tag{2.115}$$

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:

$$\text{Prox}_{tp\|\cdot\|_\infty}(\mathbf{w}) = \mathbf{w} - \text{Prox}_{\lambda\|\cdot\|_1}(\mathbf{w}), \tag{2.116}$$

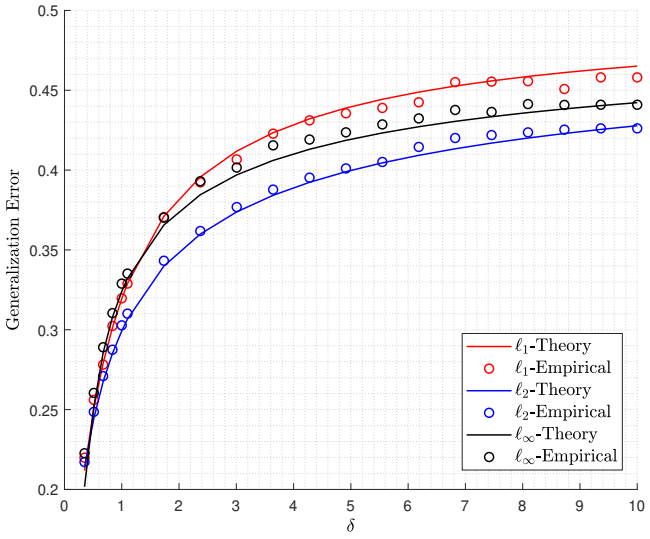


Figure 2.11: Generalization error of the general max margin classifier under three penalty functions, ℓ_1 norm with the red line (ℓ_1-GMM), ℓ_2 norm with the blue line (ℓ_2-GMM), and ℓ_∞ norm with the black line (ℓ_∞-GMM). **In this figure, the entries of $\mathbf{w}^\star$ are drawn independently from $\mathcal{N}(0, \kappa^2)$ Gaussian distribution**. Solid lines correspond to the theoretical results derived from Theorem 7, while the circles are the result of empirical simulations. For the numerical simulations, the result is the average over 100 independent trials with $p = 200$ and $\kappa = 2$.

where $\lambda := \lambda(t)$ is the smallest nonnegative number that satisfies,

$$\mathbb{E}\big[|\eta(W, \lambda)|\big] = \mathbb{E}\big[(|W| - \lambda)_+\big] \leq t. \tag{2.117}$$

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-GMM, here we can use the closed-form of the proximal operator to simplify the first three equations in the nonlinear system (2.94). 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 (2.109), and (3) the uniform binary distribution, $\Pi = \text{Unif}\big(\{\pm 1\}^p\big)$.

Numerical Simulations

In this section, we investigate the validity of our theoretical results with multiple numerical simulations applied to the three different cases of GMM classifiers elaborated in Section 2.4. For each of the three potentials discussed in the paper (i.e., ℓ_1, ℓ_2, and ℓ_∞ norms) we perform numerical simulations for three different models on

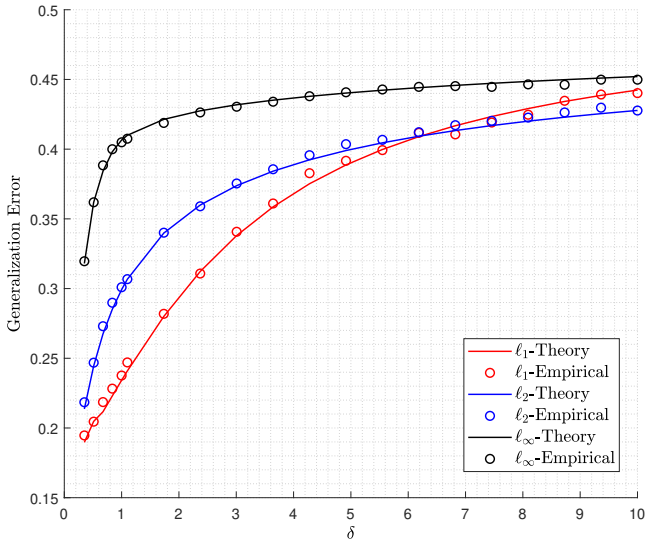


Figure 2.12: Generalization error of the general max margin classifier under three penalty functions, ℓ_1 norm with the red line (ℓ_1-GMM), ℓ_2 norm with the blue line (ℓ_2-GMM), and ℓ_∞ norm with the black line (ℓ_∞-GMM). **In this figure, the underlying vector $\mathbf{w}^\star$ is s-sparse, where the non-zero entries are drawn independently from $\mathcal{N}(0, \kappa^2/s)$ Gaussian distribution**. Solid lines correspond to the theoretical results derived from Theorem 7, and the circles are the result of empirical simulations. For the numerical simulations, the result is computed by taking the average over 100 independent trials with $p = 200$, $s = .1$ and $\kappa = 2$.

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. As will observed in our numerical simulations, the appropriate choice of the potential function in the GMM optimization (2.90) has an impact on the generlization error of the resulting classifier. The three different distribution 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 2.11 gives the generalization error when $\mathbf{w}^\star$ has Gaussian distribution. The solid lines show the theoretical results derived from Theorem 7 and Lemma 7. 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

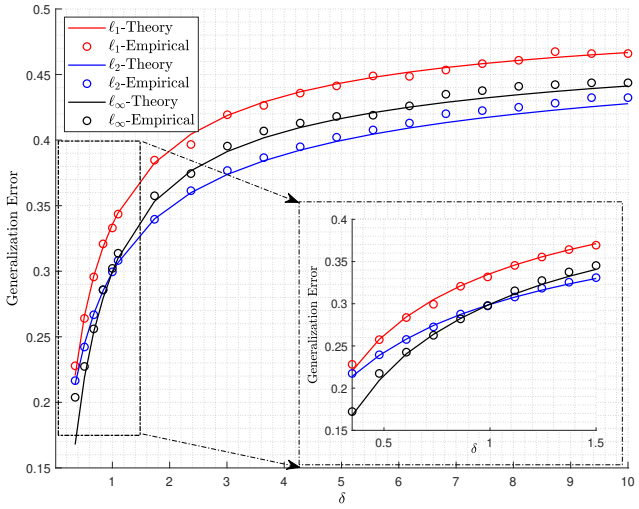


Figure 2.13: Generalization error of the general max margin classifier under three penalty functions, ℓ_1 norm with the red line (ℓ_1-GMM), ℓ_2 norm with the blue line (ℓ_2-GMM), and ℓ_∞ norm with the black line (ℓ_∞-GMM). **In this figure, the entries of $\mathbf{w}^\star$ are drawn independently from $\kappa * \textsc{Rad}(0.5)$ Rademacher distribution**. Solid lines correspond to the theoretical results derived from Theorem 7, and the circles are the result of empirical simulations. For the numerical simulations, the result is the average over 100 independend trials with $p = 200$ and $\kappa = 2$.

in this figure that the max-margin classifier (ℓ_2-GMM) outperforms the other two classifiers. We should also note that as the overparameterization ratio, δ, grows the generalization error increases which indicates that the estimator is not reliable for large values of δ.

Sparse: here, we assume that the entries of $\mathbf{w}^\star$ are drawn from the sparse distribution represented in (2.109), i.e., each entry is nonzero with probability s, and the nonzero entries have i.i.d. Gaussian distribution with appropriately-defined variance. Figure 2.12 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 $\kappa = 2$. Similar to the previous case, the empirical results match the theory. Also, it can be observed that the ℓ_1-GMM outperforms the two other classifiers in the regime of δ that 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 2.13 shows the result of numerical simulations under this model. Similar to previous cases the empirical

results ($\kappa = 2,\ p = 200$) match the theory. Also, the ℓ_∞-GMM classifier outperforms the other two classifiers for $\delta < 1$ (which corresponds to the underparameterized setting). However, the max-margin classifier performs better for larger values of δ.

2.5 Highlight of the Proof and CGMT Framework

Developed by Thrampoulidis et al., Convex Gaussian Min-Max Theorem (CGMT) is a noble framework that enables us to analyze a wide variety of problems, including the convex estimator (2.2), and is the main idea behind the proof of Theorem 1. The CGMT is an extension of a Gaussian comparison inequality proved by Gordon in 1988 [79, 80]. Starting with the works of Rudelson and Vershynin [139], Stojnic [153], Oymak [127], Chandrasekaran [40] and also Amelunxen, Maccoy [6] Gordon's original theorem has played a key role in the analysis of (underdetermined) noiseless linear inverse problems. We refer the interested reader to [165, 171] for more details and a discussion on the relation of the CGMT to the result by Gordon. Before going through the technicalities of CGMT, we present a simple understanding of how CGMT framework can be applied in practice. Recall the convex estimator (2.2)

$$\hat{\beta} = \arg\min_{\beta} \mathcal{L}\left(\mathbf{X}\beta/\sqrt{p}, \mathbf{g}(\mathbf{X}\beta_0/\sqrt{p})\right) + \lambda f(\beta)\ . \tag{2.118}$$

After a couple of steps that we will explain shortly, we can rewrite this optimization in the form of the following min-max optimizations, which we refer to as the *Primary Optimization* **(PO)**,

$$\Phi(\mathbf{X}) := \min_{\beta\in\mathcal{S}_\beta}\ \max_{\mathbf{u}\in\mathcal{S}_\mathbf{u}}\ \mathbf{u}^\mathsf{T}\mathbf{X}\beta + \Psi(\beta, \mathbf{u})\ . \tag{2.119}$$

This is a more complicated form of our initial optimization, because of the extra maximization. Besides, as we go more through the technical details, we can see that one of the main restrictions in the analysis of this optimization, in how the variables $\mathbf{u}$ and β are coupled through the features matrix $\mathbf{X}$. If we were somehow able to decouple these two, the analysis would get much easier. This is what CGMT actually does.

The Convex Gaussian Min-max Theorem associates with the primary optimization (PO) problem a simplified *Auxiliary Optimization* **(AO)** problem, from which we can tightly infer properties of the original (PO), such as the optimal cost, the optimal solution, etc.. In the other words, CGMT introduces another optimization that is much simpler to analyze, and its solution has the same properties as the solution to

the PO. Specifically, the Auxiliary optimization is given as follows

$$\phi(\mathbf{h}_1, \mathbf{h}_2) := \min_{\beta \in \mathcal{S}_\beta} \max_{\mathbf{u} \in \mathcal{S}_\mathbf{u}} \|\beta\|_2 \mathbf{h}_1^\mathsf{T} \mathbf{u} + \|\mathbf{u}\|_2 \mathbf{h}_2^\mathsf{T} \mathbf{w} + \Psi(\beta, \mathbf{u}) \ . \tag{2.120}$$

As you see, the auxiliary optimization is the same as the primary optimization, except the bi-linear form $\mathbf{u}^\mathsf{T}\mathbf{X}\beta$ which is replaces by two two terms $\|\beta\|_2 \mathbf{h}_1^\mathsf{T}\mathbf{u}$ and $\|\mathbf{u}\|_2 \mathbf{h}_2^\mathsf{T}\mathbf{w}$. This simple change will significantly help us later in the analysis.
In the other words, we will be able to analyze the properties of the solution to the auxiliary optimization using techniques in convex analysis and high dimensional probability and random matrix theory (which are not applicable to the primary optimization). Afterwards, the CGMT tells us that the primary and auxiliary optimizations shares some properties, including those of our interest. Next, we will go through a precise analysis of our theorem, using the CGMT framework.

2.6 Proof of Theorem 1

In this section, we rigorously prove Theorem 1. Recall the convex estimator (2.2),

$$\hat{\beta} = \arg\min_{\beta} \mathcal{L}\left(\mathbf{X}\beta/\sqrt{p}, \mathbf{g}(\mathbf{X}\beta_0/\sqrt{p})\right) + \lambda f(\beta) \ . \tag{2.121}$$

Note that the entries of $\mathbf{X}$ are independently drawn from $\mathcal{N}(0,1)$. Now, we rewrite the optimization by introducing a new variable,

$$\min_{\substack{\beta \in \mathbb{R}^p \\ \mathbf{w} = \mathbf{X}\beta_0/\sqrt{p}}} \mathcal{L}\left(\mathbf{w}, \mathbf{g}(\mathbf{X}\beta_0/\sqrt{p})\right) + \lambda f(\beta) \ . \tag{2.122}$$

In the next step, we bring the constraint over $\mathbf{w}$ in the objective function using Lagrange multiplier,

$$\min_{\substack{\beta \in \mathbb{R}^p \\ \mathbf{w} \in \mathbb{R}^n}} \max_{\mathbf{u} \in \mathbb{R}^n} \mathcal{L}\left(\mathbf{w}, \mathbf{g}(\mathbf{X}\beta_0/\sqrt{p})\right) + \lambda f(\beta) + \frac{1}{n}\mathbf{u}^\mathsf{T}(\mathbf{X}\beta/\sqrt{p} - \mathbf{w}) \ . \tag{2.123}$$

Note that the two terms $\mathbf{X}\beta_0$ and $\mathbf{u}^\mathsf{T}\mathbf{X}\beta$ are dependent. We would like to decompose the later term in such a way that one becomes independent of the other. This will help us use CGMT, as you will see later. To do so, we decompose the variable β into its projection in the direction of β_0 and the subspace orthogonal to β_0. In the other words, we rewrite $\beta = \alpha\beta_0 + \tilde{\beta}$, where $\tilde{\beta} \perp \beta_0$. The optimization becomes,

$$\min_{\substack{\tilde{\beta} \in \mathbb{R}^p \\ \mathbf{w} \in \mathbb{R}^n \\ \alpha \in \mathbb{R} \\ \frac{1}{p}\tilde{\beta}^\mathsf{T}\beta_0 = 0}} \max_{\mathbf{u} \in \mathbb{R}^n} \mathcal{L}\left(\mathbf{w}, \mathbf{g}(\mathbf{X}\beta_0/\sqrt{p})\right) + \lambda f(\tilde{\beta} + \alpha\beta_0) + \frac{1}{n}\mathbf{u}^\mathsf{T}(\mathbf{X}\tilde{\beta}/\sqrt{p} + \alpha\mathbf{X}\beta_0/\sqrt{p} - \mathbf{w}) \ . \tag{2.124}$$

Now, since the matrix $\mathbf{X}$ is Gaussian with iid entries, the variables $\mathbf{X}\beta_0$ and $\mathbf{X}\tilde{\beta}$ are independent. We reorder the term on optimization to get

$$\min_{\substack{\tilde{\beta}\in\mathbb{R}^p \\ \mathbf{w}\in\mathbb{R}^n \\ \alpha\in\mathbb{R} \\ \frac{1}{p}\tilde{\beta}^\mathsf{T}\beta_0=0}} \max_{\mathbf{u}\in\mathbb{R}^n} \frac{1}{n\sqrt{p}}\mathbf{u}^\mathsf{T}\mathbf{X}\tilde{\beta} + \left(\mathcal{L}\left(\mathbf{w}, \mathbf{g}(\mathbf{X}\beta_0/\sqrt{p})\right) + \lambda f(\tilde{\beta}+\alpha\beta_0) + \frac{1}{n}\mathbf{u}^\mathsf{T}(\alpha\mathbf{X}\beta_0/\sqrt{p} - \mathbf{w})\right) . \tag{2.125}$$

Now, the first bi-linear term $\mathbf{u}^\mathsf{T}\mathbf{X}\tilde{\beta}$ is independent of the rest of the objective function. Up to now, it seems like we've build a more difficult optimization to analyze by introducing new variables and a maximization. But the point of every step up to now was simply to utilize the CGMT framework. Note that what we have in (2.125) is in the form of the primary optimization (2.119), in the CGMT framework. Next, we give a precise statement of the CGMT framework and apply it to the optimization (2.125).

The Convex Gaussian Min-max Theorem

Consider the primary optimization (PO) and the Auxiliary optimization (AO) below,

$$\Phi(\mathbf{X}) := \min_{\beta\in\mathcal{S}_\beta} \max_{\mathbf{u}\in\mathcal{S}_\mathbf{u}} \mathbf{u}^\mathsf{T}\mathbf{X}\beta + \Psi(\beta,\mathbf{u}) , \quad \text{(PO)} \tag{2.126a}$$

$$\phi(\mathbf{h}_1,\mathbf{h}_2) := \min_{\beta\in\mathcal{S}_\beta} \max_{\mathbf{u}\in\mathcal{S}_\mathbf{u}} \|\beta\|_2\mathbf{h}_1^\mathsf{T}\mathbf{u} + \|\mathbf{u}\|_2\mathbf{h}_2^\mathsf{T}\mathbf{w} + \Psi(\beta,\mathbf{u}) , \quad \text{(AO)} . \tag{2.126b}$$

where $\mathbf{X}\in\mathbb{R}^{n\times p}, \mathbf{h}_1\in\mathbb{R}^n, \mathbf{h}_2\in\mathbb{R}^p, \mathcal{S}_\mathbf{u}\subset\mathbb{R}^n, \mathcal{S}_\beta\subset\mathbb{R}^p$ and $\Psi : \mathbb{R}^p\times\mathbb{R}^n\to\mathbb{R}$. We denote $\beta_\Phi := \beta_\Phi(\mathbf{X})$ and $\beta_\phi := \beta_\phi(\mathbf{h}_1,\mathbf{h}_2)$ any optimal minimizers in (2.126a) and (2.126b), respectively. Then, we have the following result.

Theorem 8 (CGMT) *In* (2.126)*, let* $\mathcal{S}_\beta, \mathcal{S}_\mathbf{u}$ *be compact sets,* Ψ *be continuous on* $\mathcal{S}_\beta\times\mathcal{S}_\mathbf{u}$*, and,* $\mathbf{X}, \mathbf{h}_1$ *and* $\mathbf{h}_2$ *all have entries iid standard normal. The following statements are true:*

1. *For all* $c\in\mathbb{R}$*:*
$$\mathbb{P}(\ \Phi(\mathbf{X}) < c\ ) \le 2\mathbb{P}(\ \phi(\mathbf{h}_1,\mathbf{h}_2)\le c\ ).$$
2. *Further assume that* $\mathcal{S}_\beta, \mathcal{S}_\mathbf{u}$ *are convex sets and* ψ *is convex-concave on* $\mathcal{S}_\beta\times\mathcal{S}_\mathbf{u}$*. Then, for all* $c\in\mathbb{R}$*,*
$$\mathbb{P}(\ \Phi(\mathbf{X}) > c\ ) \le 2\mathbb{P}(\ \phi(\mathbf{h}-1,\mathbf{h}_2)\ge c\ ).$$

In particular, for all $\mu \in \mathbb{R}, t > 0$, $\mathbb{P}(\,|\Phi(\mathbf{X}) - \mu| > t\,) \leq 2\mathbb{P}(\,|\phi(\mathbf{h}_1, \mathbf{h}_2) - \mu| \geq t\,)$.

3. *Let* $\mathcal{S}$ *be an arbitrary open subset of* $\mathcal{S}_\beta$ *and* $\mathcal{S}^c = \mathcal{S}_\beta / \mathcal{S}$ *. Denote* $\phi_{\mathcal{S}^c}(\mathbf{h}_1, \mathbf{h}_2)$ *the optimal cost of the optimization in* (2.126b) *when the minimization over* $\mathbf{w}$ *is now constrained over* $\mathbf{w} \in \mathcal{S}^c$*. If there exist constants* $\overline{\phi}$*,* $\overline{\phi}_{\mathcal{S}^c}$ *and* $\eta > 0$ *such that*

   a) $\overline{\phi}_{\mathcal{S}^c} \geq \overline{\phi} + 3\eta$*,*

   b) $\phi(\mathbf{h}_1, \mathbf{h}_2) < \overline{\phi} + \eta$ *with probability at least* $1 - p$*,*

   c) $\phi_{\mathcal{S}^c}(\mathbf{h}_1, \mathbf{h}_2) > \overline{\phi}_{\mathcal{S}^c} - \eta$ *with probability at least* $1 - p$*,*

   then,

$$\mathbb{P}(\beta_\Phi(\mathbf{X}) \in \mathcal{S}) \geq 1 - 4p.$$

The first two statements of Theorem 8 are identical to [171, Thm. 3], and, a proof is included therein.

The second statement of the theorem states that under the theorem assumption, the values of the optimal objective functions $\Phi(\mathbf{X})$ and $\phi(\mathbf{h}_1, \mathbf{h}_2)$ converge to the same values (if the later converges).

Using the third statement, we show that if the optimal solution of the (AO) lies within some set $\mathcal{S}$ with probability approaching to 1 (as the dimensions p and n increase), and optimizing (AO) over the set $\mathcal{S}^{\mathrm{c}}$ will result in a strictly larger optimal objective value, then the optimal solution in (PO) will be also in the set $\mathcal{S}$ with probability approaching to 1. Using this, we would like to show that the optimal values for α, and $\|\tilde{\beta}\|_2$ will converge to the same value for (AO) and (PO). To do so, we show that these terms in (AO) converge to a unique value, and so will they in the (PO). Rigorously applying this theorem to the primary optimization is discussed in the proof of Theorem 2 in the Appendix. For now, let's apply CGMT to the optimization (2.125), and get the auxiliary optimization from it.

Clearly, the optimization (2.125) has the desired format of (PO) in the CGMT. Therefore, the corresponding Auxiliary Optimization will be

$$\min_{\substack{\tilde{\beta}\in\mathbb{R}^p \\ \mathbf{w}\in\mathbb{R}^n \\ \alpha\in\mathbb{R} \\ \frac{1}{p}\tilde{\beta}^{\mathsf{T}}\beta_0=0}} \max_{\mathbf{u}\in\mathbb{R}^n} \frac{1}{n\sqrt{p}} \|\tilde{\beta}\|\mathbf{h}_1^{\mathsf{T}}\mathbf{u} + \frac{1}{n\sqrt{p}}\|\mathbf{u}\|\mathbf{h}_2^{\mathsf{T}}\tilde{\beta} + \mathcal{L}\left(\mathbf{w}, \mathbf{g}(\mathbf{X}\beta_0/\sqrt{p})\right) + \lambda f(\tilde{\beta} + \alpha\beta_0)$$

$$+ \frac{1}{n}\mathbf{u}^{\mathsf{T}}(\alpha\mathbf{X}\beta_0/\sqrt{p} - \mathbf{w})\,. \tag{2.127}$$

Corollary 3 (Asymptotic CGMT) *Using the same notation as in Theorem 8, suppose there exists constants $\overline{\phi} < \overline{\phi}_{\mathcal{S}^c}$ such that $\phi(\mathbf{g}, \mathbf{h}) \xrightarrow{P} \overline{\phi}$ and $\phi_{\mathcal{S}^c}(\mathbf{g}, \mathbf{h}) \xrightarrow{P} \overline{\phi}_{\mathcal{S}^c}$. Then,*

$$\lim_{n\to\infty} \mathbb{P}(\mathbf{w}_{\Phi}(\mathbf{G}) \in \mathcal{S}) = 1.$$

Analysis of the AO

Out next step would be to analyze the auxiliary optimization. The idea is to use dimension reduction techniques to reduce the optimization (2.127) to an optimization over scalars. To do so, for most of the high dimensional variables, we would like to do the optimization over their direction first.

In the first step, we would like to optimize over the direction of $\mathbf{u}$. In the other words, we rewrite $\mathbf{u}/\sqrt{n} = t \cdot \tilde{\mathbf{u}}$, where $\|\tilde{\mathbf{u}}\| = 1$ and $t \geq 0$. Then, if we solve the optimization over $\tilde{\mathbf{u}}$, we get

$$\min_{\substack{\tilde{\beta}\in\mathbb{R}^p \\ \mathbf{w}\in\mathbb{R}^n \\ \alpha\in\mathbb{R} \\ \frac{1}{p}\tilde{\beta}^{\mathsf{T}}\beta_0=0}} \max_{t\geq 0} \frac{t}{\sqrt{n}} \|\frac{\|\tilde{\beta}\|}{\sqrt{p}}\mathbf{h}_1 + \frac{\alpha}{\sqrt{p}}\mathbf{X}\beta_0 - \mathbf{w}\| + \frac{t}{\sqrt{pn}}\mathbf{h}_2^{\mathsf{T}}\tilde{\beta} + \mathcal{L}\left(\mathbf{w}, \mathbf{g}(\mathbf{X}\beta_0/\sqrt{p})\right)$$
$$+ \lambda f(\tilde{\beta} + \alpha\beta_0)\ . \tag{2.128}$$

Next, we would like to do the same trick for $\tilde{\beta}$, but it also exists inside the function f. We can pull it out using the same trick that we used to pull out $\mathbf{X}\beta$ out of the loss function. By introducing new variable $\mathbf{v}$ to rewrite the optimization as

$$\min_{\substack{\mathbf{v},\tilde{\beta}\in\mathbb{R}^p \\ \mathbf{w}\in\mathbb{R}^n \\ \alpha\in\mathbb{R} \\ \frac{1}{p}\tilde{\beta}^{\mathsf{T}}\beta_0=0 \\ \mathbf{v}=\tilde{\beta}+\alpha\beta_0}} \max_{t\geq 0} \frac{t}{\sqrt{n}} \|\frac{\|\tilde{\beta}\|}{\sqrt{p}}\mathbf{h}_1 + \frac{\alpha}{\sqrt{p}}\mathbf{X}\beta_0 - \mathbf{w}\| + \frac{t}{\sqrt{pn}}\mathbf{h}_2^{\mathsf{T}}\tilde{\beta} + \mathcal{L}\left(\mathbf{w}, \mathbf{g}(\mathbf{X}\beta_0/\sqrt{p})\right)$$
$$+ \lambda f(\mathbf{v})\ . \tag{2.129}$$

And then, using Lagrange multiplier $\mathbf{x}$ and γ to bring the constraints over $\tilde{\beta}$ to the objective function.

$$\min_{\substack{\mathbf{v},\tilde{\beta}\in\mathbb{R}^p \\ \mathbf{w}\in\mathbb{R}^n \\ \alpha\in\mathbb{R}}} \max_{\substack{t\geq 0 \\ \gamma\in\mathbb{R} \\ \mathbf{x}\in\mathbb{R}^p}} \frac{t}{\sqrt{n}} \|\frac{\|\tilde{\beta}\|}{\sqrt{p}}\mathbf{h}_1 + \frac{\alpha}{\sqrt{p}}\mathbf{X}\beta_0 - \mathbf{w}\| + \frac{t}{\sqrt{pn}}\mathbf{h}_2^{\mathsf{T}}\tilde{\beta} + \mathcal{L}\left(\mathbf{w}, \mathbf{g}(\mathbf{X}\beta_0/\sqrt{p})\right)$$
$$+ \lambda f(\mathbf{v}) + \frac{1}{p}\mathbf{x}^{\mathsf{T}}(\tilde{\beta} + \alpha\beta_0 - \mathbf{v}) + \frac{\gamma}{p}\tilde{\beta}^{\mathsf{T}}\beta_0\ . \tag{2.130}$$

Now, we would like to solve the optimization over the direction of $\tilde{\beta}$ as we did so for $\mathbf{u}$. But the optimization above doesn't look convex concave with respect to its vari-

ables and we are not by default allowed to switch minimization and maximization.

At this point, recall that the (PO) in (2.125) is itself convex. In fact, for it, all conditions of Sion's min-max Theorem [149] are met, thus, the order of min-max operations can be flipped. According to the CGMT, the (PO) and the (AO) are tightly related in an asymptotic setting. We use this, to translate the convexity properties of the (PO) to the (AO). In essence, we show that when dimensions grow, the order of min-max operations in the (AO) can be flipped. Thus, we will instead consider the following problem as the (AO):

$$\min_{\substack{\mathbf{v}\in\mathbb{R}^p\\ \mathbf{w}\in\mathbb{R}^n\\ \alpha\in\mathbb{R}\\ \sigma\geq 0}} \max_{\substack{t\geq 0\\ \gamma\in\mathbb{R}\\ \mathbf{x}\in\mathbb{R}^p}} \min_{\sigma=\frac{\|\tilde{\beta}\|}{\sqrt{p}}} \frac{t}{\sqrt{n}} \, \| \frac{\|\tilde{\beta}\|}{\sqrt{p}} \mathbf{h}_1 + \frac{\alpha}{\sqrt{p}} \mathbf{X}\beta_0 - \mathbf{w}\| + \frac{t}{\sqrt{pn}} \mathbf{h}_2^\mathsf{T} \tilde{\beta} + \mathcal{L}\left(\mathbf{w}, \mathbf{g}(\mathbf{X}\beta_0/\sqrt{p})\right)$$
$$+ \lambda f(\mathbf{v}) + \frac{1}{p}\mathbf{x}^\mathsf{T}(\tilde{\beta} + \alpha\beta_0 - \mathbf{v}) + \frac{\gamma}{p}\tilde{\beta}^\mathsf{T}\beta_0 \, . \tag{2.131}$$

Observe that the objective function remains the same; it is only the order of min-max operations that is slightly modified. Since the objective function is not necessarily convex-concave in its arguments, there is no immediate guarantee that the two problems in (2.130) and (2.131) are equivalent for all realizations of $\mathbf{h}_1$ and $\mathbf{h}_2$. However, the lemma below essentially shows that such a strong duality holds with high probability over $\mathbf{h}_1$ and $\mathbf{h}_2$ in high dimensions. Hence, the problem in (2.131) can be as well used, instead of the one in (2.130), in order to analyze the (PO). For this reason, henceforth, we refer to (2.131) as the (AO) problem.

Lemma 9 *Let $\hat{\tilde{\beta}}(\mathbf{X})$ denote an optimal solution of* (2.125)*. Consider the (AO) problem in* (2.131)*. Let σ_* be the value that the optimal σ in* (2.131) *converges to. For any $\epsilon > 0$ define the set $\mathcal{S} := \{\tilde{\beta} \mid |\|\tilde{\beta}\|_2 - \sigma_*| < \epsilon\}$, and, $\phi_{\mathcal{S}^c}(\mathbf{h}_1, \mathbf{h}_2)$ be the optimal cost of the same optimization as in* (2.131)*, only this time the minimization over $\tilde{\beta}$ is further constrained such that $\tilde{\beta} \notin \mathcal{S}$. Then,*

$$\lim_{p\to\infty} \mathbb{P}\left(\, |\|\hat{\tilde{\beta}}(\mathbf{X})\|_2 - \sigma_*| < \epsilon \, \right) = 1.$$

This lemma and its proof is similar to Lemma A.3 in [165]. Now we solve the

optimization over the direction of $\tilde{\beta}$ and let $\sigma := \frac{\|\tilde{\beta}\|}{\sqrt{p}}$. We get

$$\min_{\substack{\mathbf{v}\in\mathbb{R}^p\\ \mathbf{w}\in\mathbb{R}^n\\ \alpha\in\mathbb{R},\sigma\geq 0}} \max_{\substack{t\geq 0\\ \gamma\in\mathbb{R}\\ \mathbf{x}\in\mathbb{R}^p}} \frac{t}{\sqrt{n}}\|\sigma\mathbf{h}_1 + \frac{\alpha}{\sqrt{p}}\mathbf{X}\beta_0 - \mathbf{w}\| - \frac{\sigma}{\sqrt{p}}\|\gamma\beta_0 + \mathbf{x} + \frac{t}{\sqrt{\delta}}\mathbf{h}_2\| + \frac{1}{p}\mathbf{x}^\mathsf{T}(\alpha\beta_0 - \mathbf{v}) + \mathcal{L}\left(\mathbf{w}, \mathbf{g}(\mathbf{X}\beta_0/\sqrt{p})\right) + \lambda f(\mathbf{v}) \ . \tag{2.132}$$

Now, we have gotten closer to our goal which is to build an optimization over scalars. In the next step, we would like to do the optimization over $\mathbf{x}$. What helps us next, is the following trick,

$$x = \min_{\tau\geq 0} \frac{1}{2\tau} + \frac{x^2\tau}{2} \ . \tag{2.133}$$

Using this trick, we would like to square the $\|\cdot\|_2$ norms in the optimization and get

$$\min_{\substack{\mathbf{v}\in\mathbb{R}^p\\ \mathbf{w}\in\mathbb{R}^n\\ \alpha\in\mathbb{R}\\ \sigma,\tau_1\geq 0}} \max_{\substack{t,\tau_2\geq 0\\ \gamma\in\mathbb{R}\\ \mathbf{x}\in\mathbb{R}^p}} \frac{t}{2\tau_1} + \frac{t\tau_1}{2n}\|\sigma\mathbf{h}_1 + \frac{\alpha}{\sqrt{p}}\mathbf{X}\beta_0 - \mathbf{w}\|^2 - \frac{\sigma}{2\tau_2} - \frac{\sigma\tau_2}{2p}\|\gamma\beta_0 + \mathbf{x} + \frac{t}{\sqrt{\delta}}\mathbf{h}_2\|^2 + \frac{1}{p}\mathbf{x}^\mathsf{T}(\alpha\beta_0 - \mathbf{v}) + \mathcal{L}\left(\mathbf{w}, \mathbf{g}(\mathbf{X}\beta_0/\sqrt{p})\right) + \lambda f(\mathbf{v}) \ . \tag{2.134}$$

Now we can simply do the optimization over $\mathbf{x}$ and get

$$\min_{\substack{\mathbf{v}\in\mathbb{R}^p\\ \mathbf{w}\in\mathbb{R}^n\\ \alpha\in\mathbb{R}\\ \sigma,\tau_1\geq 0}} \max_{\substack{t,\tau_2\geq 0\\ \gamma\in\mathbb{R}}} \frac{t}{2\tau_1} + \frac{t\tau_1}{2n}\|\sigma\mathbf{h}_1 + \frac{\alpha}{\sqrt{p}}\mathbf{X}\beta_0 - \mathbf{w}\|^2 - \frac{\sigma}{2\tau_2} + \frac{1}{2\sigma\tau_2 p}\|(\alpha - \sigma\gamma\tau_2)\beta_0 + \frac{t\tau_2\sigma}{\sqrt{\delta}}\mathbf{h}_2 - \mathbf{v}\|^2 - \frac{\sigma\tau_2}{2p}\|\gamma\beta_0 + \frac{t}{\sqrt{\delta}}\mathbf{h}_2\|^2 + \mathcal{L}\left(\mathbf{w}, \mathbf{g}(\mathbf{X}\beta_0/\sqrt{p})\right) + \lambda f(\mathbf{v}) \ . \tag{2.135}$$

Now, we use the Monreau envelope notation defined in (2.4), for the minimization over $\mathbf{v}$ and $\mathbf{w}$. Recall that the Monreau envelope is defined as

$$\mathrm{e}_{\mathcal{L}}(\mathbf{x}, \mathbf{y}, \tau) := \min_{\mathbf{v}} \frac{1}{2\tau}\|\mathbf{v} - \mathbf{x}\|^2 + \mathcal{L}(\mathbf{v}, \mathbf{y}) \ ,$$

$$\mathrm{e}_f(\mathbf{x}, \tau) := \min_{\mathbf{v}} \frac{1}{2\tau}\|\mathbf{v} - \mathbf{x}\|^2 + f(\mathbf{x}) \ . \tag{2.136}$$

$$\tag{2.137}$$

Also let $\kappa = \|\beta_0\|/\sqrt{p}$. We can replace $\mathbf{X}\beta_0/\sqrt{p}$ with $\kappa\mathbf{h}_3$, where $\mathbf{h}_3$ is a random vector with iid standard Gaussians. Replacing these in (2.135) yields,

$$\min_{\substack{\alpha\in\mathbb{R}\\ \sigma,\tau_1\geq 0}} \max_{\substack{t,\tau_2\geq 0\\ \gamma\in\mathbb{R}}} \frac{t}{2\tau_1} - \frac{\sigma}{2\tau_2} - \frac{\sigma\tau_2}{2p}\|\gamma\beta_0 + \frac{t}{\sqrt{\delta}}\mathbf{h}_2\|^2 + \mathrm{e}_{\mathcal{L}}\left(\sigma\mathbf{h}_1 + \alpha\kappa\mathbf{h}_3, \mathbf{g}(\kappa\mathbf{h}_3), \frac{1}{t\tau_1}\right) + \mathrm{e}_f\left(-\frac{\sigma\tau_2 t}{\sqrt{\delta}}\mathbf{h}_2 + (\alpha - \sigma\tau_2\gamma)\beta_0, \sigma\tau_2\right) \ . \tag{2.138}$$

Up to now, analysis of the Auxiliary optimization was precise without letting the problem dimensions grow or applying any convergence lemma. Now, we would like to replace the Monreau envelope functions and and the norm of the random vector above (third term) with their corresponding point-wise convergence functions. But since these functions are inside the min-max, we are typically not allowed to do so. But in this case, we would like to use the following lemma in convex analysis which will make this possible.

Lemma 10 *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, $\inf_{x>0} M_n(x) \xrightarrow{P} \inf_{x>0} F(x)$.

It's not hard to check that the objective function in (2.138) satisfies the assumptions of lemma 10. By applying this lemma consecutively on (2.138), and using Assumptions 1, we can do the following replacements in the objective function of (2.138),

$$\begin{aligned}
&\frac{\sigma\tau_2}{2p}\|\gamma\beta_0 + \frac{t}{\sqrt{\delta}}\mathbf{h}_2\|^2 \xrightarrow{P} \frac{\sigma\tau_2 t^2}{2\delta} + \frac{\sigma\tau_2\gamma^2\kappa^2}{2} \\
&\mathrm{e}_{\mathcal{L}}\left(\sigma\mathbf{h}_1 + \alpha\kappa\mathbf{h}_3, \mathbf{g}(\kappa\mathbf{h}_3), \frac{1}{t\tau_1}\right) \xrightarrow{P} L(\sigma,\alpha,\frac{1}{t\tau_1}) \\
&\mathrm{e}_f\left(-\frac{\sigma\tau_2 t}{\sqrt{\delta}}\mathbf{h}_2 + (\alpha - \sigma\tau_2\gamma)\beta_0, \sigma\tau_2\right) \xrightarrow{P} F(-\frac{\sigma\tau_2 t}{\sqrt{\delta}}, \alpha - \sigma\tau_2\gamma, \sigma\tau_2)\ ,
\end{aligned} \tag{2.139}$$

where the functionals L and F are defined in Assumption 1. Therefore, (2.138) becomes

$$\min_{\substack{\alpha\in\mathbb{R}\\ \sigma,\tau_1\geq 0}} \max_{\substack{t,\tau_2\geq 0\\ \gamma\in\mathbb{R}}} \frac{t}{2\tau_1} - \frac{\sigma}{2\tau_2} - \frac{\sigma\tau_2 t^2}{2\delta} + \frac{\sigma\tau_2\gamma^2\kappa^2}{2} + L(\sigma,\alpha,\frac{1}{t\tau_1}) + F(-\frac{\sigma\tau_2 t}{\sqrt{\delta}}, \alpha - \sigma\tau_2\gamma, \sigma\tau_2)\ . \tag{2.140}$$

This is the scalar optimization in Theorem 1. Let's denote its solutions with the scalars $(\hat{\alpha}, \hat{\sigma}, \hat{\tau}_1, \hat{\tau}_2, \hat{t}, \hat{\gamma})$. Then, if one follows the steps of the analysis of the (AO), it can be observed that the final solution $\hat{\beta}$ in the Auxiliary optimization (2.127) in terms of the scalars $(\alpha, \sigma, \tau_1, \tau_2, t, \gamma)$, is

$$\hat{\beta}_{(\mathrm{AO})} = \mathrm{Prox}_f\left((\hat{\alpha} - \hat{\sigma}\hat{\tau}_2\hat{\gamma})\beta_0 + (\frac{\hat{\sigma}\hat{\tau}_2\hat{t}}{\sqrt{\delta}})\mathbf{h}_2\ , \hat{\sigma}\hat{\tau}_2\right) \tag{2.141}$$

Therefore, using Assumption 2, for the auxiliary optimization, we have

$$\lim_{p,n\to\infty} \psi(\hat{\beta}_{(\mathrm{AO})}, \beta_0) = \Psi(\hat{\alpha} - \hat{\sigma}\hat{\tau}_2\hat{\gamma}, \frac{\hat{\sigma}\hat{\tau}_2\hat{t}}{\sqrt{\delta}}, \hat{\sigma}\hat{\tau}_2) \ . \tag{2.142}$$

Now, let's define the set $\mathcal{S}$ to be

$$\mathcal{S} = \{(\alpha, \sigma, \tau_1, \tau_2, t, \gamma) \quad \text{s.t. } |\Psi(\alpha - \sigma\tau_2\gamma, \frac{\sigma\tau_2 t}{\sqrt{\delta}}, \sigma\tau_2) - \Psi(\hat{\alpha} - \hat{\sigma}\hat{\tau}_2\hat{\gamma}, \frac{\hat{\sigma}\hat{\tau}_2\hat{t}}{\sqrt{\delta}}, \hat{\sigma}\hat{\tau}_2)| < \epsilon\} \tag{2.143}$$

Since the solution to the optimization (2.140) is unique, if we optimize the auxiliary optimization over $\mathcal{S}^{\mathrm{c}}$, the value of optimal objective is going to increase. Thus, we can utilize the third part of Theorem 8 to show that the same convergence happens in Primary Optimization and we will have

$$\lim_{p,n\to\infty} \psi(\hat{\beta}_{(\mathrm{PO})}, \beta_0) = \Psi(\hat{\alpha} - \hat{\sigma}\hat{\tau}_2\hat{\gamma}, \frac{\hat{\sigma}\hat{\tau}_2\hat{t}}{\sqrt{\delta}}, \hat{\sigma}\hat{\tau}_2) \ . \tag{2.144}$$

This concludes the proof.

Chapter 3

GENERAL PERFORMANCE METRICS FOR THE LASSO

In this chapter[1], we extend the applicability of the CGMT framework and the precise results that it yields to more general performance metrics. For concreteness, we focus primarily on the problem of sparse recovery under ℓ_1-regularized least-squares (a.k.a LASSO). We also discuss how the results extend to more general structured signal recovery problems and to a wide family of convex recovery methods known as regularized M-estimators. We establish accurate predictions of a wide range of performance metrics that have a Lipschitz property. For illustration,this result can be used to accurately predict the probability that the LASSO successfully identifies the non-zero entries of the unknown signal; specializing the result to the high-SNR regime yields bounds that are geometric in nature and admit insightful interpretations.

There is an increasing line of work on the precise analysis of regularized M-estimators. Please see [165, Sec. 7] for an exhaustive review. We have already referred to the works that use the CGMT framework [129, 154, 165, 171]. The most general result is included in [165] which characterizes the ℓ_2-reconstruction error of general regularized M-estimators under very generic settings; yet, no other performance metrics have been considered thus far. A different line of works is based on a state evolution framework for an iterative Approximate Message Passing (AMP) algorithm inspired by statistical physics ([18, 63] and the references therein). To the best of our knowledge, the AMP framework has not been used to analyze general regularized M-estimators. Nonetheless, [18, 63, 119] have considered Lipschitz performance metrics for the LASSO; our result extends the formulae to general regularized M-estimators. Overall, the two methods of analysis are very different (and of their own value each); this is the first time that the CGMT framework is used for general performance metrics.

3.1 Problem Setup

Consider the problem of recovering a sparse signal $\beta_0 \in \mathbb{R}^p$ comprised of only k non-zero measurements from n noisy linear observations of the form $\mathbf{y} = \mathbf{X}\beta_0 + \mathbf{z} \in \mathbb{R}^n$, where $\mathbf{X}$ is the measurement matrix and $\mathbf{z}$ is the noise vector. The typical

[1] This chapter is mostly based on [4]

approach to produce an estimate $\hat{\beta}$ of β_0 is by solving an ℓ_1-regularized least-squares minimization, as follows:

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2 + \frac{\lambda}{\sqrt{p}}\|\beta\|_1. \tag{3.1}$$

Here, $\lambda > 0$ is a regularization parameter. (The normalization with $\sqrt{p}$ is for convenience in the analysis). This method is known as row Square-root LASSO in the statistics literature [22], and is a slight variation of the popular LASSO; see [129] for a discussion. Our analysis applies to both instances, but we focus on the former for concreteness. Also, for convenience, we shall often refer to (3.1) simply as the LASSO.

Measuring Performance

A "good estimate" might translate to a variety of different desired attributes associated with $\hat{\beta}$. This translates to a variety of different *performance metrics*, which we discuss here.

ℓ_2-reconstruction error: A standard and somewhat generic measure of performance is the ℓ_2-reconstruction error, which measures the deviation of $\hat{\beta}$ from the true signal β_0 in the ℓ_2-norm. Formally, the metric acts on the *reconstruction error vector* $\hat{\mathbf{w}} := \hat{\beta} - \beta_0$ and returns its Euclidean norm, i.e., $\Psi_{\ell_2}(\hat{\mathbf{w}}) := \|\hat{\mathbf{w}}\|_2 = \|\hat{\beta} - \beta_0\|_2$. The ℓ_2-error in estimating the coefficients of β_0 also controls the mean squared prediction error, i.e. the error in predicting a (future) response to a fresh (random) measurement (e.g. [130, Sec. 8.1]).

Lipschitz Metrics: Beyond the ℓ_2-reconstruction error, we consider performance metrics $\Psi : \mathbb{R}^p \to \mathbb{R}$ that act on the error vector $\hat{\mathbf{w}} := \hat{\beta} - \beta_0$ and which satisfy a Lipschitz property, i.e. $|\Psi(\mathbf{x}) - \Psi(\mathbf{y})| \leq L \cdot \|\mathbf{x} - \mathbf{y}\|_2$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$ and some L. One common such metric is $\Psi(\mathbf{w}) = \|\mathbf{w}\|_1$, [123].

Support Recovery: In the problem of sparse recovery a natural performance metric that arises in a variety of contexts (e.g. subset selection in regression, structure estimation in graphical models, sparse approximation [188]) is that of support recovery, i.e. identifying whether an entry of the unknown signal β_0 is on the support (aka is non-zero), or it is off the support (aka is zero). We take a decision based on the solution $\hat{\beta}$ of the LASSO: declare the i^{th} entry to be on the support iff $|\hat{\beta}_i| \geq \epsilon$. Here $\epsilon > 0$ is a user-defined threshold imposed on $\hat{\beta}$; such a hard-thresholding operation is practical due to machine precision inaccuracies in solving (3.1).

In Theorem 10 we accurately predict the *(per-entry) rate of successful on-support and off-support recovery*. Formally, let

$$\Phi_{\epsilon,\text{on}}(\hat{\beta}) = \frac{1}{k} \sum_{i \in S(\beta_0)} \mathbb{1}_{\{|\hat{\beta}_i| \geq \epsilon\}} \tag{3.2a}$$

$$\Phi_{\epsilon,\text{off}}(\hat{\beta}) = \frac{1}{n-k} \sum_{i \notin S(\beta_0)} \mathbb{1}_{\{|\hat{\beta}_i| \leq \epsilon\}}, \tag{3.2b}$$

where $\mathbb{1}_{\mathcal{S}}$ is the indicator function of a set $\mathcal{S}$. The metric $\Phi_{\epsilon,\text{on}}(\hat{\beta})$ (resp. $\Phi_{\epsilon,\text{off}}(\hat{\beta})$) measures the ratio of the non-zero (reps. zero) entries of β_0 that are properly identified to be on (resp. off) the support.

An equivalent way to interpret the metrics defined above is to consider their expectation. For instance, $\mathbb{E}[\Phi_{\epsilon,\text{on}}(\hat{\beta})] = (1/k) \sum_{i \in \mathcal{S}(\beta_0)} \mathbb{P}(|\hat{\beta}_i| \geq \epsilon)$ measures the *average* probability that a single non-zero entry of β_0 is correctly identified to be on the support. In particular, if the entries of $\hat{\beta}$ are iid, then in the limit $\Phi_{\epsilon,\text{on}}(\hat{\beta})$ converges to the probability that a single on-support entry is correctly identified.

Working Hypothesis

The unknown signal $\beta_0 \in \mathbb{R}^p$ is k-sparse: its first k entries are sampled iid from a distribution p_{β_0} of zero mean and of unit variance ($\mathbb{E}[(\beta_0)_i^2] = 1$), and the rest of them are zero. The measurement matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ has entries iid zero mean Gaussian random variables with variance $\frac{1}{p}$ (denote $\mathcal{N}(0, 1/p)$). The noise vector $\mathbf{z} \in \mathbb{R}^n$ has entries iid $\mathcal{N}(0, \sigma^2)$. We study the linear asymptotic regime in which the problem dimensions p, n and k all grow to infinity at proportional rates [2]: $k/p \to \rho \in (0, 1)$ and $n/p \to \delta \in (0, \infty)$. Also, the regularizer parameter λ in (3.1) is considered to be constant, in particular independent of p. Under the current setting, the Signal to Noise Ratio (SNR) becomes SNR $:= \rho/\sigma^2$.

3.2 Results

We gather our main results in this section. For a sequence of random variables $\{\mathcal{X}^{(p)}\}$ and a constant c, $\{\mathcal{X}^{(n)}\} \xrightarrow{P} c$ denotes convergence in probability as $p \to \infty$. We reserve the letters H and X_0 to denote (scalar) random variables with distributions $\mathcal{N}(0, 1)$ and p_{β_0}, respectively.

[2] The results of Section 3.2 apply on a sequence of problem instances $\{\beta_0, \mathbf{X}, \mathbf{z}, n, k\}_p$ indexed by $p \in \mathbb{N}$ such that the properties mentioned hold for all members of the sequence for all p. To keep notation clear we do not explicitly use the subscript p for symbols of the sequence.

ℓ_2-reconstruction Error

The precise characterization of the ℓ_2-reconstruction error has been performed in [164, 172] via the CGMT framework (also, [18, 63] have analyzed the problem via an alternative framework called AMP). We include a statement of the result here since it helps us set up some necessary definitions for the presentation of the more general result that follows in the next section.

ψ-distance functional: For a function $\psi : \mathbb{R} \to \mathbb{R}$, let $\text{Dist}_{\psi(.)}(\cdot,\cdot) : \mathbb{R} \times \mathbb{R}_{>0} \to \mathbb{R}$ be defined as

$$\begin{aligned}\text{Dist}_{\psi(.)}(\kappa,\lambda) := \rho \cdot \mathbb{E}[\psi(X_0 - \eta(\kappa H + X_0, \kappa\lambda))] \\ + (1-\rho) \cdot \mathbb{E}[\psi(\eta(\kappa H, \kappa\lambda))],\end{aligned} \tag{3.3}$$

where the expectation is over both $X_0 \sim p_{\beta_0}$ and $H \sim \mathcal{N}(0,1)$, and $\eta(X,\tau) = (X/|X|)\max\{|X|-\tau, 0\}$ denotes the soft-thresholding operator. The function returns the distance, with respect to the function $\psi(.)$, between a r.v. X_0 and the soft threshold operator applied to the random variable itself after adding a Gaussian noise to it. This motivates the terminology used. Also, note the implicit dependence of the functional on the rest of the problem parameters, namely ρ, δ and σ.

noindentλ_{crit}: There exists a critical value of the regularizer parameter, namely λ_{crit}, such that the error behavior is different when $\lambda \leq \lambda_{\text{crit}}$ compared to $\lambda > \lambda_{\text{crit}}$ [129]. Define the pair $(\alpha_{\text{crit}}, \lambda_{\text{crit}})$ as the solution to the following system of equations:

$$\begin{cases}\alpha_{\text{crit}}^2 = \text{Dist}_{(.)^2}(\kappa_{\text{crit}}, \lambda_{\text{crit}}), \\ \delta = \rho \cdot \mathbb{P}\{|\kappa H + X_0| \geq \lambda_{\text{crit}}\kappa_{crit}\} + 2(1-\rho)Q(\lambda_{\text{crit}}),\end{cases} \tag{3.4}$$

where $\kappa_{crit} = \sqrt{(\alpha_{\text{crit}}^2 + \sigma^2)/\delta}$ and $Q(\cdot)$ is the standard Q-function. It is shown in [164, Sec. 2.C] that if $\delta \leq 1$, then (3.4) has a unique solution. Otherwise, define $\lambda_{\text{crit}} = 0$. With these we are ready to state the first lemma.

Lemma 11 ([164]) *Under the working hypothesis of Section 3.1 and for any fixed $\lambda > 0$, define $\alpha := \alpha(\lambda)$ as the unique solution to the equation $\alpha^2 = Dist_{(.)^2}(\sqrt{(\alpha^2+\sigma^2)/\delta}, \lambda)$, if $\lambda \geq \lambda_{crit}$, and, as $\alpha = \alpha_{crit}$, otherwise. Then, it holds in probability that $\lim_{p\to\infty} \frac{1}{p}\|\hat{\beta} - \beta_0\|_2^2 = \alpha^2$.*

Extensive empirical evidence suggest that the system of the two nonlinear equations in two unknowns in (3.4) can be solved numerically very efficiently using a simple iterative fixed-point method (see also [165, Rem. 4.3.3]). Figure 3.1 below illustrates the accuracy of the lemma.

Lipschitz Performance Metrics

Theorem 9 below generalizes Lemma 11 to metrics that attain a Lipschitz property. Assumption 3 below formally defines the required properties of such metrics.

Assumption 3 (Lipschitz metrics) *We say Assumption 3 holds for the Lipschitz function* $\Psi : \mathbb{R}^p \to \mathbb{R}$ *if*

• *For all constants* $c > 0$*, there exists a constant* $C > 0$ *such that for all* $\beta \in \mathbb{R}^p$ *that* $\|\beta\| \leq c\sqrt{p}$*, we have* $|\Psi(\beta)| \leq C\sqrt{p}$*.*

• *For all* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$*,* $|\Psi(\mathbf{x}) - \Psi(\mathbf{y})| \leq \frac{L}{\sqrt{p}}\|\mathbf{x} - \mathbf{y}\|_2$*, for a constant* L *independent on* p*.*

• *For all* $\alpha, \lambda > 0$ *and* $\mathbf{h} \sim \mathcal{N}(0, \mathbf{I}_p)$*, there exists function* $\Gamma : \mathbb{R}_{>0} \times \mathbb{R}_{>0} \to \mathbb{R}$ *such that*

$$\Psi(\beta_0 - \eta(\kappa\mathbf{h} + \beta_0, \lambda\kappa)) \xrightarrow{P} \Gamma(\kappa, \lambda). \tag{3.5}$$

Here, η *is the "vector" soft-threshold operator acting element-wise on the entries of its first argument.*

The first is a simple scaling requirement such that $\Psi(\mathbf{x}) = O(1)$. The second imposes a growth condition on the Lipschitz constant with respect to p (this is necessary for the asymptotic analysis but can potentially be relaxed). The third requirement of Assumption 3 is easier to interpret in the "separable-case" in which $\Psi(\mathbf{x}) = (1/b)\sum_i \psi(\mathbf{x}_i)$ for some L-Lipschitz scalar function ψ. Then, condition (3.5) holds by the WLLN for $\Gamma(\kappa, \lambda) = \text{Dist}_\psi(\kappa, \lambda)$ (recall (3.3)).

Theorem 9 (Lipschitz performance of LASSO) *Under the working hypothesis of Section 3.1 and with* α *and* λ_{crit} *defined as in Lemma 11, fix* $\lambda > 0$*, let* $\hat{\lambda} = max\{\lambda, \lambda_{crit}\}$ *and* $\kappa = \sqrt{\alpha^2 + \sigma^2}/\sqrt{\delta}$*. Then, for any Lipschitz function* $\Psi(x)$ *that satisfies Assumption 3, it holds in probability that,* $\lim_{p\to\infty} \Psi(\hat{\beta} - \beta_0) = \Gamma(\kappa, \hat{\lambda})$.

Evaluating the prediction only involves identifying the function Γ as per Assumption 3, and calculating the parameters α and λ_{crit} as per Lemma 11. Of course, Lemma 11 follows from Theorem 9 when applied for $\Psi(\hat{\beta} - \beta_0) = \frac{1}{\sqrt{p}}\|\tilde{\beta} - \beta_0\|_2$, since the latter is easily shown to satisfy Assumption 3 for $\Gamma(\kappa, \lambda) = \sqrt{\text{Dist}_{(.)^2}(\kappa, \lambda)}$. A different Lipschitz performance metric that is often of interest in practice is the ℓ_1-reconstruction errof $\Psi(\hat{\beta} - \beta_0) = (1/p)\|\hat{\beta} - \beta_0\|_1$. This is an example of a separable

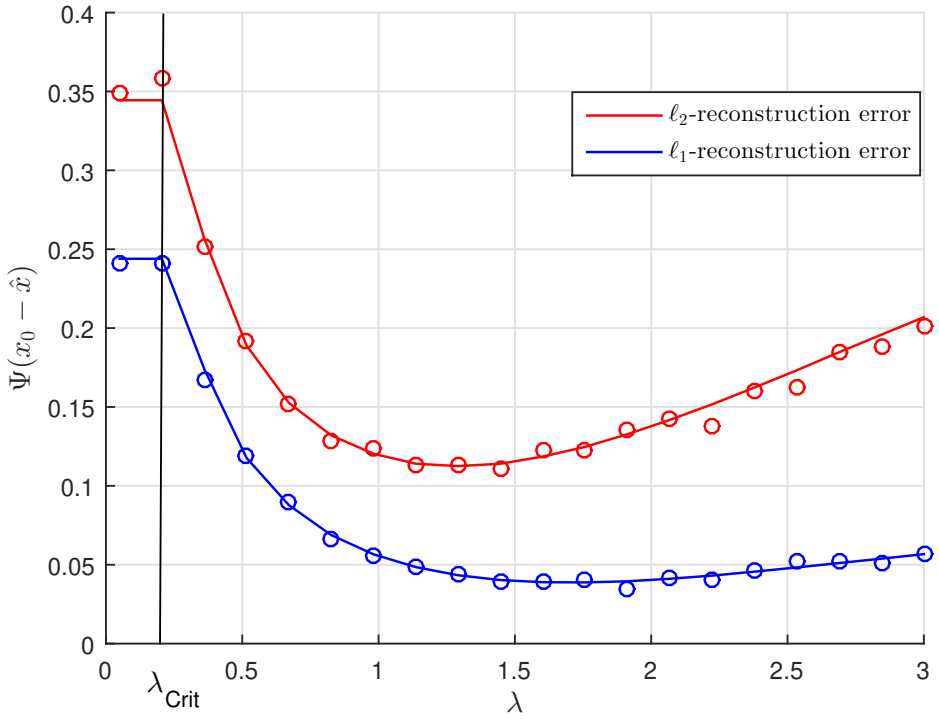


Figure 3.1: Performance of Square root Lasso with respect to $\Psi(\mathbf{x}) = \frac{1}{\sqrt{p}}\|\mathbf{x}\|_2$ (Red Line) and $\Psi(\mathbf{x}) = \frac{1}{p}\|\mathbf{x}\|_1$ (Blue line) as a function of λ. The theoretical prediction comes from Theorem 9. For the simulations, we used $p = 256$, $\delta = 0.8$, $\rho = 0.1$, SNR=0.5 and the data are averaged over 5 independent realization of the Problem.

metric, thus it satisfies Assumption 3 for $\Gamma(\kappa, \lambda) = \text{Dist}_{|\cdot|}(\kappa, \lambda)$. See Figure 3.1 for an illustration. Observe that the prediction of the theorem (although asymptotic) is accurate for problem dimensions of only a few hundreds. Also, the precise nature of the predictions allows optimal tuning of the regularizer parameter λ, the number of measurements δ, etc..

Support Recovery

Theorem 10 below characterizes the support recovery metrics introduced in (3.2). Recall that $\epsilon > 0$ is a fixed hard threshold imposed on the entries of the solution $\hat{\beta}$ to the LASSO in order to decide whether an entry is on or off the support.

Theorem 10 (Probability of support recovery) *Under the working hypothesis of Section 3.1 and with α and λ_{crit} defined as in Lemma 11, fix $\lambda > 0$, let $\kappa = \sqrt{(\alpha^2 + \sigma^2)/\delta}$ and $\hat{\lambda} = \max\{\lambda_{crit}, \lambda\}$. Then, for any $\epsilon > 0$, it holds in probability that $\lim_{p\to\infty} \Phi_{\epsilon,on}(\hat{\beta}) = \mathbb{P}\{|\kappa H + X_0| \geq \epsilon + \hat{\lambda}\kappa\}$ and*

$\lim_{p\to\infty} \Phi_{\epsilon,off}(\hat{\beta}) = \mathbb{P}\{|\kappa H| \leq \epsilon + \hat{\lambda}\kappa\}$.

The metrics in (3.2) are not Lipschitz. Hence, they don't satisfy all requirements of Assumption 3 of Section 3.2, and Theorem 9 is not directly applicable.

Nonetheless, the core idea behind the proof of the theorem is similar to that of Theorem 9 and requires only a few extra arguments (see Section 3.3). Figure 3.2 illustrates the validity of the prediction.

Remark 23 (Off-support) *When $\epsilon \ll \hat{\lambda}\kappa$, the formula of the theorem for $\Phi_{\epsilon,off}(\hat{\beta})$ reduces to $\mathbb{P}\{|\kappa H| \leq \epsilon + \hat{\lambda}\kappa\} \sim \mathbb{P}\{|H| \leq \hat{\lambda}\}$, which is independent of the problem parameters δ, ρ and SNR. This simple observation is verified in Figure 3.2: the off-support recovery probability is the same for different values of under-sampling parameter δ as long as $\lambda \geq \lambda_{crit}$.*

Remark 24 (Large/Small λ) *It is easy to conclude from Theorem 10 that as λ becomes large $\Phi_{\epsilon,off}$ (reps. $\Phi_{\epsilon,on}$) converge to one (resp. zero). Of course, this behavior is expected since large values for the regularizer parameter put more emphasis on the ℓ_1-regularization term in* (3.1)*, thus promoting sparser solutions. Reversed behavior is observed when λ takes values close to zero.*

Remark 25 (Optimal λ) *A natural question becomes that of determining the optimal value of the regularizer parameter. In order to balance between on- and off- support recovery probabilities a reasonable performance metric becomes $\Phi_\epsilon = \omega\Phi_{\epsilon,on} + (1-\omega)\Phi_{\epsilon,off}$ for $\omega \in [0,1]$. Theorem 10 precisely characterizes the behavior of this as a function of λ; thus, it determines the optimal value of λ that minimizes Φ_ϵ.*

Remark 26 (High-SNR Regime) *Here, we analyze the probability of support recovery at SNR $\gg 1$ (eqv. $\sigma^2 \to 0$). In this regime, λ_{crit} takes a simple form: if $\delta < 1$, then $\lambda_{crit} = Q^{-1}(\frac{\delta-\rho}{2(1-\rho)})$ where Q^{-1} is the inverse Q-function, otherwise, $\lambda_{crit} = 0$ [129, Sec. 8]. Let us first examine the behavior of "off-support" recovery probability. When $\sigma^2 \ll 1$, the formula of Theorem 10 reduces to the following simpler one:*

$$\lim_{p\to\infty} \Phi_{\epsilon,off}(\hat{\beta}) \sim 1 - 2Q\left(\hat{\lambda} + \frac{\epsilon}{\sigma}\sqrt{\delta - D(\hat{\lambda})}\right), \tag{3.6}$$

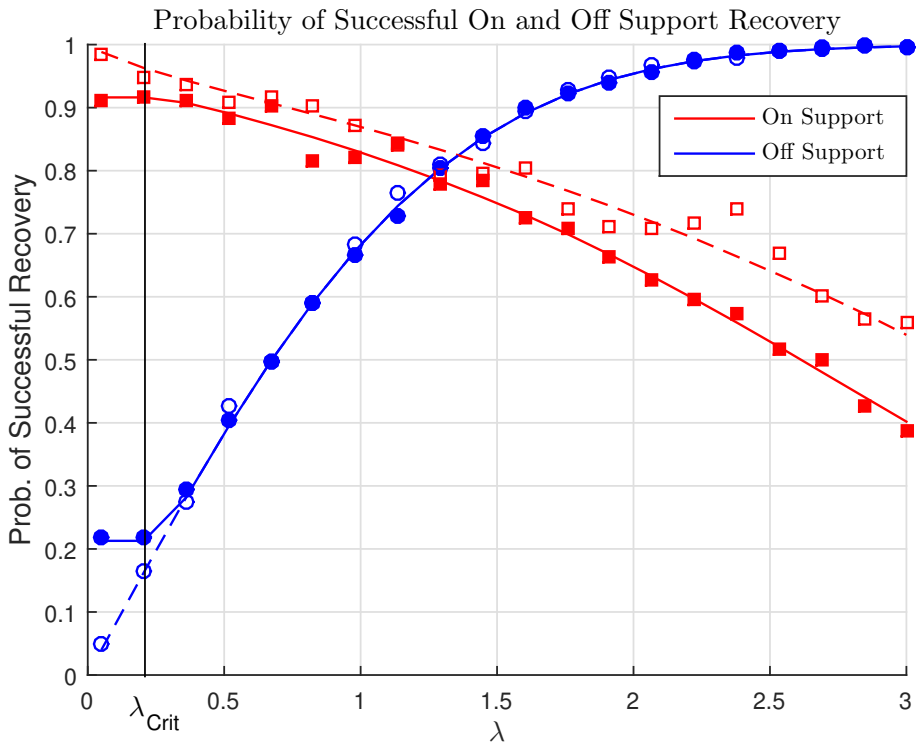


Figure 3.2: Probability of successful detection of support and off support entries as a function of λ for two different problem setup. The theoretical prediction (Solid and dashed lines) comes from Theorem10. For the simulations (Squares and Circles), we used $p = 256$, SNR= 0.5, $\epsilon = 10^{-3}$, $\rho = 0.1$ and the data are averaged over 5 independent realizations of problem. For solid lines and squares and circles, we used $\delta = 0.8$, while for dashed lines and empty squares and circles $\delta = 1.2$.

for λ such that $\delta > D(\hat{\lambda})$, $\hat{\lambda} = \max\{\lambda, \lambda_{crit}\}$, and

$$D(\lambda) = \rho \cdot \mathbb{E}[(H-\lambda)^2 \big| H > 0] + (1-\rho) \cdot \mathbb{E}[\eta^2(H,\lambda)].$$

Several remarks are in place here. First, when the threshold ϵ does not scale with σ and $\sigma \to 0$, then naturally the probability converges to one. The same is true as λ grows large, which is again expected. Finally, the term $D(\lambda)$ is known as the "Gaussian squared distance" in the relevant literature of noiseless linear inverse problems and admits insightful geometric interpretations [6, 73, 129]. In particular, $\min_{\lambda>0} D(\lambda)$ is known to be an asymptotically tight approximation of the exact phase transition threshold of ℓ_1-minimization [6, 153]. This can also be seen in (3.6) *which requires $\delta > \min_{\lambda>0} D(\lambda)$. Also, the formula is valid for λ such that $\delta > D(\hat{\lambda})$.*

For the on-support probability, we can show that it behaves as $\lim_{p\to\infty} \Phi_{\epsilon,On}(\hat{\beta}) \sim \mathbb{P}\left\{|\hat{\kappa}H + X_0| \geq \epsilon + \hat{\lambda}\hat{\kappa}\right\}$ for $\hat{\kappa} = \sigma/\sqrt{\delta - D(\hat{\lambda})}$. Similar remarks can be made.

Generalizations

The results of Section 3.2 on Lipschitz-like performance metrics for the LASSO extend to general regularized M-estimators. We outline the result here and defer a detailed treatment to the extended version of the paper. Regularized M-estimators solve

$$\hat{\beta} = \arg\min_{\beta} \mathcal{L}(\mathbf{y} - \mathbf{X}\beta) + \lambda f(\beta), \tag{3.7}$$

where $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}$ is a proper continuous convex loss function and f a convex regularizer. Clearly, the LASSO is an instance of (3.7) for $\mathcal{L}(\cdot) = \|\cdot\|_2$ and $f(\cdot) = \|\cdot\|_1$. Depending on the noise distribution and on the particular structure of β_0, different choices for the loss and regularizer function might be more appropriate [165]. Here, for simplicity we focus on separable functions of the form $\mathcal{L}(\mathbf{x}) = (1/p)\sum_{i=1}^{n} \ell(\mathbf{x}_i)$ (wlog, $\ell(0) = 0$) and $f(\mathbf{x}) = (1/p)\sum_{i=1}^{p} \tilde{f}(\mathbf{x}_i)$ for non-negative convex functions $\tilde{f}, \ell : \mathbb{R} \times \mathbb{R}$. Also, extending on the assumption of Section 3.1 we assume that the entries of the noise vector $\mathbf{z}$ are sampled iid from a distribution p_Z (not necessarily Gaussian). We only require the (rather mild) assumptions $\mathbb{E}\left[|\ell'(cH + \mathbf{Z})|^2\right] < \infty$ and $\mathbb{E}\left[\left|\tilde{f}'(cH + X_0)\right|^2\right] < \infty$ (see [165, Sec. 4] for details). where the expectations are taken over $Z \sim p_Z$, $X_0 \sim p_X$ and $H, G \sim \mathcal{N}(0,1)$. Here, $\tilde{f}'(x) = \sup_{s \in \partial \tilde{f}(x)} |s|$ where $\partial \tilde{f}(x)$ is the subdifferential of $\tilde{f}$ at x, and similar for ℓ'.

We also need to recall the notion of the Moreau Envelope function. For a convex function $\phi : \mathbb{R} \to \mathbb{R}$, the Moreau Envelope function of ϕ at x with parameter $\tau > 0$ is defined as $\mathrm{e}_f(x;\tau) := \min_y \frac{1}{2\tau}\|x-y\|^2 + \phi(y)$ We denote the optimal value of y above as $\mathrm{prox}_{\phi}(x,\tau)$. With these, [165, Thm. 4.1] characterizes the ℓ_2-reconstruction error of general regularized M-estimators as follows. There exists a unique α for which it holds in probability that $\lim_{p\to\infty} \frac{1}{p}\|\hat{\beta} - \beta_0\|^2 = \alpha^2$, where α is the solution to the following system of four equations in four unknowns (if such a solution does not exist then $\alpha = 0$):

$$\begin{cases} \alpha^2 = \mathbb{E}\left[\left(X_0 - \mathrm{prox}_f(\frac{\gamma}{\nu}H + X_0; \frac{\lambda}{\nu})\right)^2\right] \\ \gamma^2 = \delta\mathbb{E}\left[\left(\mathrm{e}'_{\ell}(\alpha H + \mathbf{Z}; \kappa)\right)^2\right] \\ \nu\alpha = \delta\mathbb{E}\left[H \cdot \mathrm{e}'_{\ell}(\alpha H + \mathbf{Z}; \kappa)\right] \\ \kappa\gamma = \mathbb{E}\left[H \cdot \mathrm{prox}_f(\frac{\gamma}{\nu}H + X_0; \frac{\lambda}{\nu})\right] \end{cases} \tag{3.8}$$

Imposing an extra requirement that the loss function be strongly convex and following similar steps as in the proof of Theorem 9 (see Section 3.3), we extend the result to Lipschitz performance metrics. In particular, for any Ψ satisfying Assumption 3 it holds in probability $\lim_{p\to\infty} \Psi(\hat{\beta} - \beta_0) = \Gamma_f(\gamma/\nu, \lambda/\gamma)$, where $(\alpha, \gamma, \nu, \kappa)$ are as in (3.8) and the function $\Gamma_f : \mathbb{R}_{>0} \times \mathbb{R}_{>0} \to \mathbb{R}$ is defined as $\Psi(X_0 - \text{prox}_f(\kappa H + X_0; \lambda) \xrightarrow{P} \Gamma_f(\kappa, \lambda)$.

3.3 Proof Outline

Our analysis is based on the recently developed Convex Gaussian Min-max Theorem (CGMT) framework [165, 171]. The following lemma specializes the general result of [165, Thm. 6.1] to the LASSO method in (3.1).

Theorem 11 (CGMT for LASSO) *Let $\mathbf{X}, \mathbf{z}, \beta_0$ be as in Section 3.1, $\mathbf{g} \in \mathbb{R}^n, \mathbf{h} \in \mathbb{R}^p$ have entries iid $\mathcal{N}(0,1)$ and all be independent of each other. Consider the optimizations:*

$$\min_{\mathbf{w}} \frac{1}{\sqrt{p}}\|\mathbf{z} - \mathbf{X}\mathbf{w}\|_2 + \frac{\lambda}{p}\|\beta_0 + \mathbf{w}\|_1, \tag{3.9}$$

$$\min_{\mathbf{w}} \max_{0<\gamma\leq 1} \gamma \frac{\|\mathbf{g}\|_2}{\sqrt{p}} \sqrt{\frac{\|\mathbf{w}\|_2^2}{p} + \sigma^2} - \gamma \frac{\mathbf{h}^T\mathbf{w}}{\sqrt{p}} + \frac{\lambda}{p}\|\beta_0 + \mathbf{w}\|_1. \tag{3.10}$$

Denote $\phi(\mathbf{g}, \mathbf{h})$ the optimal cost of the latter. Further, for an open set $\mathcal{S} \subset \mathbb{R}^n$ denote $\phi_{\mathcal{S}^c}(\mathbf{g}, \mathbf{h})$ its optimal cost when the minimization over $\mathbf{w}$ is now constrained over $\mathbf{w} \in \mathcal{S}$. Suppose there exists constants $\overline{\phi} < \overline{\phi}_{\mathcal{S}}$ such that $\phi(\mathbf{g}, \mathbf{h}) \xrightarrow{P} \overline{\phi}$ and $\phi_{\mathcal{S}^c}(\mathbf{g}, \mathbf{h}) \xrightarrow{P} \overline{\phi}_{\mathcal{S}^c}$. Then, for any minimizers $\mathbf{w}_\Phi$ and $\mathbf{w}_\phi$ of (3.9) and (3.10), respectively, the events $\{\mathbf{w}_\phi \in \mathcal{S}\}$ and $\{\mathbf{w}_\Phi \in \mathcal{S}\}$ occur with probability 1 in the limit of $p \to \infty$.

The minimization in (3.9) corresponds to the LASSO, only now the optimization variable is the error vector $\mathbf{w} := \beta - \beta_0$. To see how the theorem is applicable, suppose we are interested in showing $\Psi(\hat{\beta} - \beta_0) \xrightarrow{P} \alpha_*$ (eqv. $\Psi(\mathbf{w}_\Phi) \xrightarrow{P} d_*$) for some constant d_*. Then, we need to apply Theorem for the set $\mathcal{S}_\Psi := \{\mathbf{w} \big| |\Psi(\mathbf{w}) - d_*| < \delta\}$, where $\delta > 0$ is an arbitrarily small constant. The theorem suggests that d_* is the converging limit of $\Psi(\mathbf{w}_\phi)$, the solution to the *Auxiliary Optimization* (AO) in (3.10). The strategy becomes now clear [165]. First, we need to analyze the (AO) problem in (3.10) and find the converging limit of $\Psi(\mathbf{w}_\phi)$, say d_*. The second step consists of showing that the objective function of the (AO) strictly increases when

$\mathbf{w}$ is constrained such that $\Psi(\mathbf{w}_\phi)$ is far from d_*. Of course, the premise of this machinery is that these two tasks are much simpler to complete for the (AO) rather than for the original LASSO minimization [165]. The basics of these steps are outlined in the next two sections; details are deferred to the extended version of the paper.

Proving Theorem 9

It is shown in [164] that $\mathbf{w}_{\phi,i} = \beta_{0,i} - \eta(\kappa(\mathbf{g},\mathbf{h})\mathbf{h}_i + \beta_{0,i}, \kappa(\mathbf{g},\mathbf{h})\cdot\lambda)$ where $\kappa(\mathbf{g},\mathbf{h}) := \sqrt{\alpha^2(\mathbf{g},\mathbf{h})+\sigma^2}/\sqrt{\delta}$ and $\alpha(\mathbf{g},\mathbf{h})$ can be expressed as the minimizer of a random optimization problem (e.g. [164, eqn. (46)]). Moreover, it is shown in [164] that $\alpha(\mathbf{g},\mathbf{h})$ converges to α as this is defined in Lemma 11. Conditioning on the h.p. event that $\alpha(\mathbf{g},\mathbf{h}) \to \alpha$, we show that $\Psi(\mathbf{w}_{\phi,i}) \xrightarrow{P} \Psi(\beta_0 - \overrightarrow{\eta}(\kappa\mathbf{h}+\beta_0, \kappa\cdot\lambda))$. But the latter term converges to $\Gamma(\kappa,\lambda)$ by assumption, thus showing that the formula of Theorem 9 holds for the solution of the (AO) problem. It is also worth mentioning that the above argument shows the entries of $\mathbf{w}_\phi$ to be asymptotically iid.

Next, we need to verify that the objective function of the (AO) strictly increases when $\mathbf{w}$ is such that $|\Psi(\mathbf{w}) - \Gamma(\kappa,\lambda)| > 2\delta > 0$. First, if $\mathbf{w}$ is such, then using the result of the previous section it follows that $|\Psi(\mathbf{w}) - \Psi(\mathbf{w}_\phi)| > \delta$ with probability approaching 1. Then, the Lipschitzness property of Ψ implies that $\|\mathbf{w}-\mathbf{w}_\phi\|_2/\sqrt{p} > \delta/L$. The desired conclusion then follows by showing that the objective function in (3.10) is strongly convex in $\mathbf{w}$ and recalling optimality of $\mathbf{w}_\phi$. In particular, $\gamma\|\mathbf{g}\|_2\sqrt{\|\mathbf{w}\|_2^2/n+\sigma^2}$ is strongly convex with coefficient τ/p for some constant $\tau > 0$ (independent of p). Thus for the objective function of the optimization in (3.10) (say $F(\cdot)$) it holds $F(\mathbf{w}) \geq F(\mathbf{w}_\phi) + C\frac{\|\mathbf{w}-\mathbf{w}_\phi\|_2^2}{p} \geq F(\mathbf{w}_\phi) + \frac{\tau\delta}{L}$.

On the Support Recovery

The two metric defined in (3.2) do not satisfy the Lipschitz property. Nevertheless the proof of Theorem 10 follows from Theorem 9 when combined with a weak-convergence argument. Let $\psi : \mathbb{R} \to \mathbb{R}$ be arbitrary L-Lipschitz function. By Theorem 9, $(1/p)\sum_i \psi(\mathbf{w}_{\Phi,i}) \xrightarrow{P} \mathrm{Dist}_\psi(\kappa,\lambda)$. Since this holds for all Lipschitz functions, the empirical probability measure of $\mathbf{w}_\Phi$ converges [25, Thm. 25.8]. Hence, it follows (almost identically as in [25, Thm. 19]) that $\Psi_{\epsilon,on} \xrightarrow{P} \Gamma(\kappa,\lambda)$, where Γ as in Assumption 3 for the function $\Psi(\mathbf{w}) = 1/k\sum_{i=1}^k \mathbb{1}_{\{|\mathbf{w}_i-\beta_{0,i}|\geq\epsilon\}}$. Simplifying the "$\Gamma(\kappa,\lambda)$-term" yields the statement of Theorem 10.

Chapter 4

SPARSE COVARIANCE ESTIMATION FROM QUADRATIC MEASUREMENTS: A PRECISE ANALYSIS

The problem of covariance estimation[1] arises in many areas of modern statistics and information processing systems such as finance [5], when the underlying signal is high-dimensional, and/or the memory or computation power is limited. Therefore, in the current big data era, finding an efficient algorithm (in terms of sample complexity and memory requirements) to accurately estimate the covariance matrix is of great importance.

Recently, in [44, 48, 145], a framework for covariance estimation using phaseless, or energy measurements has been proposed. This type of measurements find applications in covariance estimation of data streams [44, 120], spectrum estimation of stochastic processes from energy measurements, noncoherent subspace detection from energy measurements [145], and others. Mathematically, given an unknown covariance matrix $\mathbf{\Sigma}_0 \in \mathbb{R}^{p\times p}$, the problem reduces to estimating $\mathbf{\Sigma}_0$ from a number of *m quadratic samples* of the form $\mathbf{a}_i^\top \mathbf{\Sigma}_0 \mathbf{a}_i + z_i$, $i = 1, \ldots, m$. Here, the measurement vectors $\mathbf{a}_i$'s are given and z_i's represent noise. [44, 48, 145] provide different convex and non-convex optimization algorithms for the recovery of $\mathbf{\Sigma}_0$. in this section, we provide a convex optimization formulation that is similar to that of [44], and precisely characterize its performance. Moreover, in the noiseless setting, our analysis framework provides the *exact phase transition*, i.e., the necessary and sufficient number of measurements for perfect recovery of the underlying covariance matrix. We are particularly interested in analyzing the underdetermined case, where we have $n < \frac{p(p+1)}{2}$ measurements and n denotes the number of variables. In such settings, the problem is ill-posed and the recovery is not possible unless the covariance matrix belongs to a low-dimensional set.

In many practical settings, the covariance matrix possesses certain structures. For instance, one common assumption is that the pairwise correlation has small magnitude for many pairs of entries of the underlying random vector, and hence, the covariance matrix has many (near)zero entries. in this section, we focus on the problem of recovery of a sparse covariance matrix. The convex optimization formu-

[1] This chapter in mainly based on the work in [2]

lation includes a regularization term that only enforces sparsity on the non-diagonal entries of the matrix. We then analyze the error performance of the solution. An order-wise analysis of a similar convex estimator for this problem exists in the works of [44, 48]. However, they provide upper bounds on the error of the estimated covariance matrix, with order-wise phase transitions. The key contribution of this section is to precisely characterize the error in the estimate and to present the necessary and sufficient number of measurements required for perfect recovery of the covariance matrix. In practice, having a precise theoretical understanding is extremely useful in designing the proper measurement settings.

The organization of this chapter is as follows. In Section 4.1 we introduce the main notations and mathematically set up the problem. Section 4.2 includes the main result of the paper followed by discussions and numerical simulations. In this section, we also present a major outcome of our main theorem which is the the characterization of the phase transition in the noiseless setting. Finally, Section 4.3 concludes the paper by describing the key steps of the proof of our main theorem.

4.1 Problem Setup

Notations

We gather here the basic notations that are used throughout the paper. Bold lower letters are reserved for vectors and upper letters are used for matrices. For a vector $\mathbf{v}$, v_i denotes its i^{th} entry and $\|\mathbf{v}\|$ is its ℓ_2-norm. $(\cdot)^\mathsf{T}$ is used to denote the transpose. $X \sim p_X$ implies that the random variable X has a density p_X, $\mathbb{E}[X]$ denotes its expected value. For a sequence of random variables $\{X^{(i)}\}_{i\in\mathbb{N}}$, $X^{(i)} \xrightarrow{\mathbb{P}} C$ indicates convergence in probability, i.e., $\lim_{i\to\infty}\mathbb{P}\{|X^{(i)} - C| > \epsilon\} = 0$. $\mathcal{N}(\mu, \Sigma)$ denotes the multivariate Gaussian distribution, with mean μ and covariance matrix Σ. $\mathbf{I}_d$ represents the identity matrix in dimension d. $\mathbb{S}_p$ refers to the set of $p \times p$ symmetric matrices. For $\mathbf{X} \in \mathbb{S}_p$, $\|\mathbf{X}\|_F$, $\mathrm{Tr}(\mathbf{X})$ and $\|\mathbf{X}\|_0$, respectively represent the Frobenius norm, the trace, and the number of non-zero entries. We also define $\|\mathbf{X}\|_1^- = \sum_{i\neq j} |X_{i,j}|$ as the ℓ_1-norm of the non-diagonal entries of matrix $\mathbf{X}$. The function $\psi : \mathbb{S}_p \to \mathbb{R}$ is said to be *Lipschitz* if $|\psi(\mathbf{X}) - \psi(\mathbf{Y})| \leq \frac{L}{p}\|\mathbf{X} - \mathbf{Y}\|_F$, for some constant $L > 0$. For a function $f : \mathcal{S} \to \mathbb{R}$, we define its proximal operator as following,

$$\mathrm{Prox}_f(\mathbf{v}, t) = \arg\min_{\mathbf{x}\in\mathcal{S}} \frac{1}{2t}\|\mathbf{x} - \mathbf{v}\|^2 + f(\mathbf{x}), \quad \forall \mathbf{v} \in \mathcal{S},\ t \in \mathbb{R}_+.$$

Setup

Following [44], we consider the problem of recovering an unknown symmetric matrix $\mathbf{\Sigma}_0 \in \mathbb{S}_p$, from n (noise-corrupted) quadratic measurements of the form,

$$y_i = \frac{1}{p}\,\mathbf{a}_i^T \mathbf{\Sigma}_0 \mathbf{a}_i + z_i = \frac{1}{p}\mathrm{Tr}\big(\mathbf{\Sigma}_0 \mathbf{a}_i \mathbf{a}_i^\mathsf{T}\big) + z_i, \quad i = 1 \ldots, n\,, \tag{4.1}$$

where $\{\mathbf{a}_i \in \mathbb{R}^p\}_{i=1}^n$, is the set of known measurement vectors, and $\mathbf{z} = [z_1, z_2, \ldots, z_m]^\mathsf{T} \in \mathbb{R}^n$ is the noise vector. Throughout this section, for our analysis purposes, we assume that $\mathbf{a}_i$'s are independently drawn from the Gaussian distribution with mean zero and covariance matrix $\mathbb{I}_p$, the noise vector is independent of all the measurement vectors, and has independent zero mean entries with variance σ^2. The normalization $\frac{1}{p}$ in (4.1) ensures that the measurement matrices $\mathbf{a}_i \mathbf{a}_i^\mathsf{T}$ are approximately unit-(Frobenius)norm.

Our result is asymptotic which assumes a fixed oversampling ratio, $\delta := \frac{2n}{p(p+1)} \in (0, \infty)$, while $p \to \infty$. Our interest is in studying the case where $\delta < 1$, in which the problem is ill-posed. Therefore, one needs to efficiently exploit the low-dimensional structure of the underlying covariance matrix, $\mathbf{\Sigma}_0$. Here, we focus on the setting where the covariance matrix is sparse, i.e., it has a few non-zero entries and define $\kappa = \frac{\|\mathbb{S}_p\|_0}{p^2}$ as the sparsity factor. This happens in practical applications when a large number of entries have small pairwise correlations. We analyze the following convex optimization formulation for recovery of a sparse covariance matrix,

$$\hat{\mathbf{\Sigma}} = \arg\min_{\mathbf{\Sigma} \in \mathbb{S}_p} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \frac{1}{p}\mathbf{a}_i^T \mathbf{\Sigma} \mathbf{a}_i \right)^2 + \frac{\lambda}{p^2} \|\mathbf{\Sigma}\|_1^-. \tag{4.2}$$

Recall from our notations in Section 4.1, the regularization term, $\|\mathbf{\Sigma}\|_1^-$, enforces sparsity only on the non-diagonal entries of $\mathbf{\Sigma}$, as the diagonals of a covariance matrix consist of positive entries. Due to the positive-definiteness of the covariance matrix, researchers often restrict the optimization program to the cone of positive semidefinite matrices [36, 44]. Here, we relax that constraint and let the feasible set contain all symmetric matrices. This relaxation not only simplifies the optimization program (by removing some constraints), but at the same time we will show that, when λ is tuned properly, it will provide a positive-definite estimate for $\mathbf{\Sigma}_0$, which indeed is equal to the solution of the semidefinite program.

Contribution

The optimization (4.2) resembles the well-known *LASSO* problem in the literature [176]. When the measurement matrix in LASSO is Gaussian, there exist

powerful tools such as CGMT [165] and AMP [62] that could precisely analyze its performance. For example, the CGMT framework has been successfully applied to analyze the performance in a number of applications including analysis of regularized M-estimators [165], massive MIMO [1, 166, 174], and PhaseMax in phase retrieval[54, 142]. Roughly speaking, for a convex optimization problem over an instance of a Gaussian process, the CGMT associates a secondary optimization with similar performance yet often much simpler to analyze.

Unfortunately, these frameworks do not apply to (4.2), because the measurement matrix is $\mathbf{a}_i\mathbf{a}_i^\mathsf{T}$ whose entries are neither Gaussian nor independent. To provide a precise analysis, we introduce a novel comparison lemma, that associates an equivalent optimization with our initial optimization problem (see Lemma 12). The equivalent optimization has i.i.d. Gaussian entries in its measurement, which makes it suitable to analyze via CGMT. Lemma 12 claims that under some conditions, the performance of these two optimizations is asymptotically the same. Therefore, analysis of the equivalent optimization via CGMT, characterizes the performance of our initial problem. To the best of our knowledge, this is the first work that introduces an equivalent optimization to analyze the performance of the problem of signal reconstruction from quadratic Gaussian measurements.

4.2 Main Results

The goal is to analyze the performance of our estimate, $\hat{\mathbf{\Sigma}}$ in (4.2). Let $\phi(\cdot)$ be a function which is Lipschitz and either convex or concave. Popular convex instances include $\phi(\mathbf{X}) = \frac{1}{p^2}\|\mathbf{X}\|_1$, and $\phi(\mathbf{X}) = \|\mathbf{X}\|_F/p$, and a concave one is $\phi(\mathbf{X}) = \lambda_{\min}(\mathbf{X})$. Our result is asymptotic in the sense that given a sequence of problem instances indexed by p, $\big(\mathbf{\Sigma}_0^{(p)}, \{\mathbf{a}_i^{(p)}\}_{i=1}^n, \mathbf{z}^{(p)}\big)_{p\in\mathbb{N}}$, it characterizes the error performance as the limiting behavior of the sequence $\{\phi(\hat{\mathbf{\Sigma}}^{(p)} - \mathbf{\Sigma}_0^{(p)})\}_{p\in\mathbb{N}}$. To streamline the notations, we often drop the superscript (p) when understood from the context.

Let $\Sigma_0 = [\sigma_{i,j}] \sim p_{\mathbf{\Sigma}_0}$, where $p_{\mathbf{\Sigma}_0}$ denote the distribution of the underlying covariance matrix, which also incorporates the sparsity structure of $\mathbf{\Sigma}_0$. For instance, one can assume that the off-diagonal entries of $\mathbf{\Sigma}_0$ are non-zero with probability $\kappa \in [0, 1]$, so that the average sparsity of the resulting covariance is $\kappa \cdot p(p-1)$. It will be observed in Theorem 12, that $p_{\mathbf{\Sigma}_0}$ plays a role in the performance of (4.2) via the summary functionals $F_\lambda(\cdot,\cdot)$, $G_\lambda(\cdot,\cdot)$, and $\Phi_\lambda(\cdot,\cdot)$, as follows.

Assumption 4 *Let* $\mathbf{\Sigma}_0 \sim p_{\mathbf{\Sigma}_0}$, *and* $\mathbf{H} = [h_{i,j}]$ *such that* $h_{i,j} = h_{j,i} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$,

for $1 \leq i \leq j \leq n$. *We say that Assumption 4 holds, if there exist functions* $F_\lambda : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}_+$, $G_\lambda : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$, *and* $\Phi_\lambda : \mathbb{R} \times \mathbb{R}^{>0} \to \mathbb{R}$ *such that for all* $s \in \mathbb{R}$ *and* $\tau > 0$,

$$\begin{aligned}
&\frac{1}{p^2}\ \|Prox_{\|\cdot\|_1^-}\big(\boldsymbol{\Sigma}_0 + s\mathbf{H}, \lambda\tau\big) - \boldsymbol{\Sigma}_0\|_F^2 \xrightarrow{\mathbb{P}} F_\lambda(s,\tau),\\
&\frac{1}{p^2}\ Tr\Big(\mathbf{H} \cdot Prox_{\|\cdot\|_1^-}\big(\boldsymbol{\Sigma}_0 + s\mathbf{H}, \lambda\tau\big)\Big) \xrightarrow{\mathbb{P}} G_\lambda(s,\tau), \quad \textit{and,}\\
&\phi\Big(Prox_{\|\cdot\|_1^-}\big(\boldsymbol{\Sigma}_0 + s\mathbf{H}, \lambda\tau\big) - \boldsymbol{\Sigma}_0\Big) \xrightarrow{\mathbb{P}} \Phi_\lambda(s,\tau),
\end{aligned} \tag{4.3}$$

where the convergence is in probability, over the distributions of $\boldsymbol{\Sigma}_0$ *and* $\mathbf{H}$.

Later in this section, we argue that Assumption 4 holds under very generic settings, and the the functions, F_λ, G_λ, and Φ_λ, capture the role of λ and $p_{\boldsymbol{\Sigma}_0}$ in our analysis of error performance. We now present the main result of the paper which characterizes the limiting behavior of $\phi(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_0)$ in terms of δ and σ.

Theorem 12 *Let Assumption 4 holds, and* $\hat{\boldsymbol{\Sigma}} \in \mathbb{S}_p$ *denote the solution to the convex optimization* (4.2)*, given n quadratic observations of the form* (4.1)*. Then, as* $p \to \infty$,

$$\phi(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_0) \xrightarrow{\mathbb{P}} \Phi_\lambda\left(\sqrt{\frac{\sigma^2 + 2\alpha^{\star 2}}{2\delta}}, \frac{\sqrt{\sigma^2 + 2\alpha^{\star 2}}}{2\beta^\star}\right), \tag{4.4}$$

where $(\alpha^\star, \beta^\star)$ *is the unique solution to the following system of non-linear equations with two unknowns,* α *and* β,

$$\begin{cases}
\alpha^2 = F_\lambda\left(\sqrt{\frac{\sigma^2+2\alpha^2}{2\delta}}, \frac{\sqrt{\sigma^2+2\alpha^2}}{2\beta}\right),\\
\beta = \sqrt{\sigma^2 + 2\alpha^2} - \sqrt{\frac{2}{\delta}} \cdot G_\lambda\left(\sqrt{\frac{\sigma^2+2\alpha^2}{2\delta}}, \frac{\sqrt{\sigma^2+2\alpha^2}}{2\beta}\right),
\end{cases} \tag{4.5}$$

and the functions $F_\lambda(\cdot,\cdot)$ *and* $G_\lambda(\cdot,\cdot)$ *are defined in* (4.3).

A few remarks are in place regarding Theorem 12:

[**Frobenius Norm of the Error**] Here, we show that the parameter $\alpha^\star$ in the Theorem 12, represents the limiting value of the Frobenius of the error of (4.2),

$\frac{1}{p}\|\hat{\mathbf{\Sigma}} - \mathbb{S}_p\|_F$. Since, by choosing $\phi(\mathbf{X}) = \frac{1}{p^2}\|\mathbf{X}\|_F^2$, from the definitions of $\Phi_\lambda(s,\tau)$ and $F_\lambda(s,\tau)$ in (4.3), we get

$$\Phi_\lambda(s,\tau) = F_\lambda(s,\tau). \tag{4.6}$$

Therefore, applying Theorem 12, the error performance can be computed as follows,

$$\begin{aligned}\frac{1}{p^2}\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0\|_F^2 \xrightarrow{p} \; & \Phi_\lambda\left(\sqrt{\frac{\sigma^2 + 2\alpha^{\star 2}}{2\delta}}, \frac{\sqrt{\sigma^2 + 2\alpha^{\star 2}}}{2\beta^\star}\right) \\ = \; & F_\lambda\left(\sqrt{\frac{\sigma^2 + 2\alpha^{\star 2}}{2\delta}}, \frac{\sqrt{\sigma^2 + 2\alpha^{\star 2}}}{2\beta^\star}\right) = \alpha^{\star 2}\,,\end{aligned} \tag{4.7}$$

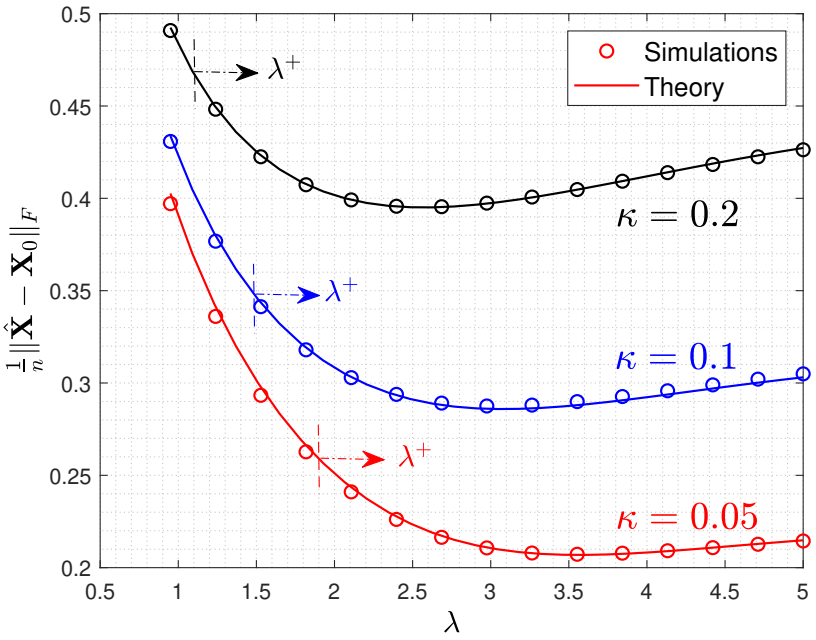


Figure 4.1: Performance of the optimization (4.2) with respect to $\phi(\mathbf{W}) = \|\mathbf{W}\|_F/p$, as a function of λ. Circles represent numerical simulations, and solid lines are theoretical predictions from Theorem 12. For simulations, we used $p = 120$, $\delta = .8$, $\mathbb{E}[z_i^2] = 1$, and three choices of sparsity factor; $\kappa = .05$ in red, $\kappa = .1$ in blue, and $\kappa = .2$ in black. The results are averaged over 80 random realizations of data. For $\lambda > \lambda^+$, the output of (4.2) will be positive definite.

where the last two equalities come from (4.6) and (4.5). Figure 4.1 demonstrates that the empirical result well matches the theoretical result derived from Theorem 12 . For our numerical simulations, the underlying covariance matrix, $\mathbf{\Sigma}_0$, is chosen to be a (uniformly) subsampled symmetric Gaussian matrix added by a multiple of the identity matrix (such that $\lambda_{\min}(\Sigma_0) = 0.1$ and $\kappa = \frac{\|\mathbb{S}_p\|_0}{p^2}$).

[**Tuning** λ] In order to guarantee that the solution of (4.2) is positive definite, we need to tune the regularization parameter λ. Here, we show that if λ is chosen such that the matrix $\hat{\mathbf{M}} := \text{Prox}_{\|\cdot\|_1^-}(\mathbf{\Sigma}_0 + s^\star \cdot \mathbf{H}, \lambda\tau^\star)$ is positive definite, then $\hat{\mathbf{S}}$ will also be positive definite. Here $s^\star = \sqrt{\frac{\sigma^2+2\alpha^{\star 2}}{2\delta}}$, and $\tau^\star = \frac{\sqrt{\sigma^2+2\alpha^{\star 2}}}{2\beta^\star}$, with $(\alpha^\star, \beta^\star)$ being the solution of (4.5).

To show this, we define the performance measure of optimization (4.2) to be the concave Lipschitz function $\phi(\mathbf{X}) = \frac{1}{\sqrt{p}}\lambda_{\min}(\mathbf{X} + \mathbf{\Sigma}_0)$. Applying Theorem 12 yields,

$$\frac{1}{\sqrt{p}}\lambda_{\min}(\hat{\mathbf{S}}) - \frac{1}{\sqrt{p}}\lambda_{\min}\left(\text{Prox}_{\|\cdot\|_1^-}\left(\mathbf{\Sigma}_0 + s^\star \cdot \mathbf{H}, \lambda\tau^\star\right)\right) \xrightarrow{\mathbb{P}} 0\,,$$

As a result, when λ is tuned properly such that $\hat{\mathbf{M}} := \text{Prox}_{\|\cdot\|_1^-}(\mathbf{\Sigma}_0 + s^\star \cdot \mathbf{H}, \lambda\tau^\star) \succeq \mathbf{0}$, then $\hat{\mathbf{S}}$ would be a positive definite matrix.

Computing the appropriate range of λ for an arbitrary distribution of $p_{\mathbf{\Sigma}_0}$ is beyond the scope of this thesis. For the case where $\mathbf{\Sigma}_0$ is chosen as the addition of a subsampled Gaussian plus a multiple of identity, using results from random matrix theory, we can show that $\lambda > C\log(\frac{p}{\lambda_{\min}(\mathbf{\Sigma}_0)})$, where C is a constant independent of p and $\mathbf{\Sigma}_0$, is sufficient for $\hat{\mathbf{S}}$ being PD. Figure 4.1 also specifies the range of λ under which the estimate $\hat{\mathbf{S}}$ is positive definite, for three choices of the sparsity factor, κ.

Phase Transition

Using the result of Theorem 12, we are able to characterize the phase transition of our convex optimization program, which provides us with the necessary and sufficient number of measurements for the perfect recovery of the underlying covariance matrix. Here, we assume that the measurements are *noiseless* ($\mathbf{z} = \mathbf{0}$). The goal is to identify $\delta_{\text{rec}} = \frac{2n_{\text{rec}}}{p(p+1)}$ that indicates the exact phase transition, such that , as $p \to \infty$, $\delta > \delta_{\text{rec}}$ is the necessary and sufficient condition for perfect recovery of $\mathbf{\Sigma}_0$. We state the following corollary (without proof) that characterizes the exact phase transition of the the optimization program (4.2):

Corollary 4 *Consider the optimization program* (4.2)*, given n* ***noiseless*** *measurements (*$\mathbf{z} = \mathbf{0}$*) of the form* (4.1)*. For a fixed oversampling ratio* $\delta = \frac{2n}{p(p+1)}$ *and the sparsity factor* $\kappa = \frac{\|\mathbf{S}_0\|_0}{p^2}$*, the optimization program* (4.2) *perfectly recovers the true covariance matrix (in the sense that* $\lim_{p\to\infty}\mathbb{P}\{||\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0|| > \epsilon\} = 0$*, for any fixed* $\epsilon > 0$*), if and only if* $\delta > \delta_{rec}$*, where* δ_{rec} *is the unique solution of the following*

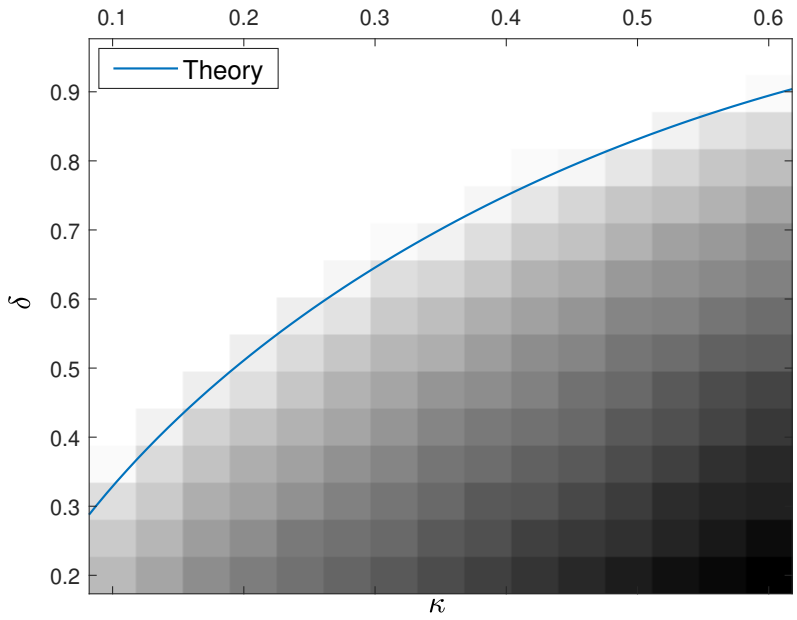


Figure 4.2: Phse transition regimes for the optimization (4.2), in terms of the oversampling ratio $\delta = \frac{2n}{p(p+1)}$, and the sparsity factor κ. Solid line comes from (4.8). For the empirical results, we used $p = 40$. The results are averaged over 20 independent realization of measurement vectors.

equation,

$$\delta\, Q^{-1}\left(\frac{2\delta-\kappa}{2-2\kappa}\right) = (1-\kappa)\varphi\left(Q^{-1}\left(\frac{2\delta-\kappa}{2-2\kappa}\right)\right). \tag{4.8}$$

Here, $\varphi(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$, *and* $Q(x) = \int_x^\infty \varphi(z)dz$, *represent the probability density and the tail distribution of the standard normal distribution, respectively.*

Corollary 4 specifies that the optimization (4.2) achieves perfect recovery w.p.a. 1, if and only if $n > \delta_{\text{rec}} \cdot p(p+1)/2$. Note that δ_{rec} is only a function of the sparsity factor κ, and is independent of other statistics of $\mathbf{S}_0$. Figure 4.2 illustrates validity of the Corollary 4. For numerical simulations, we used the same model to general $\mathbf{\Sigma}_0$, as for the Figure 4.1. As observed in the figure, the error becomes zero only in the regime where $n > \delta_{\text{rec}} \cdot p(p+1)/2$.

4.3 Proof Outline

The complete proof of this theorem can be found in the (15) and (16).

Here, we outline the fundamental ideas behind the proof of Theorem 12. Consider the optimization problem (4.2). To get a direct handle on the error term, it is

convenient to rewrite the optimization in terms of a new variable, $\mathbf{W} := \Sigma - \mathbf{\Sigma}_0$. Thus, (4.2) is equivalent to the following,

$$\min_{\mathbf{W}\in\mathbb{S}_p} \frac{1}{2n}\sum_{i=1}^{n}\left(z_i - \frac{1}{p}\mathrm{Tr}(\mathbf{W}\mathbf{A}_i)\right)^2 + \frac{\lambda}{p^2}\|\mathbf{W}+\mathbf{\Sigma}_0\|_1^-, \tag{4.9}$$

where $\mathbf{A}_i = \mathbf{a}_i\mathbf{a}_i^T$, and the goal is to analyze $\phi(\hat{\mathbf{W}})$. To have a simpler notation, we rewrite the optimization in a vectorized format. Let,

$$\mathbf{A} = \begin{bmatrix} \mathrm{Vec}^\intercal(\mathbf{A}_1) \\ \vdots \\ \mathrm{Vec}^\intercal(\mathbf{A}_m) \end{bmatrix},\ \mathbf{w} = \mathrm{Vec}(\mathbf{W}),\ \mathbf{x}_0 = \mathrm{Vec}(\mathbf{\Sigma}_0)\ , \tag{4.10}$$

where $\mathrm{Vec}(\mathbf{X})$ is the vectorization of $\mathbf{X}$ and $\mathrm{Mat}(\mathbf{x})$ is its inverse transform. For a vector $\mathbf{x}$, we also define $\|\mathbf{x}\|_1^- = \|\mathrm{Mat}(\mathbf{x})\|_1^-$. Using these notations, (4.9) can be rewritten as

$$\begin{aligned} \Psi(\mathbf{A}) &= \min_{\mathrm{Mat}(\mathbf{w})\in\mathbb{S}_n} \psi(\mathbf{A},\mathbf{w}) = \frac{1}{2n}\|\mathbf{z}-\frac{1}{p}\mathbf{A}\mathbf{w}\|^2 + \frac{\lambda}{p^2}\|\mathbf{w}+\mathbf{x}_0\|_1^-, \\ \hat{\mathbf{W}}(\mathbf{A}) &= \mathrm{Mat}\big(\arg\min_{\mathrm{Mat}(\mathbf{w})\in\mathbb{S}_p} \phi(\mathbf{A},\mathbf{w})\big)\ . \end{aligned} \tag{4.11}$$

We proceed onward by analyzing the optimization (4.11) which is similar to the popular LASSO problem. As stated before, the main bottleneck in analyzing this optimization is the fact that the entries of $\mathbf{A}$ are not i.i.d. Gaussian. Instead, we prove the desired indirectly, via the following two steps. First, we show that the properties of $\Psi(\cdot)$ are preserved asymptotically, as we replace $\mathbf{A}$, with a carefully-designed *Gaussian* matrix $\mathbf{B}$ with *independent* entries. In other words, $\hat{\mathbf{W}}(\mathbf{A})$ and $\hat{\mathbf{W}}(\mathbf{B})$ have the same asymptotic performance. Consequently, we utilize the CGMT framework to analyze $\hat{\mathbf{W}}(\mathbf{B})$. Leaving some technical details for section (15) and (16), the mechanics are easy to explain and provide valuable intuition regarding our approach.

Step 1. We start by introducing some new notations. Let $\mathbf{G}_i$, for $i = 1, 2, \ldots, n$, be a *symmetric* matrix whose diagonal and (upper) non-diagonal entries are drawn independently from distributions $\mathcal{N}(1,2)$ and $\mathcal{N}(0,1)$, respectively. Also, define,

$$\mathbf{B} = \begin{bmatrix} \mathrm{Vec}^\intercal(\mathbf{G}_1) \\ \vdots \\ \mathrm{Vec}^\intercal(\mathbf{G}_n) \end{bmatrix} \in \mathbb{R}^{n\times p^2}\ . \tag{4.12}$$

The following comparison lemma forms the heart of our proof, allowing us to have the same performance replacing $\mathbf{A}$ with $\mathbf{B}$ in (4.11).

Lemma 12 *Consider the functions* $\Psi(.)$ *and* $\hat{\mathbf{W}}(.)$ *defined in* (4.11) *and the random matrices* $\mathbf{A}$ *and* $\mathbf{B}$ *from* (4.10) *and* (4.12)*. Let the assumptions in Theorem 12 hold and both parameters* $\psi(\mathbf{W}(\mathbf{A}))$ *and* $\psi(\mathbf{W}(\mathbf{B}))$ *converge in probability as* $p \to \infty$*. Then,* $\psi\big(\mathbf{W}(\mathbf{A})\big) - \psi\big(\mathbf{W}(\mathbf{B})\big) \xrightarrow{\mathbb{P}} 0$.

Lemma 12 essentially states that replacing matrices $\mathbf{a}_i\mathbf{a}_i^\mathsf{T}$, for $i = 1, 2, \ldots, n$, in the optimization (4.9) with Gaussian matrices $\mathbf{G}_i$ does not alter the performance measure. It is worth noting that the stated result is only valid when all the stated conditions hold. We defer the technical details in the proof to (15) and (16), but once it is established we only need to analyze the performance of $\mathbf{W}(\mathbf{B})$.

Step 2. We utilize the CGMT framework to analyze the performance of $\mathbf{W}(\mathbf{B})$.

Lemma 13 *Let* $(\alpha^\star, \beta^\star)$ *be the unique solution to the system of equations* (4.5)*, then as* $p \to \infty$*, we have*

$$\phi(\mathbf{W}(\mathbf{B})) \xrightarrow{\mathbb{P}} \Phi\left(\sqrt{\frac{\sigma^2 + 2\alpha^{\star 2}}{2\delta}}, \frac{\sqrt{\sigma^2 + 2\alpha^{\star 2}}}{2\beta^\star}\right). \tag{4.13}$$

To show this lemma we need to apply the CGMT. We refer the interested reader to section V.D. of [165]. So in the next section, we will focus on the proof of the Lemma 12.

Proof of Lemma 12

Proving that $\psi(\hat{\mathbf{W}}(\mathbf{A}))$ and $\psi(\hat{\mathbf{W}}(\mathbf{B}))$ converge to the same value is one step away from proving that for every $C > 0$

$$\left| \min_{\substack{\mathrm{Mat}(\mathbf{w}) \in \mathbb{S}_p \\ \frac{1}{2p^2}\|\mathbf{w}\|^2 \le C}} \psi(\mathbf{A}, \mathbf{w}) - \min_{\substack{\mathrm{Mat}(\mathbf{w}) \in \mathbb{S}_p \\ \frac{1}{2p^2}\|\mathbf{w}\|^2 \le C}} \psi(\mathbf{B}, \mathbf{w}) \right| \xrightarrow{p} 0 \,. \tag{4.14}$$

Once we have this lemma, if $\frac{1}{2p^2}\|\hat{\mathbf{W}}(\mathbf{A})\|_F^2$ and $\frac{1}{2p^2}\|\hat{W}b(\mathbf{B})\|_F^2$ converge to different values s_1 and s_2, choosing $C = (s_1 + s_2)/2$ in (4.14) results in a contradiction. Thus $\frac{1}{2p^2}\|\mathbf{W}(\mathbf{A})\|_F^2$ and $\frac{1}{2p^2}\|\mathbf{W}(\mathbf{B})\|_F^2$ converge to the same value. Then, using Lipschitsness and convexity of $\phi(.)$ and the same set of arguments as in the Section IV-B of [4] shows that $\psi(\mathbf{W}(\mathbf{A}))$ and $\psi(\mathbf{W}(\mathbf{B}))$ converge to the same value.

It remains to prove (4.14). We define

$$\eta(x) = \begin{cases} |x| & \text{if } |x| > 1 \\ \frac{3}{8} + \frac{6}{8}x^2 - \frac{1}{8}x^4 & \text{o.w.} \end{cases}. \tag{4.15}$$

$\eta(.)$ can be also applied to matrices, where it acts element-wise. It is twice differentiable and convex. Besides, $t \cdot \eta(\mathbf{X}/t)$ is jointly convex in $\mathbf{X}$ and $t > 0$, since it is the perspective function of $\eta(\cdot)$. Next, we introduce a new optimization,

$$\Psi_{t,\epsilon}(\mathbf{A}) = \min_{\mathbf{w}\in\mathbb{S}_p} \frac{1}{2n}\|\mathbf{z} - \frac{1}{p}\mathbf{A}\mathbf{w}\|^2 + \frac{\lambda t}{p^2} \cdot \eta\big(\frac{\mathbf{w}+\mathbf{x}_0}{t}\big) + \frac{\epsilon}{2p^2}\|\mathbf{w}\|^2 \ . \tag{4.16}$$

This function is convex in t and concave in ϵ. Furthermore,

$$\inf_{t>0}\sup_{\epsilon>0} -C\epsilon + \Psi_{t,\epsilon}(\mathbf{A}) = \min_{\substack{\mathbf{w}\in\mathbb{S}_p \\ \frac{1}{2p^2}\|\mathbf{w}\|^2 \le C}} \psi(\mathbf{A},\mathbf{w}) \ . \tag{4.17}$$

Thus, if we show that

$$\big|\Psi_{t,\epsilon}(\mathbf{A}) - \Psi_{t,\epsilon}(\mathbf{B})\big| \xrightarrow{p} 0 \ , \tag{4.18}$$

we can apply Lemma 14 on (4.17), twice, to prove (4.14). The key ingredient of proving (4.18) is the Lindeberg replacement principle and the strong convexity of the regularizer in (4.16). We omit the details due to the limited space, but such similar results has been shown in [130, 131]. We now present lemma bellow which is also known as the convexity lemma in the literature [136].

Lemma 14 *Consider a series of convex functions $f_p : \mathbb{R}^{>0} \to \mathbb{R}$ that converges point-wise to the function $f : \mathbb{R}^{>0} \to \mathbb{R}$. Besides, there exists $M > 0$ such that for all $x > M$, we have $f(x) > \inf_{s>0} f(s)$. Then $f(.)$ is also convex and $\inf_{s>0} f_p(s) \xrightarrow{p} \inf_{s>0} f(s)$.*

Chapter 5

SCALABLE COVARIANCE ESTIMATION IN GRAPHICAL MODELS WITH PROVABLE GUARANTEES

5.1 Introduction

variance matrices play a fundamental role in behavioral analysis of multivariate random variables in a variety of fields including finance and economics, engineering, and environmental and physical sciences. In modern such inference applications, one is faced with the problem of estimating covariance matrices associated with data, of a very large dimension p, from a few (and potentially less than p), number of observations n. Also, it is often the case that the true unknown matrix possesses some low-dimensional structure. This might be a property directly associated with the covariance matrix (examples including sparse or approximately low-rank covariance matrices), but it can also arise in an indirect form. For example, a very well-encountered structural model, which is relevant to graphical models for Gaussian random variables, is one in which the inverse of the covariance matrix (rather than itself), also termed the *precision matrix*, is sparse. Here, the sparsity pattern of the concentration matrix implies the structure of the associated graph in the Gaussian Markov Random Field (GMRF), and is thus critical to estimate.

Overall, modern inference procedures for covariance estimation need to have the following favorable properties:

- on the *statistical/theoretical* side, it is important that they *provably* reveal the underlying structures of the true desired matrices (such as the support pattern of the precision matrix) while having access to only few number of observations.
- on the *computational* side, it is critical that their computational complexity *scales* with the increasing problem dimensions, thus allowing to solve practical instances in which n and p are on the range of (at least) a few hundrends. In a similar flavor, algorithms that allow performing operations in a *parallel* fashion on different machines thus speeding up the total performance are also desirable.

in this section, we consider the classical problem of estimating the *sparse* precision matrix of a multivariate zero-mean random variable, given n iid observations $\{\mathbf{x}_i\}_{i=1,\ldots,n} \in \mathbb{R}^p$. While this problem has attracted lots of attention over the past decade or so, and several algorithms have been proposed and analyzed in the relevant literature, it appears that none of them enjoys *both* the computational *and* the theoretical features discussed above. We propose a *novel algorithm* that combines all these; specifically, it (i) is *scalable*, (ii) is *parallelizable*, and (iii) *provably estimates the desired structure of the underlying graphical model from only a few number of observations*.

Background and Motivation

We consider the problem of estimating the covariance matrix Σ of a multivariate zero-mean random variable, given n i.i.d. observations $\{\mathbf{x}_i\}_{i=1,\ldots,n} \in \mathbb{R}^p$. We focus on the regime of high-dimensions in which both n and p are large. Of particular, interest is the case of limited number of observations $n \ll p$. In this high-dimensional setting the classical *sample covariance estimator* $\mathbf{R} = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\mathsf{T}$ has been shown to perform poorly[92][93]. Besides, it cannot accommodate for any prior knowledge on the structure of Σ. Many different kinds of structures have been considered in the literature. For example, for the case of banded covariance (or concentration) matrices, where the entries decrease as a function of their distance from the diagonal, popular estimators include banded estimators [195][24][32], and *shrinkage estimators* [102][76].

in this section, we focus on a widely-studied model where the *concentration matrix* Σ^{-1} is sparse. This model shows up in graphical models for Gaussian random variables. It is classically known that in graphical models, the sparsity pattern of the concentration matrix implies the structure of the associated graph in the Gaussian Markov random field (GMRF).

ℓ_1-penalized log det-optimization. A popular approach to estimate a sparse precision matrix is via solving the following *convex* optimization program:

$$\hat{\mathbf{\Omega}} = \arg\min_{\mathbf{\Omega}>0} \operatorname{Tr}(\mathbf{\Omega}\mathbf{R}) - \log\det(\mathbf{\Omega}) + \lambda \|\mathbf{\Omega}^-\|_1, \tag{5.1}$$

where recall that $\mathbf{R} := \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\mathsf{T}$ is the sample correlation matrix, and $\|\mathbf{\Omega}^-\|_1 = \sum_{i\neq j} |\mathbf{\Omega}_{ij}|$ is the ℓ_1-norm of the off-diagonal entries of its argument. When the observations are generated from a Gaussian distribution, then the objective function

in (5.1) is nothing but the ℓ_1-penalized (negative) log-likelihood function. The ℓ_1-regularization is known to promote sparse solutions under several settings [68].

On the analysis side, taking advantage of the convex nature of (5.1), the estimator $\hat{\mathbf{\Omega}}$ has been well-analyzed in the literature and it has been proved to enjoy favorable properties. When the $\mathbf{x}_i$'s are Gaussian, Rothman et al. [138] showed, under standard assumptions on Σ, that $\|\hat{\mathbf{\Omega}} - \Sigma^{-1}\|_F^2 \approx O(\sqrt{\frac{(s+p)\log p}{n}})$, where s denotes the number of non-zero non-diagonal entries in the concentration matrix; this implies consistency in a Frobenius-norm sense, as long as $n > O((s+p)\log p)$. Later in 2008, Lam and Fan [100] further showed that (5.1) additionally recovers the *sparsity pattern* of the concentration matrix as long as $s = O(\sqrt{p})$ and $n = \Omega((s+p)\log p)$. Since then, Ravikumar et. al. [134] have extended these results beyond Gaussians to a general class of random variables with appropriate tail bounds (which includes sub-Gaussians and random variables with bounded moments), under additional *coherence assumptions* on Σ. Importantly, they proved consistency with respect to the stronger notion of the *max-norm* $\|\hat{\mathbf{\Omega}} - \Sigma^{-1}\|_{\max} = \max_{i,j} |(\hat{\mathbf{\Omega}} - \Sigma^{-1})_{i,j}|$.

These rich set of performance guarantees apply to the solution $\hat{\mathbf{\Omega}}$ of the convex optimization (5.1). However, the computational task of obtaining (ϵ-close approximations of) $\hat{\mathbf{\Omega}}$ via standard generic convex solvers, such as interior point methods and other second-order-methods [28], does not scale well with the problem dimensions, which makes solving (5.1) prohibitive in any practical scenario where p is on the order of even a few hundreds!

GLASSO. The computational challenge of scaling the minimization in (5.1), has led to significant research activity on deriving fast alternative algorithms [49][199][75]. A very popular method towards this direction is the Graphical LASSO (GLASSO) [75]. GLASSO is an iterative algorithm, which starts with an initialization (say) $\bar{\mathbf{\Sigma}}$, and at each step, it solves a specific LASSO problem (aka ℓ_1-regularized least-squares) with the goal of updating a specific row/column of this matrix. After p iterations the entire matrix is updated once and the process continues until convergence (in the sense that the updated matrix remains in some ϵ-neighborhood of the previous estimate). While the details of the updates are not important for our discussion, it should be emphasized that solving LASSO problems can be done very efficiently using off-the-shelf solvers (e.g. [74]). Hence, GLASSO is scalable and is used in practice, instead of (5.1).

Unfortunately, despite its obvious algorithmic advantages when compared to (5.1), no matching theoretical guarantees for GLASSO exist in the literature. This is

despite the fact that GLASSO can be interpreted as a block-wise coordinate descent approach for solving (5.1). For example, the answers to the following questions are unknown: "What is the rate of convergence of GLASSO?", or "If you were to run the algorithm for (say) p steps (during which the entire initial matrix $\bar{\mathbf{\Sigma}}$ is updated once), how good an estimate is obtained?". It should also be mentioned that more recent proposed extensions of GLASSO such as the P-GLASSO and DP-GLASSO [111], although apparently faster, still suffer by the lack of any analytical guarantees on their rate of convergence and on the performance after a fixed number of iterations.

Adding to this, note that even though GLASSO is scalable, it is *not* parallelizable since solving the LASSO at each iteration depends on the previous one.

Contribution

In this chapter, we propose a new algorithm that combines the virtues of both GLASSO and of the convex estimator in (5.1), i.e., it is fast and *scalable* with problem dimensions, and it *provably attains the same order-wise statistical guarantees as* (5.1). Moreover, it involves solving only p *independent* LASSO problems which can be performed on different machines; thus it is also *parallelizable*.

The algorithm starts with a *shrinkage estimator* $\bar{\mathbf{\Sigma}}$ of Σ[1]. Then, it solves p *independent* LASSO problems. Each LASSO problem uses $\bar{\mathbf{\Sigma}}$ to obtain an estimate of the i^{th} row/column of the concentration matrix. These p estimates are combined in the last step to obtain a final estimate $\hat{\mathbf{\Omega}}$ of the concentration matrix $\mathbf{\Omega} = \Sigma^{-1}$. We give the details in Section 5.2.

The scalability of the algorithm is obvious: its computational complexity is the same as that of solving p LASSO problems. On top of that (something that is not true for GLASSO), the LASSO problems are independent of each other, thus they can be run over parallel machines to achieve further improved computational performance. Finally, when the random variables are Gaussians, we prove recovery bounds that coincide with state of the art corresponding results on the performance of the convex program in (5.1). In particular, we show that with the correct tuning of the input parameters, the algorithm exactly recovers the zero-pattern of the concentration matrix, which here corresponds to the structure of the underlying Gaussian graphical model. Furthermore, we provide consistency guarantees on the max-norm as in [134].

[1] Other appropriate initializations are also possible, we discuss these in later sections.

Paper Organization

The rest of the paper is organized as follows. In Section 5.2 we describe how the proposed algorithm operates, and we present its computational features and accompanying theoretical results on its statistical performance. A more detailed discussion, a thorough comparison of the results to the relevant literature and numerical simulations are included in Section 5.3. The proofs, and also an illustration of the intuition behind the algorithm are deferred to the Appendix.

5.2 The Algorithm: Computational & Statistical Guarantees

For convenience, denote the concentration matrix as $\mathbf{\Omega} := \Sigma^{-1}$ and $[p] = \{1, \ldots, p\}$. For a matrix $\mathbf{M}$ and a vector $\mathbf{v}$ we denote $\mathbf{M}_{k,\ell}$ and $\mathbf{v}_k$ the $(k,\ell)^{th}$ entry of $\mathbf{M}$ and the k^{th} entry of $\mathbf{v}$, respectively.

FGL Algorithm

We call our algorithm the Fast Graphical-LASSO (FGL) algorithm. In this section, we describe how the algorithm operates.

Initialization. Let $\mathbf{R} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\mathsf{T}$ be the sample covariance matrix and $\mathbf{D_R}$ be a diagonal matrix with the same diagonal entries as $\mathbf{R}$. For a constant $0 \leq \mu \leq 1$, the *shrinkage estimator* $\mathbf{R}_\mu$ of Σ is defined as

$$\mathbf{R}_\mu := \mu \mathbf{D_R} + (1-\mu)\mathbf{R}. \tag{5.2}$$

Of course, $\mathbf{R}_0 = \mathbf{R}$. Further note that $\mathbf{R}_\mu$ is always positive definite for all $0 < \mu \leq 1$, even in the regime of few observations $n < p$. We will see that this property is critical for the satisfactory performance of the algorithm in the presence of few observations.

FGL takes as input a parameter μ and utilizes $\bar{\mathbf{\Sigma}} = \mathbf{R}_\mu$ as an estimate of Σ. Our main theorem specifies appropriate values for tuning μ that guarantee good statistical performance. (For example, we will see that naively setting $\mu = 0$, eqv. initializing the FGL with just $\mathbf{R}$, guarantees good performance only in the case $n > p$).

p independent LASSOs. The main body of the algorithm is solving p independent LASSO optimization problems. Each one of them uses the initialization $\mathbf{R}_\mu$ and outputs an estimate of a certain row/column of the precision matrix (a total p of them).

Each one of the p main operations of the algorithm is independent of one other, but can all be described in a common language. For this, fix any $i = 1, \ldots, p$.

Construct matrix $\bar{\mathbf{\Sigma}}^{(i)}$ by permuting the i^{th} and p^{th} rows and columns of $\bar{\mathbf{\Sigma}}$ such that the i^{th} (resp. p^{th}) row and column of $\bar{\mathbf{\Sigma}}^{(i)}$ is the p^{th} (resp. i^{th}) row and column of $\bar{\mathbf{\Sigma}}$ [2]

Consider the following partition of $\bar{\mathbf{\Sigma}}^{(i)}$:

$$\bar{\mathbf{\Sigma}}^{(i)} = \begin{bmatrix} \bar{\mathbf{\Sigma}}_{11} & \bar{\boldsymbol{\sigma}}_{12} \\ \bar{\boldsymbol{\sigma}}_{12}^{\mathsf{T}} & \bar{\sigma}_{22} \end{bmatrix}, \tag{5.4}$$

where we have suppressed the dependence of the elements of the partition on the index i in order to keep the notation simple. Using this notation, and for an input parameter $\lambda > 0$, solve the following LASSO problem,

$$\hat{\beta}^{(\mathrm{i})} := \arg\min_{\beta} \frac{1}{2}\beta^{\mathsf{T}}\bar{\mathbf{\Sigma}}_{11}\beta + \bar{\boldsymbol{\sigma}}_{12}^{\mathsf{T}}\beta + \lambda\|\beta\|_1. \tag{5.5}$$

Observe that $\hat{\beta} \in \mathbb{R}^{p-1}$. Use this to construct $\boldsymbol{\omega}^{(i)} \in \mathbb{R}^p$ as follows

$$\boldsymbol{\omega}_k^{(i)} := (\bar{\sigma}_{22} + \hat{\beta}^{\mathsf{T}}\bar{\mathbf{\Sigma}}_{11}\hat{\beta})^{-1} \times \begin{cases} 1, & k = i, \\ \hat{\beta}_i^{(i)}, & k = p, \\ \hat{\beta}_k^{(i)}, & k \notin \{i, p\}, \end{cases} \qquad k = 1, \ldots, p. \tag{5.6}$$

Output. The FGL algorithm outputs an estimate $\hat{\mathbf{\Omega}}$ of the precision matrix $\mathbf{\Omega}$, based on the previously computed vectors $\boldsymbol{\omega}^{(i)}$, $i = 1, \ldots, p$ as follows:

$$\hat{\mathbf{\Omega}}_{k,\ell} := (\boldsymbol{\omega}_\ell^{(k)} + \boldsymbol{\omega}_k^{(\ell)})/2, \qquad k, \ell = 1, \ldots, p. \tag{5.7}$$

All these are outlined in Algorithm 1 below.

Algorithm 1 FGL Algorithm

Input: Observations $\{\mathbf{x}_j \in \mathbb{R}^p\}_{j=1,\ldots,n}$. Parameters $0 \le \mu \le 1$, $\lambda > 0$.
Output: Estimate $\hat{\mathbf{\Omega}}$ of the precision matrix $\mathbf{\Omega} = \Sigma^{-1}$.

Set $\bar{\mathbf{\Sigma}} = \mathbf{R}_\mu = \mu\mathbf{D}_{\mathbf{R}} + (1-\mu)\mathbf{R}$ as in (5.2).
for all $i = 1, \ldots, p$ **do**
Solve the LASSO problem in (5.5) to find $\hat{\beta}^{(i)}$, where $\bar{\mathbf{\Sigma}}_{11}$ and $\bar{\boldsymbol{\sigma}}_{12}$ are as in (5.4).
Use $\hat{\beta}^{(i)}$ to form $\boldsymbol{\omega}^{(i)}$ as in (5.6).
end for
Using the $\boldsymbol{\omega}^{(i)}$'s, construct $\hat{\mathbf{\Omega}}$ as in (5.7).

[2] Formally, define permutation matrices $\mathbf{P}^{(i)}, i = 1, \ldots, p$ such that

$$\mathbf{P}_{k,\ell}^{(i)} = \begin{cases} 1, & k = \ell \notin \{i, p\}, \\ 1, & (k, \ell) \in \{(i, p), (p, i)\}, \\ 0, & \text{else.} \end{cases} \tag{5.3}$$

Then, $\bar{\mathbf{\Sigma}}^{(i)} = \mathbf{P}^{(i)}\bar{\mathbf{\Sigma}}\mathbf{P}^{(i)}$.

Computational Performance

It is clear that the main computational burden of the FGL algorithm is that of solving p LASSO problems. This makes the algorithm scalable. Moreover, observe that the LASSO problems are all independent of each other. The outputs $\hat{\beta}^{(i)}$ of each are combined only at the last step of the algorithm to form the final estimate $\hat{\mathbf{\Omega}}$. Thus, each LASSO problem can be solved separately on a different machine. Each machine needs only have access to the matrix $\bar{\mathbf{\Sigma}}$ computed at the first stage of the algorithm. (See however Section 5.3 where it is shown that storing $\bar{\mathbf{\Sigma}}$ is not necessary and all operations can be performed via sole access to the observation vectors, when $n > p$).

Statistical Performance

Aside from the obvious computational virtues discussed above, it is further shown in this section that the FGL algorithm enjoys provable performance guarantees.

As is typical, the guarantees require some conditions on the true unknown covariance matrix Σ. We start with these and state the main result in Theorem 13. The theorem holds for the case of Gaussian random variables, but we expect the result to generalize to wider classes, such as sub-gaussians and variables with bounded moments. We leave these to a future long version of the paper.

Some notation & Assumptions Let $\mathbf{M} \in \mathbb{R}^{p\times p}$. For sets $\mathcal{S}, \mathcal{R} \subseteq [p]$, $\mathbf{M}_{\mathcal{S},\mathcal{R}}$ denotes a sub-matrix of $\mathbf{M}$ with entries $(\mathbf{M}_{i,j})_{i\in\mathcal{S},j\in\mathcal{R}}$. Also, we write $\|\mathbf{M}\|_{\max} = \max_{i,j} |M_{i,j}|$ for the maximum element-wise norm of $\mathbf{M}$ and $\|\mathbf{M}\|_\infty = \max_i \sum_{j=1}^p |\mathbf{M}_{i,j}|$ for the induced infinity norm.

Our analysis keeps explicit track of the positive quantities κ, γ, $\bar{\lambda}$ and $\underline{\lambda}$ defined below, so that they can scale in a non-trivial manner with the problem dimension p.

- First we define the parameter $\kappa > 0$ that corresponds to the diagonal dominance of the covariance matrix. We have,

$$\forall i \in [p], \quad \sum_{j\neq i} |\mathbf{\Omega}_{i,j}| \leq \kappa \mathbf{\Omega}_{i,i}. \tag{5.8}$$

- We also define $\gamma > 0$ such that

$$\forall i \in [p], \quad \|(\Sigma_{\mathcal{S}_i,\mathcal{S}_i})^{-1}\|_\infty \leq \gamma,$$

corresponding to the ℓ_∞-norm of the inverse sub-covariance matrices.

- We use $0 < \underline{\lambda} \leq \bar{\lambda} < \infty$ to denote the minimum and maximum eigenvalues of Σ,

$$\lambda_{max}\{\Sigma\} \leq \bar{\lambda}$$
$$\lambda_{min}\{\Sigma\} \geq \underline{\lambda}.$$

We require the following assumption on the covariance matrix. See Section 5.3 for a discussion of its interpretation.

Assumption 5 *There exists a constant* $0 < \alpha < 1$ *such that,*

$$\forall i \in [p], \quad \|\Sigma_{\bar{\mathcal{S}}_i,\mathcal{S}_i}(\Sigma_{\mathcal{S}_i,\mathcal{S}_i})^{-1}\|_\infty \leq 1 - \alpha.$$

Main Result. We are now ready to present our main theorem characterizing the statistical performance of the FGL algorithm in terms of both support recovery and max-norm consistency.

Theorem 13 (Zero-pattern & max-norm Guarantees) *Let the observations* $\{\mathbf{x}_i\}_{i=1,\ldots,n}$ *follow a zero mean Gaussian distribution of covariance* $\Sigma \in \mathbb{R}^{p\times p}$ *that further satisfies Assumption 5. Let* $\mathbf{\Omega} = \Sigma^{-1}$*, and* $\hat{\mathbf{\Omega}}$ *be the output of the FGL Algorithm 1 with input parameters* $\mu = \mu^\star$ *defined in* (5.11) *and* $\lambda = \frac{8-2\alpha}{\alpha}(\kappa+1)\bar{\lambda}\sqrt{\frac{\tau \log p}{n}}$*. Then, for any* $\tau > 2$*, the following statements hold with probability at least* $1 - 5/p^{\tau-2}$*.*

(i) $\mathrm{Supp}(\hat{\mathbf{\Omega}}) \subseteq \mathrm{Supp}(\mathbf{\Omega})$.

(ii) Further suppose that

$$n > M\left(\frac{(4-\alpha)\bar{\lambda}\gamma(C_1+\kappa)}{\alpha\underline{\lambda}}\right)^2 d^2\tau \log p, \tag{5.9}$$

for some constant M *and* $C_1 = 1 + \frac{(8-2\alpha)(\kappa+1)}{\alpha}$*. Then,*

$$\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\|_{\max} \leq C_2\sqrt{\frac{\tau \log p}{n}}, \tag{5.10}$$

for $C_2 = \frac{2\bar{\lambda}}{\underline{\lambda}}\gamma(C_1+\kappa) + \bar{\lambda}\kappa(1+\kappa+\kappa^2+C_1(2\kappa+1))$.

In the statement of the theorem, the bounds and the value of the regularizer parameter λ are both specified in terms of a parameter $\tau > 2$. Larger values of τ lead to looser

bounds on the error performance, but yield faster rates of probability convergence. Further, we suggest the following tuning for μ [3]

$$\mu^\star = \min\left\{0.5, \frac{40\sqrt{2}\bar{\lambda}}{\max_{i\neq j}\{|\Sigma_{i,j}|\}}\sqrt{\frac{\tau\log p}{n}}, 16\sqrt{\frac{p\tau\log p}{n}}\right\}. \tag{5.11}$$

Statement (i) of the theorem guarantees *recovery of the zero pattern* of the precision matrix $\mathbf{\Omega}$. Equivalently, for Gaussian graphical models, it guarantees recovery of the structure of the underlying graphical model. A further refinement of this result to *perfect signed-support recovery* is possible with an additional assumption on the minimum on-support entry of $\mathbf{\Omega}$, as stated in Corollary 5 in Section 5.3.

The second statement of Theorem 13 proves consistency of the FGL estimate in the max-norm sense as long as $n = O(\log p)$. This is the same order-wise result as the one obtained by Ravikumar et al. [134] regarding the log det-minimization in (5.1); see Section 5.3. Also, the bound of Theorem 13(ii) can be further used to conclude bounds on the Frobenius and on the Spectral norm of the error. We discuss these in Section 5.3.

5.3 Discussion and Numerical Experiments

We start in Section 5.3 with a couple of extensions of Theorem 13 to signed-support recovery and bounds on the Frobenius and on the spectral norm. In Section 5.3, (i) we discuss the main differences of the proposed algorithm to the GLASSO, (ii) we compare Theorem 13 to corresponding results on the performance of the convex log det minimization (5.1) derived in[75, 102, 134] and show that the FGL algorithm enjoys the same order-wise guarantees as the latter and (iii) we provide an interpretation of Assumption 5. Finally, in Section 5.3 we present the results of numerical experiments.

Extensions

Signed-support Recovery. A further refinement of Theorem 13(i) to *perfect signed-support recovery* is possible with an additional assumption on the minimum on-support entry of $\mathbf{\Omega}$ as shown in the corollary 5 below.

[3] It turns out from the analysis that the initial matrix $\bar{\mathbf{\Sigma}}$ of Algorithm 1 needs be positive definite. As discussed, the shrinkage estimators $\mathbf{R}_\mu$ enjoys that property even when $n \ll p$. The larger the value of μ is the better the bound on the minimum eigenvalue of $\mathbf{R}_\mu$, but at the same time as the parameter μ increases, we loose control of error of the estimator in terms of the operator norm and element-wise maximum norm. It turns out that the tuning suggested in (5.11) serves both of the aforementioned goals well enough.

Corollary 5 (Perfect signed-support recovery) *In addition to the assumptions of Theorem 13, suppose that n and p further satisfy* $C_2\sqrt{\frac{\tau \log p}{n}} \leq \min_{\substack{\mathbf{\Omega}_{i,j}\neq 0 \\ i\neq j}} |\mathbf{\Omega}_{i,j}|$. *Then, with probability at least* $1 - \frac{5}{p^{\tau-2}}$, *we achieve perfect signed-support recovery, i.e.,* $sign(\hat{\mathbf{\Omega}}_{i,j}) = sign(\mathbf{\Omega}_{i,j}), \forall (i,j)$.

Rates in spectral and Frobenius norms. The bound on the max-norm of Theorem 13(ii) can be used to derive bounds on the Frobenius and spectral norms of the error, as well.

Corollary 6 (Spectral and Frobenius norm bounds) *Under the same setup and assumptions as in Theorem 13(ii), with probability at least* $1 - \frac{5}{p^{\tau-2}}$, *it holds*

$$\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\|_F = O\left(\sqrt{\frac{(s+p)\tau \log p}{n}}\right) \quad \textit{and}$$
$$\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\|_2 = O\left(\sqrt{\frac{\min\{(s+p), d^2\}\tau \log p}{n}}\right). \tag{5.12}$$

Using bounds on spectral norm, we can also provide guarantees for positive definiteness of the matrix. Note that the covariance matrix $\mathbf{\Omega}$ is positive definite with $\lambda_{min}\{\mathbf{\Omega}\} \geq \bar{\lambda}^{-1}$. Thus, any matrix $\hat{\mathbf{\Omega}}$ for which $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\|_2 < \underline{\bar{\lambda}}^{-1}$ holds, is also positive definite. This in turn implies that, if $\bar{\lambda}$ is constant, then enough samples on the order of $O(\min\{s+p, d^2\}\tau \log p)$ guarantee that $\hat{\mathbf{\Omega}}$ is positive definite. In this case, it can be inverted to further obtain an estimate of the covariance matrix.

Further remarks

On Initializations of $\bar{\mathbf{\Sigma}}$. Theorem 13 characterizes the performance of the FGL algorithm under the initialization $\bar{\mathbf{\Sigma}} = \mathbf{R}_\mu$ for appropriate tuning of the parameter μ as in (5.11). It turns out from the analysis that the same desired performance is attained as long as the initialization $\bar{\mathbf{\Sigma}}$ is positive definite, and is appropriately close to the true Σ in both the spectral norm and the max-norm. For example, the sample covariance matrix $\mathbf{R}$ satisfies these conditions, but only when $n > p$. This suggests, that $\mathbf{R}$, which otherwise might have been a standard candidate for initialization of such an iterative covariance estimation algorithms, is a good initialization only when the number of observations is relatively large. However, when this is the case, initializing with $\mathbf{R}$ leads to reduced space complexity . To see this note that

calculating the LASSO objective in (5.5) does *not* require computing and storing $\mathbf{R}$, but rather, it can be done directly solely via access on the $\mathbf{x}_i$'s since for any two vectors $\mathbf{u}$ and $\mathbf{v}$, it holds $\mathbf{u}^\top \mathbf{R}\mathbf{v} = \frac{1}{n}\sum_{i=1}^n (\mathbf{u}^\top \mathbf{x}_i)(\mathbf{v}^\top \mathbf{x}_i)$, and $\mathbf{R}\mathbf{v} = \frac{1}{n}\sum_{i=1}^n (\mathbf{v}^\top \mathbf{x}_i)\mathbf{x}_i$.

Comparison to the GLASSO. The Graphical-LASSO algorithm [75] is an iterative algorithm. It starts with an initial estimate $\bar{\mathbf{\Sigma}}_0$ (say) of Σ and it iteratively updates its rows and columns. At each iteration, it solves a LASSO problem (as the name of the algorithm suggests). This is similar to the LASSO problem in (5.5) of the FGL algorithm, but other than that there are important differentiating features between the two algorithms as discussed next. First, the GLASSO at each iteration updates the last (after permutation) row/column of the covariance estimate $\bar{\mathbf{\Sigma}}$. Instead, the FGL algorithm never operates directly on $\bar{\mathbf{\Sigma}}$, but rather outputs an estimate of the precision matrix. Each iteration of the GLASSO results in the last row/column of the $\bar{\mathbf{\Sigma}}^{-1}$ that is sparse (owing to the structure-promoting nature of the LASSO), but at the same time it ruins any structure of the rest of the blocks of $\bar{\mathbf{\Sigma}}^{-1}$ obtained in previous iterations. In contrast, since the FGL algorithm operates directly on the precision matrix, it does not suffer from this issue. Besides, updating $\bar{\mathbf{\Sigma}}$ at each iteration makes the analysis of the GLASSO hard. The fact that the solutions of the LASSO problems in our algorithm are independent allows us to obtain the theoretical guarantees in Theorem 13. To the best of our knowledge, analogous results for the GLASSO are not available in the literature

Regarding time complexity, both the algorithms are scalable. Yet, FGL is a "one-shot" algorithm in the sense that it only requires solving p LASSO problems. Algorithms like the GLasso [75], the P-GLasso and the DP-Glasso [111] require at least the same complexity. An additional feature of FGL, as mentioned before, is the fact that it can be parallelized over different machines for even faster performance.

Comparison with existing guarantees on the log det**-optimization.** As discussed in the introduction, the convex nature of the estimator in (5.1) allows analyzing its statistical performance [134, 138]. In particular, Ravikumar et al.[134, Thm. 1] used a primal-dual witness approach to prove that, under an appropriate incoherence assumption on the Hessian of the log det term and under enough number of observations $n > Cd^2\tau \log p$ (for some C a constant that depends on the incoherence parameter), the optimal solution $\hat{\mathbf{\Omega}}$ of (5.1) satisfies $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\|_{\max} \leq M_3\sqrt{\frac{\tau \log p}{n}}$,

with probability at least $1 - \frac{1}{p^{\tau-2}}$. Here, M_3 depends on the model parameters such as γ, κ, etc. (for example, when these are constants then $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\|_{\max} = O(\sqrt{(\tau \log p)/n})$). Of course, this bound coincides with the result of Theorem

13(ii). Moreover, it is shown in [134] that $\text{Supp}(\hat{\mathbf{\Omega}}) \subseteq \text{Supp}(\mathbf{\Omega})$, which is also guaranteed by Theorem 13(i). Hence, for Gaussian random variables, Theorem 13 shows that the FGL algorithm attains the same order-wise performance bounds as state-of the-art corresponding results on the convex optimization in (5.1). It should be noted however that the results in [134] go beyond Gaussians; we defer such extensions for our setting to future work. Also, the two results hold under different incoherent conditions (compare Assumption 5 to [134, Ass. 1]), which are not directly comparable. Please refer to the new remark for a discussion on the interpretation of Assumption 5.

It is worth mentioning that the error bound on the Frobenius norm derived in Corollary 6 also coincides with corresponding best known results in the literature. In particular, Rothman et al. [138] showed that under a mild restriction on the minimum eigenavalue of the covariance matrix it holds with high probability that $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\|_F \leq M_1\sqrt{\frac{(s+p)\log p}{n}}$, for M depending on the parameters $\bar{\lambda}$ and $\underline{\lambda}$.

Overall, it has been shown that the algorithm proposed in this section enjoys performance guarantees (at least in the Gaussian case), which are as strong as the best known ones in the literature regarding the performance of the ℓ_1-regularized log det minimization. However, the former has superior computational performance and can scale with increasing high-dimensions of modern applications.

On Assumption 5. The incoherence Assumption 5 is similar (but not the same) to standard assumptions imposed on Σ in the performance analysis of the LASSO [115, 177]. Intuitively, this assumption limits the influence between different random variables in the following form. Suppose $\mathbf{x}$ is a zero-mean Gaussian random vector with covariance Σ. Then Assumption 5 is equivalent to the following,

$$\forall i \in [p], \quad \max_{\|\mathbf{x}_{\mathcal{S}_i}\|_\infty \leq 1} \|\mathbb{E}\left[\mathbf{x}_{\bar{\mathcal{S}}_i} | \mathbf{x}_{\mathcal{S}_i}\right]\|_\infty \leq 1 - \alpha. \tag{5.13}$$

Since $\mathbb{E}[\mathbf{x}_{\bar{\mathcal{S}}_i}] = 0$, this can be interpreted as a requirement that the influence of the variables in $\mathcal{S}_i$ on the variables in $\bar{\mathcal{S}}_i$ is not large.

Numerical Experiments

In this section, we illustrate the validity of the predictions of Theorem 13 via numerical simulations. For two different structures of the underlying graphical model, namely a line-graph and a star-graph, and for varying parameters p and n, we produce realizations of Gaussian observations and report the probability of successful signed-support recovery (see Corollary 6) and the max-norm of the

estimation error. The structure of the underlying graph determines the sparsity pattern of the concentration matrix.

Chain graph. For the chain graph, the precision matrix $\mathbf{\Omega}$ has non-zero entries only on the diagonal and on the upper and lower diagonals, i.e., it is tri-diagonal. For the simulations, we set $\mathbf{\Omega}_{i,j} = 0.5$ for $|i - j| = 1$ and $\mathbf{\Omega}_{i,i} = 2$. Note that for the chain graph $d = 2$, and for the specific values of the precision matrix, Assumption 5 is satisfied with parameter $\alpha = 0.732$.

For each one of the simulated values of pairs (n, p), we draw $N = 40$ batches of n Gaussian random observations with covariance $\Sigma = \Omega^{-1}$. We run the FGL algorithm on each data batch with input parameters $\lambda = \sqrt{\frac{\log p}{n}}$ and $\mu = \sqrt{\frac{\log p}{n}}$ (in consistency with the scaling suggested by Theorem 13.). In Figure 5.1a we have plotted the probability of signed support recovery as a function of the number of observations. As expected, more samples are needed as the problem size p increases. In Figure 5.1b the data are plotted against the rescaled sample size $n/\log p$. As suggested by Corollary 5, the curves corresponding to different values of p pile up, verifying that a sample size of $O(d^2 \log p)$ is sufficient for successful signed support recovery.

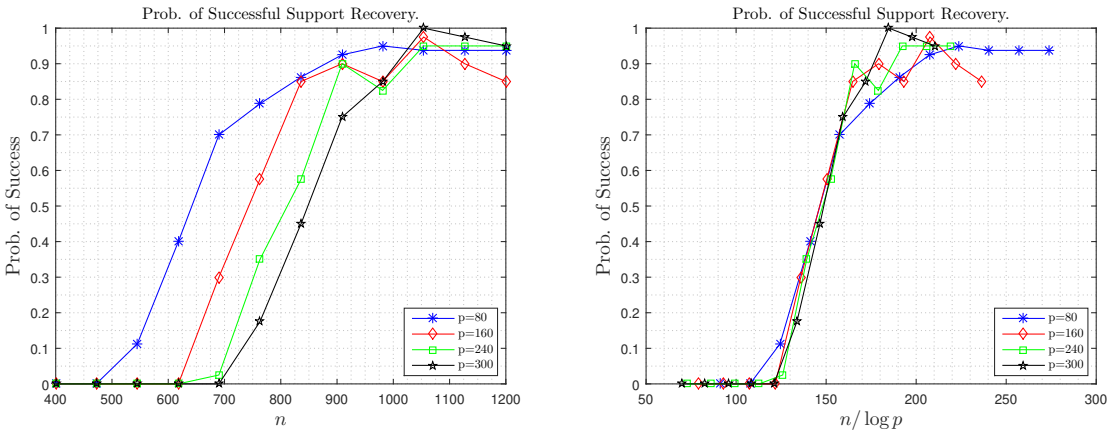
Star graph. We build the star graph by connecting its central node (first random variable) to $p/10$ other nodes and the rest of the nodes are disconnected. In fact, in the first row/column of the concentration matrix we set the first $p/10$ entries to be 0.5 and similarly, we add a scaled identity matrix to make the smallest eigen value to 1. Thus, for the star graph $d = p/10$, and for the specific values of the precision matrix, Assumption 5 is satisfied with parameter $\alpha = 0.8$.

5.4 Proof of the Main Results

Proof of Theorem 13

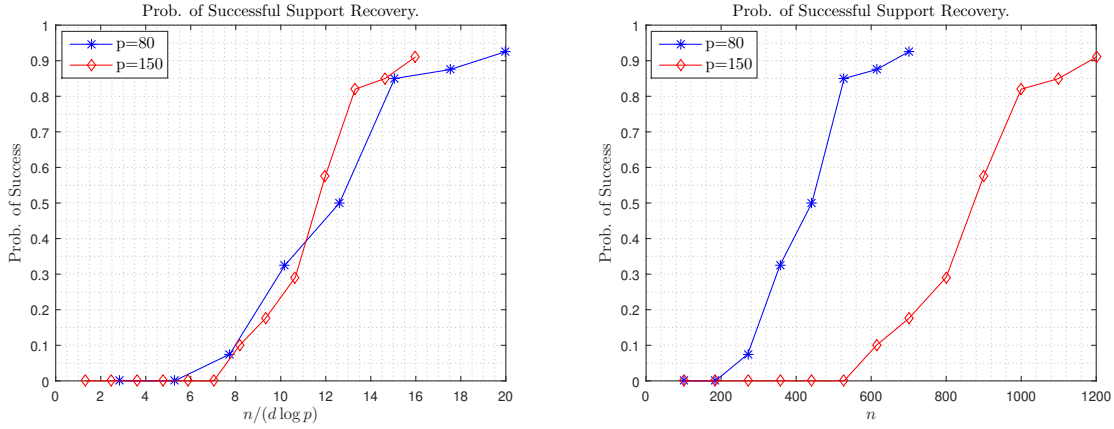
Proof 4 *In this section we prove Theorem 13 through several steps and lemmas. Before anything, we mention this result which is a simple modification of the results in [187, Proposition 2.1] and [134, Lemma 1.]*

Lemma 15 *Consider a zero-mean Gaussian random variable with covariance matrix Σ where $\lambda_{\max}(\Sigma) = \bar{\lambda}$. Given observations $\{\mathbf{x}_i\}_{i=1,\dots,n}$ of the random variable, we construct the sample covariance as $\mathbf{R} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$. Then with probability at*



(a) (b)

Figure 5.1: Plots of the probability of signed support recovery of the precision matrix corresponding to a chain graph, and for different values of p as a function of (a) the number of observations n, and, (b) of the scaled sample size $n/\log p$. The different curves pile up in (b) as predicted by Corollary 5. Each simulation point corresponds to an average over $N = 40$ points.



(a) (b)

Figure 5.2: Plots of the probability of signed support recovery of the precision matrix corresponding to a star graph with parameter $d = p/10$, and for two different values of p as a function of (a) the number of observations n, and, (b) of the scaled sample size $n/\log p/d$. The different curves pile up in (b) as predicted by Corollary 5. Each simulation point corresponds to an average over $N = 40$ points.

least $1 - \frac{5}{p^{\tau-2}}$ *we have*

$$\|\mathbf{R} - \Sigma\|_{\max} \leq 80\sqrt{2}\bar{\lambda}\sqrt{\frac{\tau \log p}{n}},$$
$$\|\mathbf{R}_{\mathcal{S}_i,\mathcal{S}_i} - \Sigma_{\mathcal{S}_i,\mathcal{S}_i}\|_2 \leq 32\bar{\lambda}\sqrt{\frac{|\mathcal{S}_i|\tau \log p}{n}}, \tag{5.14}$$

if n satisfies (5.9).

In the next step, we desire to get similar bounds on the error of the shrinkage estimator with parameter chosen as in (5.11). *We prove the following lemma,*

Lemma 16 *Consider a zero-mean Gaussian random variable with covariance matrix* Σ *where* $\lambda_{\max}(\Sigma) = \bar{\lambda}$. *Given observations* $\{\mathbf{x}_i\}_{i=1,\dots,n}$ *of the random variable, we construct the shrinkage estimator with parameter* μ *set to be as in* (5.11). *Then with probability at least* $1 - \frac{5}{p^{\tau-2}}$ *we have*

$$\|\mathbf{R}_\mu - \Sigma\|_{\max} \leq 160\sqrt{2}\bar{\lambda}\sqrt{\frac{\tau \log p}{n}},$$
$$\|\mathbf{R}_{\mu,(\mathcal{S}_i,\mathcal{S}_i)} - \Sigma_{\mathcal{S}_i,\mathcal{S}_i}\|_2 \leq 96\bar{\lambda}\sqrt{\frac{|\mathcal{S}_i|\tau \log p}{n}},$$

if n satisfies (5.9).

Proof 5 *First, we desire to bound* $\|\mathbf{R} - \Sigma\|_{\max}$. *Note that due to* (5.2) *we have,*

$$\begin{aligned}
\|\mathbf{R}_\mu - \Sigma\|_{\max} &\leq \mu\|\mathbf{D}_\mathbf{R} - \Sigma\|_{\max} + (1-\mu)\|\mathbf{R} - \Sigma\|_{\max} \\
&\leq \mu \max\left\{\|\mathbf{R}_\mu - \Sigma\|_{\max}, \max_{i,j}|\Sigma_{i,j}|\right\} \\
&+ \|\mathbf{R} - \Sigma\|_{\max} \\
&\leq 160\sqrt{2}\bar{\lambda}\sqrt{\frac{\tau \log p}{n}}
\end{aligned} \tag{5.15}$$

where the last inequality is due to lemma 15 and the way we chose μ *in* (5.11). *Now for the spectral norm we have,*

$$\begin{aligned}
&\|\mathbf{R}_{\mu,(\mathcal{S}_i,\mathcal{S}_i)} - \Sigma_{\mathcal{S}_i,\mathcal{S}_i}\|_2 \leq \mu\|\mathbf{D}_{\mathbf{R},(\mathcal{S}_i,\mathcal{S}_i)} - \Sigma_{\mathcal{S}_i,\mathcal{S}_i}\|_2 \\
&+ (1-\mu)\|\mathbf{R}_{\mathcal{S}_i,\mathcal{S}_i} - \Sigma_{\mathcal{S}_i,\mathcal{S}_i}\|_2 \\
&\leq \mu\big(\|\Sigma_{\mathcal{S}_i,\mathcal{S}_i}\|_2 + \|\mathbf{D}_{\mathbf{R},(\mathcal{S}_i,\mathcal{S}_i)} - \mathbf{D}_{\Sigma,(\mathcal{S}_i,\mathcal{S}_i)}\|_2 \\
&+ \|\mathbf{D}_{\Sigma,(\mathcal{S}_i,\mathcal{S}_i)}\|_2\big) + \|\mathbf{R}_{\mathcal{S}_i,\mathcal{S}_i} - \Sigma_{\mathcal{S}_i,\mathcal{S}_i}\|_2 \\
&\leq \mu\left(2\bar{\lambda} + \|\mathbf{R}_{\mathcal{S}_i,\mathcal{S}_i} - \Sigma_{\mathcal{S}_i,\mathcal{S}_i}\|_2\right) + \|\mathbf{R}_{\mathcal{S}_i,\mathcal{S}_i} - \Sigma_{\mathcal{S}_i,\mathcal{S}_i}\|_2 \\
&\leq 96\bar{\lambda}\sqrt{\frac{|\mathcal{S}_i|\tau \log p}{n}}
\end{aligned} \tag{5.16}$$

where the last inequality is due to lemma 15 and the value of μ *in* (5.11).

In the next step, we prove that the shrinkage estimator $\mathbf{R}_\mu$ has similar assumptions as Σ including the incoherence assumption.

Lemma 17 *Consider a zero-mean Gaussian random variable with covariance matrix Σ where $\lambda_{\max}(\Sigma) = \bar{\lambda}$. Given observations $\{\mathbf{x}_i\}_{i=1,\dots,n}$ of the random variable, we construct the shrinkage estimator with parameter μ set to be as in* (5.11)*. Suppose that Σ satisfies assumption [A1] with parameter α. Then with probability at least $1 - \frac{5}{p^{\tau-2}}$ the shrinkage estimator satisfies assumption [A1] with parameter $\alpha/2$ and also inequality* (5.8) *with parameter 2γ if n satisfies* (5.9)*.*

Proof 6 *According to lemma 16 with probability at least $1 - \frac{5}{p^{\tau-2}}$ we have*

$$\begin{aligned}
&\|\mathbf{R}_\mu - \Sigma\|_{\max} \leq 160\sqrt{2}\bar{\lambda}\sqrt{\frac{\tau \log p}{n}}, \\
&\|\mathbf{R}_{\mu,(\mathcal{S}_i,\mathcal{S}_i)} - \Sigma_{\mathcal{S}_i,\mathcal{S}_i}\|_2 \leq 96\bar{\lambda}\sqrt{\frac{|\mathcal{S}_i|\tau \log p}{n}}.
\end{aligned} \tag{5.17}$$

Let us denote $\mathbf{E} := \mathbf{R}_\mu - \Sigma$ and rename $\bar{\mathbf{\Sigma}} := \mathbf{R}_\mu$ to avoid unnecessary notations. Then for all i we have

$$\begin{aligned}
&\|\bar{\mathbf{\Sigma}}_{\bar{\mathcal{S}}_i,\mathcal{S}_i}(\bar{\mathbf{\Sigma}}_{\bar{\mathcal{S}}_i,\mathcal{S}_i})^{-1}\|_\infty = \|(\mathbf{E}_{\bar{\mathcal{S}}_i,\mathcal{S}_i} + \Sigma_{\bar{\mathcal{S}}_i,\mathcal{S}_i})(\mathbf{E}_{\mathcal{S}_i,\mathcal{S}_i} + \Sigma_{\mathcal{S}_i,\mathcal{S}_i})^{-1}\|_\infty \\
&\leq \|\Sigma_{\bar{\mathcal{S}}_i,\mathcal{S}_i}(\Sigma_{\bar{\mathcal{S}}_i,\mathcal{S}_i})^{-1}\|_\infty . \|(\mathbb{I} + (\Sigma_{\mathcal{S}_i,\mathcal{S}_i})^{-1}\mathbf{E}_{\mathcal{S}_i,\mathcal{S}_i})^{-1}\|_\infty \\
&+ \|\mathbf{E}_{\bar{\mathcal{S}}_i,\mathcal{S}_i}\|_\infty \|(\bar{\mathbf{\Sigma}}_{\mathcal{S}_i,\mathcal{S}_i})^{-1}\|_\infty \\
&\leq (1-\alpha).\frac{1}{1 - \|(\Sigma_{\mathcal{S}_i,\mathcal{S}_i})^{-1}\mathbf{E}_{\mathcal{S}_i,\mathcal{S}_i}\|_\infty} + 2\gamma.|\mathcal{S}_i|\|\mathbf{E}_{\bar{\mathcal{S}}_i,\mathcal{S}_i}\|_{\max}
\end{aligned} \tag{5.18}$$

Now we bound each term above, (note that $|\mathcal{S}_i| \leq d$)

$$\begin{aligned}
&\|(\Sigma_{\mathcal{S}_i,\mathcal{S}_i})^{-1}\mathbf{E}_{\mathcal{S}_i,\mathcal{S}_i}\|_\infty \leq \sqrt{d}\|(\Sigma_{\mathcal{S}_i,\mathcal{S}_i})^{-1}\mathbf{E}_{\mathcal{S}_i,\mathcal{S}_i}\|_2 \\
&\leq \|(\Sigma_{\mathcal{S}_i,\mathcal{S}_i})^{-1}\|_2 \|\mathbf{E}_{\mathcal{S}_i,\mathcal{S}_i}\|_2 \\
&\leq \frac{96\bar{\lambda}\sqrt{d}}{\underline{\lambda}}\sqrt{\frac{d\tau \log p}{n}} \leq \frac{3\alpha}{4-\alpha}
\end{aligned} \tag{5.19}$$

where the last inequality is because of (5.9)*. On the other hand,*

$$2\gamma.|\mathcal{S}_i|\|\mathbf{E}_{\bar{\mathcal{S}}_i,\mathcal{S}_i}\|_{\max} \leq 320\sqrt{2}\gamma d\bar{\lambda}\sqrt{\frac{\tau \log p}{n}} \leq \frac{\alpha}{4} \tag{5.20}$$

And the last inequality is due to the choice of n (5.9) *for large enough constant M. Now, combining* (5.18)*,* (5.19) *and* (5.20) *implies the following,*

$$\|\bar{\mathbf{\Sigma}}_{\bar{\mathcal{S}}_i,\mathcal{S}_i}(\bar{\mathbf{\Sigma}}_{\bar{\mathcal{S}}_i,\mathcal{S}_i})^{-1}\|_\infty \leq 1 - \frac{\alpha}{2} \tag{5.21}$$

as desired. Now we just need to bound the following

$$\begin{aligned}\|(\bar{\mathbf{\Sigma}}_{\bar{\mathcal{S}}_i,\mathcal{S}_i})^{-1}\|_\infty &\leq \|(\Sigma_{\bar{\mathcal{S}}_i,\mathcal{S}_i})^{-1}\|_\infty.\|(\mathbb{I}+(\Sigma_{\mathcal{S}_i,\mathcal{S}_i})^{-1}\mathbf{E}_{\mathcal{S}_i,\mathcal{S}_i})^{-1}\|_\infty \\ &\leq \gamma\frac{1}{1-\|(\Sigma_{\mathcal{S}_i,\mathcal{S}_i})^{-1}\mathbf{E}_{\mathcal{S}_i,\mathcal{S}_i}\|_\infty} \end{aligned} \tag{5.22}$$

Now we need to bound the later term,

$$\begin{aligned}\|(\Sigma_{\mathcal{S}_i,\mathcal{S}_i})^{-1}\mathbf{E}_{\mathcal{S}_i,\mathcal{S}_i}\|_\infty &\leq \sqrt{d}\|(\Sigma_{\mathcal{S}_i,\mathcal{S}_i})^{-1}\mathbf{E}_{\mathcal{S}_i,\mathcal{S}_i}\|_2 \\ &\leq \|(\Sigma_{\mathcal{S}_i,\mathcal{S}_i})^{-1}\|_2\|\mathbf{E}_{\mathcal{S}_i,\mathcal{S}_i}\|_2 \\ &\leq \frac{96\bar{\lambda}\sqrt{d}}{\underline{\lambda}}\sqrt{\frac{d\tau\log p}{n}} \leq \frac{1}{2}\end{aligned} \tag{5.23}$$

and the last inequality is similarly due to the choice of n. Combining (5.22) *and* (5.23) *leads to*

$$\|(\bar{\mathbf{\Sigma}}_{\bar{\mathcal{S}}_i,\mathcal{S}_i})^{-1}\|_\infty \leq 2\gamma \tag{5.24}$$

Now from now on suppose $\bar{\mathbf{\Sigma}} = \mathbf{R}_\mu$ *and with probability at least* $1-\frac{5}{p^{\tau-2}}$*, lemma 17 holds. Recall that the FGL algorithm performs p steps of iterations* $i = 1,\ldots,p$ *and in the step i, it estimates the* i^{th} *row/column of the concentration matrix* $\mathbf{\Omega}$*. All these steps are independent and can be performed in parallel. Thus, we just analyze how well the algorithm performs in recovering the* p^{st} *row/column of* $\mathbf{\Omega}$ *and then the same goes for the other steps as well. So for now assume that the we are in the* p^{th} *step of the algorithm that recovers* p^{th} *row/column of* $\mathbf{\Omega}$*. The algorithm first solves the following Lasso problem,*

$$\hat{\beta} = \arg\min_\beta \frac{1}{2}\beta^\intercal\bar{\mathbf{\Sigma}}_{11}\beta + \bar{\boldsymbol{\sigma}}_{12}^\intercal\beta + \lambda\|\beta\|_1. \tag{5.25}$$

Recall that the output of this problem was supposed to be an estimation $\hat{\beta}$ *of the quantity* $\beta_0 := \frac{\omega_{12}}{\omega_{22}}$ *and now we analyze its performance in recovering the structure of* β_0 *and also in terms of the maximum norm of the error* $\|\hat{\beta}-\beta_0\|_{max}$*.*

Performance of the Lasso

To analyze the Lasso problem (5.25)*, we utilize a well-known technique called primal-dual witness. Note that based of the assumption [B2],* $\bar{\mathbf{\Sigma}}_{11}$ *is positive definite which implies strict convexity of the objective function in* (5.25) *and also uniqueness of the optimizer* $\hat{\beta}$*. Besides,* $\hat{\beta}$ *satisfies the optimality condition*

$$\bar{\mathbf{\Sigma}}_{11}\hat{\beta} + \bar{\boldsymbol{\sigma}}_{12} + \lambda\hat{\mathbf{x}} = 0, \quad \hat{\mathbf{x}} \in \partial\|\hat{\beta}\|_1, \tag{5.26}$$

where $\partial\|\hat{\beta}\|_1$ is sub-differential of the ℓ_1-norm calculated at $\hat{\beta}$. Now we proceed by constructing a pair $(\tilde{\beta}, \tilde{x})$ that satisfies (5.26) *and due to uniqueness of the optimizer $\tilde{\beta} = \hat{\beta}$. We are going to do this in two steps, first constructing the pair $(\tilde{\beta}, \tilde{x})$ and then proving that this pair satisfies* (5.26)*.*

***1. Constructing the pair** $(\tilde{\beta}, \tilde{x})$**:** We denote support of the vector β_0 by $\mathbf{S}$ and its compliment by $\mathbf{S}^c$ and for a vector $\mathbf{v} \in \mathbb{R}^{p-1}$, $\mathbf{v}_\mathbf{S} \in \mathbb{R}^{|\mathbf{S}|}$ is a sub-vector of $\mathbf{v}$ with entries $\{v_i\}_{i\in\mathbf{S}}$.*
In order to construct $(\tilde{\beta}, \tilde{x})$, we set $\tilde{\beta}_{\mathbf{S}^c} = 0$ and

$$\tilde{\beta}_\mathbf{S} = \arg\min_{\beta_\mathbf{S}} \frac{1}{2}\beta_\mathbf{S}^\mathsf{T}\bar{\mathbf{\Sigma}}_{\mathbf{S},\mathbf{S}}\beta_\mathbf{S} + \bar{\sigma}_{12\,\mathbf{S}}^\mathsf{T}\beta_\mathbf{S} + \lambda\|\beta_\mathbf{S}\|_1.$$

Thus $\tilde{\beta}_\mathbf{S}$ satisfies the following optimality condition,

$$\bar{\mathbf{\Sigma}}_{\mathbf{S},\mathbf{S}}\tilde{\beta}_\mathbf{S} + \bar{\sigma}_{12\,\mathbf{S}} + \lambda\tilde{x}_\mathbf{S} = 0, \quad \tilde{x}_\mathbf{S} \in \partial\|\tilde{\beta}_\mathbf{S}\|_1. \tag{5.27}$$

Note that in (5.27) *we also set $\tilde{x}_\mathbf{S}$. At last we choose $\tilde{x}_{\mathbf{S}^c}$ to be*

$$\tilde{x}_{\mathbf{S}^c} = -\frac{1}{\lambda}(\bar{\mathbf{\Sigma}}_{\mathbf{S}^c,\mathbf{S}}\tilde{\beta}_\mathbf{S} + \bar{\sigma}_{12\,\mathbf{S}^c}). \tag{5.28}$$

Because of the way we constructed the pair $(\tilde{\beta}, \tilde{x})$, we already have

$$\bar{\mathbf{\Sigma}}_{11}\tilde{\beta} + \bar{\sigma}_{12} + \lambda\tilde{x} = 0, \quad \tilde{x}_\mathbf{S} \in \partial\|\tilde{\beta}_\mathbf{S}\|_1.$$

We just need to check if $\tilde{x}_{\mathbf{S}^c} \in \partial\|\tilde{\beta}_{\mathbf{S}^c}\|_1$ or equivalently if $\|\tilde{x}_{\mathbf{S}^c}\|_\infty \leq 1$ because $\tilde{\beta}_{\mathbf{S}^c} = 0$.

***2. Verifying Optimality Conditions:** First we define error of estimation in $\bar{\mathbf{\Sigma}}$ to be $\mathbf{Z} := \bar{\mathbf{\Sigma}} - \Sigma$ and also $\Delta := \bar{\mathbf{\Sigma}}_{11}\tilde{\beta} - \mathbf{\Sigma}_{11}\beta_0$. Thus*

$$\begin{cases} \Delta_\mathbf{S} = \bar{\mathbf{\Sigma}}_{\mathbf{S},\mathbf{S}}(\tilde{\beta}_\mathbf{S} - \beta_{0\,\mathbf{S}}) + \mathbf{Z}_{\mathbf{S},\mathbf{S}}\beta_{0\,\mathbf{S}} \\ \Delta_{\mathbf{S}^c} = \bar{\mathbf{\Sigma}}_{\mathbf{S}^c,\mathbf{S}}(\tilde{\beta}_\mathbf{S} - \beta_{0\,\mathbf{S}}) + \mathbf{Z}_{\mathbf{S}^c,\mathbf{S}}\beta_{0\,\mathbf{S}} \end{cases}$$
$$\Rightarrow \Delta_{\mathbf{S}^c} = \mathbf{W}\Delta_\mathbf{S} - \mathbf{W}\mathbf{Z}_{\mathbf{S},\mathbf{S}}\beta_{0\,\mathbf{S}} + \mathbf{Z}_{\mathbf{S}^c,\mathbf{S}}\beta_{0\,\mathbf{S}}, \tag{5.29}$$

where $\mathbf{W} := \bar{\mathbf{\Sigma}}_{\mathbf{S}^c,\mathbf{S}}(\bar{\mathbf{\Sigma}}_{\mathbf{S},\mathbf{S}})^{-1}$. Now, combining (5.28), (5.29) *and $\bar{\sigma}_{12\,\mathbf{S}^c} = -\Sigma_{\mathbf{S}^c,\mathbf{S}}\beta_{0\,\mathbf{S}} + \mathbf{z}_{12\,\mathbf{S}^c}$ implies the following,*

$$\tilde{x}_{\mathbf{S}^c} = \frac{1}{\lambda}\mathbf{W}\Delta_\mathbf{S} + \frac{1}{\lambda}\mathbf{W}\mathbf{Z}_{\mathbf{S},\mathbf{S}}\beta_{0\,\mathbf{S}} - \frac{1}{\lambda}\mathbf{Z}_{\mathbf{S}^c,\mathbf{S}}\beta_{0\,\mathbf{S}} - \frac{1}{\lambda}\mathbf{z}_{12\,\mathbf{S}^c}. \tag{5.30}$$

We also make use the equation (5.27) *and* $\bar{\sigma}_{12\,\mathbf{S}} = -\Sigma_{\mathbf{S},\mathbf{S}}\beta_{0\,\mathbf{S}} + \mathbf{z}_2 12\,\mathbf{S}$ *to rewrite* (5.30) *as*

$$\tilde{x}_{\mathbf{S}^c} = \frac{1}{\lambda}\mathbf{W}\mathbf{z}_{12\,\mathbf{S}} + \mathbf{W}\tilde{x}_{\mathbf{S}} + \frac{1}{\lambda}\mathbf{W}\mathbf{Z}_{\mathbf{S},\mathbf{S}}\beta_{0\,\mathbf{S}} - \frac{1}{\lambda}\mathbf{Z}_{\mathbf{S}^c,\mathbf{S}}\beta_{0\,\mathbf{S}} - \frac{1}{\lambda}\mathbf{z}_{12\,\mathbf{S}^c}. \tag{5.31}$$

We desire to bound $\|\tilde{x}_{\mathbf{S}^c}\|_\infty$ *and so*

$$\begin{aligned}\|\tilde{x}_{\mathbf{S}^c}\|_\infty \leq \frac{1}{\lambda}\|\mathbf{W}\|_\infty\|\mathbf{z}_{12\,\mathbf{S}}\|_\infty + \|\mathbf{W}\|_\infty\|\tilde{x}_{\mathbf{S}}\|_\infty \\ + \frac{1}{\lambda}\|\mathbf{W}\|_\infty\|\mathbf{Z}_{\mathbf{S},\mathbf{S}}\|_{max}\|\beta_{0\,\mathbf{S}}\|_1 \\ + \frac{1}{\lambda}\|\mathbf{Z}_{\mathbf{S}^c,\mathbf{S}}\|_{max}\|\beta_{0\,\mathbf{S}}\|_1 + \frac{1}{\lambda}\|\mathbf{z}_{12\,\mathbf{S}^c}\|_{max}.\end{aligned} \tag{5.32}$$

From the assumptions we have a bound on all the variables above including $\|\mathbf{Z}\|_{max} \leq f(n,p)$, $\|\mathbf{W}\|_\infty \leq 1-\bar{\alpha}$, $\|\beta_0\|_1 \leq \kappa$ *and* $\lambda = \frac{4-2\bar{\alpha}}{\bar{\alpha}}(\kappa+1)f(n,p)$ *that we can use in* (5.32) *which results in* $\|\tilde{x}_{\mathbf{S}^c}\|_\infty \leq 1$.

3. Deriving Performance of the Lasso: *The previous two steps showed that* $\hat{\beta} = \tilde{\beta}$ *is the unique answer to the lasso problem* (5.25) *with the property that Support*$\{\tilde{\beta}\} \subset$*Support*$\{\beta_0\}$. *This already proves the first part of the theorem Since in the way we estimate* $\hat{\boldsymbol{\omega}}_{12}$ *from* $\hat{\beta}$ *in the algorithm, Support*$\{\tilde{\beta}\} \subset$*Support*$\{\beta_0\}$ *implies Support*$\{\hat{\boldsymbol{\omega}}_{12}\} \subset$*Support*$\{\omega_{12}\}$.

On the other hand, due to equation (5.26)

$$\Delta + \mathbf{z}_{12} + \lambda\hat{x} = 0 \Rightarrow \|\Delta\|_\infty \leq \lambda + \|\mathbf{z}_{12}\|_\infty \leq C_1 f(n,p), \tag{5.33}$$

where $C_1 = 1 + \frac{(4-2\bar{\alpha})(\kappa+1)}{\bar{\alpha}}$. *We also can rewrite* (5.27) *as*

$$\begin{aligned}&\bar{\mathbf{\Sigma}}_{\mathbf{S},\mathbf{S}}\tilde{\beta}_{\mathbf{S}} - \bar{\mathbf{\Sigma}}_{\mathbf{S},\mathbf{S}}\beta_{0\,\mathbf{S}} + \bar{\mathbf{\Sigma}}_{\mathbf{S},\mathbf{S}}\tilde{\beta}_{0\,\mathbf{S}} + \bar{\sigma}_{12\,\mathbf{S}} + \lambda\tilde{x}_{\mathbf{S}} = 0 \Rightarrow \\ &\tilde{\beta}_{\mathbf{S}} - \beta_{0\,\mathbf{S}} = \bar{\mathbf{\Sigma}}^{-1}_{\mathbf{S},\mathbf{S}}\mathbf{Z}_{\mathbf{S},\mathbf{S}}\beta_{0\,\mathbf{S}} - \bar{\mathbf{\Sigma}}^{-1}_{\mathbf{S},\mathbf{S}}\mathbf{z}_{12\,\mathbf{S}} - \lambda\bar{\mathbf{\Sigma}}^{-1}_{\mathbf{S},\mathbf{S}}\tilde{x}_{\mathbf{S}}.\end{aligned}$$

Thus,

$$\begin{aligned}\|\tilde{\beta} - \beta_0\|_\infty = \|\tilde{\beta}_{\mathbf{S}} - \beta_{0\,\mathbf{S}}\|_\infty \leq \gamma(\kappa+1)f(n,p) \\ + \gamma(\kappa+1)\frac{(4-2\bar{\alpha})}{\bar{\alpha}}f(n,p) = \gamma(C_1+\kappa)f(n,p),\end{aligned} \tag{5.34}$$

where we used $\tilde{\beta}_{\mathbf{S}^c} = \beta_{0\,\mathbf{S}^c} = 0$ *and* $\|\bar{\mathbf{\Sigma}}^{-1}_{\mathbf{S},\mathbf{S}}\|_\infty \leq \gamma$ *from the assumptions. Besides,* $(\tilde{\beta} - \beta_0)$ *has at most* $s = |\mathbf{S}|$ *non-zeros entries because Support*$\{\tilde{\beta}\} \subset \mathbf{S}$ =*Support*$\{\beta_0\}$, *so*

$$\|\tilde{\beta} - \beta_0\|_1 \leq s\|\tilde{\beta} - \beta_0\|_\infty \leq s\gamma(C_1+\kappa)f(n,p) \leq 1, \tag{5.35}$$

where last inequality is because of the additional assumption in the theorem.

Estimating ω_{12} and ω_{22} from $\frac{\omega_{12}}{\omega_{22}}$

We know from previous section that how well the Lasso problem (5.25) *performs in estimating $\frac{\omega_{12}}{\omega_{22}}$. The FGL algorithm uses the following equations to estimate ω_{12} and ω_{22} using the answer $\hat{\beta}$ to* (5.25)*,*

$$\begin{aligned}\hat{\omega}_{22} &= \frac{1}{\bar{\sigma}_{22} + \hat{\beta}^{\mathsf{T}}\bar{\boldsymbol{\Sigma}}_{11}\hat{\beta}},\\ \hat{\boldsymbol{\omega}}_{12} &= \frac{\hat{\beta}}{\bar{\sigma}_{22} + \hat{\beta}^{\mathsf{T}}\bar{\boldsymbol{\Sigma}}_{11}\hat{\beta}}.\end{aligned} \tag{5.36}$$

We are interested in bounding the error terms $|\hat{\omega}_{22} - \omega_{22}|$ and $\hat{\boldsymbol{\omega}}_{12} - \omega_{12}\|_\infty\|$. As the first step,

$$\begin{aligned}&(\sigma_{22} + \beta_0^{\mathsf{T}}\boldsymbol{\Sigma}_{11}\beta_0) - (\bar{\sigma}_{22} + \hat{\beta}^{\mathsf{T}}\bar{\boldsymbol{\Sigma}}_{11}\hat{\beta}) = (\sigma_{22} - \bar{\sigma}_{22})\\ &+ 2\beta_0^{\mathsf{T}}\Delta + (\hat{\beta} - \beta_0)^{\mathsf{T}}\Delta\\ &+ \beta_0^{\mathsf{T}}(\bar{\boldsymbol{\Sigma}}_{11} - \boldsymbol{\Sigma}_{11})\beta_0 + \beta_0^{\mathsf{T}}(\bar{\boldsymbol{\Sigma}}_{11} - \boldsymbol{\Sigma}_{11})(\hat{\beta} - \beta_0)\end{aligned}$$

Therefore, Cauchy–Schwarz and triangle inequality implies

$$\begin{aligned}&|(\sigma_{22} + \beta_0^{\mathsf{T}}\boldsymbol{\Sigma}_{11}\beta_0) - (\bar{\sigma}_{22} + \hat{\beta}^{\mathsf{T}}\bar{\boldsymbol{\Sigma}}_{11}\hat{\beta})| \le |\sigma_{22} - \bar{\sigma}_{22}|\\ &+ 2\|\beta_0\|_1\|\Delta\|_\infty + \|\hat{\beta} - \beta_0\|_1\|\Delta\|\\ &+ \|\beta_0\|_1^2\|\bar{\boldsymbol{\Sigma}}_{11} - \boldsymbol{\Sigma}_{11}\|_{max} + \|\beta_0\|_1\|\bar{\boldsymbol{\Sigma}}_{11} - \boldsymbol{\Sigma}_{11}\|_{max}\|\hat{\beta} - \beta_0\|_1\\ &\le (1 + 2\kappa C_1 + C_1 + \kappa^2 + \kappa) f(n,p)\end{aligned} \tag{5.37}$$

On the other hand, because of the extra assumption in the theorem $|\bar{\sigma}_{22} - \sigma_{22}| \le f(n,p) \le \underline{\lambda}/2$ and therefore

$$\begin{aligned}&(\bar{\sigma}_{22} + \hat{\beta}^{\mathsf{T}}\bar{\boldsymbol{\Sigma}}_{11}\hat{\beta}) \ge \bar{\sigma}_{22} \ge \sigma_{22} - |\bar{\sigma}_{22} - \sigma_{22}|\\ &\ge \underline{\lambda} - \underline{\lambda}/2 = \underline{\lambda}/2\end{aligned} \tag{5.38}$$

Now, combining (5.37) *and* (5.38) *gives us the following*

$$\begin{aligned}\hat{\omega}_{22} - \omega_{22} &= \frac{1}{\bar{\sigma}_{22} + \hat{\beta}^{\mathsf{T}}\bar{\boldsymbol{\Sigma}}_{11}\hat{\beta}} - \frac{1}{\sigma_{22} + \beta_0^{\mathsf{T}}\boldsymbol{\Sigma}_{11}\beta_0}\\ &\le \frac{2}{\underline{\lambda}^2}(1 + 2\kappa C_1 + C_1 + \kappa^2 + \kappa) f(n,p)\end{aligned} \tag{5.39}$$

Finally,

$$\begin{aligned}&\|\hat{\boldsymbol{\omega}}_{12} - \omega_{12}\|_\infty = \|\hat{\beta}\hat{\omega}_{22} - \beta_0\hat{\omega}_{22} + \beta_0\hat{\omega}_{22} - \beta_0\omega_{22}\|_\infty\\ &\le \|\hat{\beta} - \beta_0\|_\infty\hat{\omega}_{22} + |\hat{\omega}_{22} - \omega_{22}|\|\beta_0\|_\infty\\ &\le \frac{2}{\underline{\lambda}}\gamma(C_1 + \kappa) f(n,p) + \kappa(1 + 2\kappa C_1 + C_1 + \kappa^2 + \kappa) f(n,p),\end{aligned} \tag{5.40}$$

where the last inequality is from (5.38)*,* (5.34) *and* (5.39)*. This implies our final result in the theorem,*

$$\begin{aligned}&\|\hat{\boldsymbol{\omega}}_{12} - \omega_{12}\|_\infty \\ &\leq \left(\frac{2}{\underline{\lambda}}\gamma(\kappa + C_1) + \kappa(1 + \kappa + \kappa^2 + C_1(2\kappa + 1))\right) f(n, p). \end{aligned} \tag{5.41}$$

Chapter 6

UNIVERSALITY IN LEARNING FROM LINEAR MEASUREMENTS

Recovering[1] 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 $\mathbf{x}_0 \in \mathbb{R}^p$, from observations of the form $y_i = \mathbf{a}_i^\mathsf{T}\mathbf{x}_0$, for $i = 1, \ldots, n$. In many modern applications, the ambient dimension of the signal, p, is often (overwhelmingly) larger than the number of observations, p. In such cases, there are infinitely many solutions that satisfy the linear equations arising from the observations, and therefore to obtain a unique solution one must assume some prior structure on the unknown vector. Common examples of structured signals are sparse and group-sparse vectors [35, 61], low-rank matrices [34, 135], and simultaneously-structured matrices [39, 128]. To this end, we use a convex penalty function $f : \mathbb{R}^p \to \mathbb{R}$, that captures the *structure* of the structured signal, in the sense that signals that do not adhere to the desired structure will have a higher cost. Therefore, the following estimator is used to recover $\mathbf{x}_0$,

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \ f(\mathbf{x}) \quad \text{subject to,} \quad y_i = \mathbf{a}_i^\mathsf{T}\mathbf{x}, \ \ i = 1, \ldots, n \ . \tag{6.1}$$

Popular choices of $f(\cdot)$ include the ℓ_1-norm for sparse vectors [176], and the nuclear norm for low-rank matrices [135]. 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 [6, 40, 156] 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 iid 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 [6, 40, 156]). For instance, it has been shown that $2k \log(p/k)$ linear measurements

[1] This chapter is mainly based on the work in [3]

is required to recover a $k-$sparse signal [58], and $3rp$ measurements suffice for the recovery of a symmetric $p \times p$ rank-r matrix [40, 127]. Recently, Oymak et al [130] 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 [131]. Although these works are of great interest, the independence assumption on the entries of the measurement vectors can be restrictive. In certain applications in communications, phase retrieval, covariance estimation, the entries of the measurement vectors $\mathbf{a}_i$ have correlations. in this section, we show a much stronger universality result which holds for a broader class of measurement distributions. Here is an informal description of our result:

> *Assume the measurement vectors* $\mathbf{a}_i$ *are drawn iid from some given distribution. In other words, the measurement vectors are iid random, but their entries are not necessarily so. Then the minimum number of observations needed to recover* $\mathbf{x}_0$ *from* (6.1) *with high probability, depends only on the first two statistics of the* $\mathbf{a}_i$*, i.e., their mean vector* μ*, and covariance matrix* $\mathbf{\Sigma}$*.*

We anticipate that this universality result will have many practical ramifications. in this section we focus on the ramifications to the problem of recovering a structured matrix, $\mathbf{X}_0 \in \mathbb{R}^{p\times p}$, from quadratic measurements (a.k.a. rank-one projections). In this problem, we are given observations of the form $y_i = \mathbf{a}_i^\top \mathbf{X}_0 \mathbf{a}_i = \mathrm{Tr}(\mathbf{X}_0(\mathbf{a}_i\mathbf{a}_i^\top)) = \mathrm{vec}(X)^t \mathrm{vec}(\mathbf{a}_i\mathbf{a}_i^t)$ for $i = 1, \ldots, m$.[2] Such measurement schemes appear in a variety of problems [31, 44, 104, 105, 194]. An interesting application of learning from quadratic measurements is the PhaseLift algorithm [36] for phase retrieval. In phase retrieval, the goal is to recover the signal $\mathbf{x}_0$ from quadratic measurements of the form, $y_i = |\mathbf{a}_i^\top \mathbf{x}_0|^2 = \mathbf{a}_i^\top(\mathbf{x}_0\mathbf{x}_0^\top)\mathbf{a}_i$. Note that $\mathbf{x}_0\mathbf{x}_0^t$ is a low-rank (in this case rank-1) matrix and PhaseLift relaxes this constraint to a non-negativity constraint and minimizes nuclear norm to encourage a low rank solution. Quadratic measurements also appears in non-coherent energy measurements in communications and signal processing [8, 180], sparse covariance estimation [44, 194], and sparse phase retrieval [104, 147]. Recently, Chen et al [44] proved sufficient bounds on the number of measurements for various structures on the matrix $\mathbf{X}_0$. However, to the best of our knowledge, prior to this work, the precise number of required measurements for

[2] The reader should pardon the abuse of notation as the measurement vectors are now $\mathrm{vec}(\mathbf{a}_i\mathbf{a}_i^t)$.

perfect recovery was unknown.

For example, when the $\mathbf{a}_i$ have iid Gaussian entries (note that the measurement vectors, which are now $\text{vec}(\mathbf{a}_i\mathbf{a}_i^t)$, are no longer iid Gaussian) we show that $3pr$ measurement is necessary and sufficient for the perfect recovery of a rank-r matrix from quadratic measurements. In the special case of phase retrieval, we therefore demonstrate that $3p$ measurements is necessary and sufficient for perfect recovery of $\mathbf{x}_0$, which settles the long standing open question of the recovery threshold for PhaseLift. In particular, this indicates that $2p$ extra phaseless measurements is all that is needed to compensate the missing phase information.

The remainder of the paper is structured as follows. The problem setup and definitions are given in Section 6.1. In Section 6.2, we introduce our universality framework, which states that the number of required observations for the recovery of an unknown model depends only on the first two statistics of the measurement vectors. As an applications, in Section 6.3, we apply this universality theorem to derive tight bounds (i.e., necessary and sufficient conditions) on the required number of observations for matrix recovery via quadratic measurements.

6.1 Preliminaries

Notations

We start by introducing some notations that are used throughout the paper. Bold lower letters $\mathbf{x}, \mathbf{y}, \ldots$ are used to denote vectors, and bold upper letters $\mathbf{X}, \mathbf{Y}, \ldots$ are for matrices. For a matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\text{Vec}(\mathbf{X}) \in \mathbb{R}^{pn}$ returns the vectorized form of the matrix. $\|\mathbf{X}\|_2$, $\|\mathbf{X}\|_F$, $\|\mathbf{X}\|_\star$ and $\text{Tr}(\mathbf{X})$ represent the operator norm, the Frobenius norm, the nuclear norm and the trace of the matrix $\mathbf{X}$, respectively. $\|\mathbf{x}\|_{\ell_p}$ denotes the ℓ_p-norm of the vector $\mathbf{x}$ and for matrices, $\|\mathbf{X}\|_{\ell_p} = \|\text{Vec}(\mathbf{X})\|_{\ell_p}$. For both vectors and matrices, $\|\cdot\|_0$ indicates the number of non-zero entries. The set of $p \times p$ positive definite matrices and positive semi-definite matrices are denoted by $\mathbb{S}^p_{++}$ and $\mathbb{S}^n_+$, respectively. The letters $\mathbf{g}$ and $\mathbf{G}$ are reserved for a Gaussian random vector and matrix with i.i.d. standard normal entries. The letter $\mathbf{H}$ is reserved for a random Gaussian *Wigner* matrix, that is a *symmetric* matrix whose upper-diagonal entries drawn independently from $\mathcal{N}(0,1)$ whose its diagonals entries are drawn independently from $\mathcal{N}(0,2)$. Finally, the letter $\mathbf{I}$ is reserved for the identity matrix. For a random vector $\mathbf{a}$, $\mathbb{E}[\mathbf{a}]$ and $\text{Cov}[\mathbf{a}]$ represent the expected value and the covariance matrix of $\mathbf{a}$.

Problem Setup

We consider the problem of recovering the unknown vector $\mathbf{x}_0 \in \mathcal{S} \subseteq \mathbb{R}^p$ from n observations of the form $y_i = \mathbf{a}_i^\mathsf{T}\mathbf{x}_0, i = 1, \ldots, n$. Here, the *known* measurement vectors $\mathbf{a}_i \in \mathbb{R}^p$'s are drawn independently and identically from a random distribution. These observations can be reformulated as

$$\mathbf{y} = \mathbf{A}\mathbf{x}_0 \,, \tag{6.2}$$

where $\mathbf{y} = [y_1, \ldots, y_n]^\mathsf{T} \in \mathbb{R}^n$ and $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n]^\mathsf{T} \in \mathbb{R}^{n\times p}$. We focus on the high-dimensional setting where both n and p grow large. We use the notation $n = \theta(p)$, to fix the rate at which n grows compared to p. Of special interest is the underdetermined case where the number of measurement is smaller than the ambient dimension. In this case, the problem of signal reconstruction is generally ill-posed unless some prior information is available regarding the structure of $\mathbf{x}_0$. Some popular cases of structures include, *sparse* vectors, *low-rank* matrices, and simultaneously-structured matrices.

Convex estimator: To recover the structured vector $\mathbf{x}_0$, we minimize a convex function $f : \mathbb{R}^p \to \mathbb{R}$ that enforces this structure. We do this minimization for all feasible points $\mathbf{x} \in \mathcal{S}$, that satisfy $\mathbf{y} = \mathbf{A}\mathbf{x}$. We formally define such estimators as follows,

Definition 3 *Let $\mathbf{x}_0 \in \mathcal{S}$ where $\mathcal{S} \subseteq \mathbb{R}^p$ is a convex set. For a convex function $f : \mathbb{R}^p \to \mathbb{R}$ and a measurement matrix $\mathbf{A} \in \mathbb{R}^{n\times p}$, we define the convex estimator $\mathcal{E}\{\mathbf{x}_0, \mathbf{A}, \mathcal{S}, f(\cdot)\}$ as following,*

$$\hat{\mathbf{x}} = \arg \min_{\substack{\mathbf{x}\in\mathcal{S} \\ \mathbf{A}\mathbf{x}=\mathbf{A}\mathbf{x}_0}} f(\mathbf{x}) \,. \tag{6.3}$$

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

Note that we are given the observation vector $\mathbf{y} = \mathbf{A}\mathbf{x}_0$ in the constraint of (6.3). We aim to characterize the perfect recovery criteria for this estimator. Given a structured vector $\mathbf{x}_0$, the perfect recovery of an estimator $\mathcal{E}\{\mathbf{x}_0, \mathbf{A}, \mathcal{S}, f(\cdot)\}$ depends on three factors; the number of observations n compared to the dimension of the ambient space p, properties of the measurement vectors $\{\mathbf{a}_i\}_{i=1}^n$, and the penalty function, $f(\cdot)$. We briefly explain each factor, below.

The rate function $\theta(\cdot)$: We work in the high dimensional regime where both n and p grow to infinity with a fixed rate $n = \theta(p)$. Finding the minimum number of

measurements to recover $\mathbf{x}_0$ via (6.3), translates to finding the *smallest rate function* $\theta^\star(\cdot)$, for which our estimator has perfect recovery. This optimal rate function depends on the problem settings and varies in different problems. For instance, in order to recover a rank-r matrix in $\mathbb{S}_+^p$, we will need the measurements to be of order $n = O(p)$, while in the case of k-sparse matrices, the measurements will be of order $n = O(k \log(p^2/k))$, where in many applications k is a fraction of p^2.

The penalty function: We use a convex function $f(\cdot)$ that promotes the particular structure of $\mathbf{x}_0$. Exploiting a convex penalty for the recovery of structured signals has been studied extensively [6, 33, 40, 62, 156, 165]. Chandrasekaran et. al. [40] introduced the concept of the atomic norm, which is a convex surrogate defined based on a set of (so-called) "atoms". For instance, the corresponding atomic norm for sparse recovery is the ℓ_1-norm and for low-rank matrix recovery the nuclear norm. Another interesting scenario is when the underlying parameter $\mathbf{x}_0$ simultaneously exhibits multiple structures such as being low-rank and sparse. For simultaneously structured signals building the set of atoms is often intractable. Therefore, it has been proposed [43, 128] to use a weighted sum of corresponding atomic norms for each structure as the penalty.

The measurement vectors: We consider a random ensemble, where the vectors $\{\mathbf{a}_i\}_{i=1}^n$ are drawn *independently and identically* from a random distribution. Later in Section 6.1, we formally present the required assumptions on this distribution. It has been observed that the estimator (6.3) exhibits a *phase transition* phenomenon, i.e., there exist a phase transition rate $\theta^\star(p)$, such that when $n > \theta^\star(p)$ the optimization program (6.3) successfully recover $\mathbf{x}_0$ with high probability, otherwise, when $n < \theta^\star(p)$ it fails with high probability [6, 40]. The question is that *how is this phase transition is related to the properties of the measurement vectors* $\mathbf{a}_i$*'s?*

Universality in learning: Directly calculating the precise phase transition behavior of the estimator $\mathcal{E}(\mathbf{x}_0, \mathbf{A}, \mathcal{S}, f(\cdot))$, for a general random distribution on the measurement vectors is very challenging. Recently, as an extension of Gaussian comparison lemmas due to Gordon [79, 80] and earlier work in [6, 40, 153, 156], a new framework, known as CGMT [165, 171], has been developed which made this analysis possible when the measurement vectors $\{\mathbf{a}_i\}_{i=1}^n$, are independently drawn from the Gaussian distribution, $\mathcal{N}(0, \mathbf{I}_p)$. Another parallel work that makes this analysis

possible under the same conditions is known as AMP [62]. However, the Gaussian assumption is critical in the analysis through these frameworks, which restricts us from investigating a vast variety of practical problems.

As our main result, we show that, for a broad class of distributions, the phase transition of $\mathcal{E}(\mathbf{x}_0, \mathbf{A}, \mathcal{S}, f(\cdot))$ depends only on the first two statistics of the distribution on the measurement vectors $\{\mathbf{a}_i\}_{i=1}^n$. As a result, the phase transition of the estimator remains unchanged when we replace the measurement vectors with the ones drawn from a Gaussian distribution with the same mean vector and covariance matrix. As the phase transition is the same as the one with Gaussian measurements, we can use the CGMT framework to analyze the latter and get the desired result.

Equivalent Gaussian Problem: Let $\mu := \mathbb{E}[\mathbf{a}_i]$ and $\bar{\mathbf{\Sigma}} := \mathrm{Cov}[\mathbf{a}_i]$ for $i = 1, 2, \ldots, n$, and consider the following problem:

1. We are given n observations of the form $\tilde{y}_i = \mathbf{g}_i^\mathsf{T}\mathbf{x}_0$ and the measurement vectors $\{\mathbf{g}_i\}_{i=1}^n$.
2. The rows of the measurement matrix $\mathbf{G} = [\mathbf{g}_1, \ldots, \mathbf{g}_n]^\mathsf{T} \in \mathbb{R}^{n\times p}$ are independently drawn from the multivariate Gaussian distribution $\mathcal{N}(\mu, \bar{\mathbf{\Sigma}})$.
3. We use the estimator $\mathcal{E}(\mathbf{x}_0, \mathbf{G}, \mathcal{S}, f(\cdot))$, as in Definition 3, to recover $\mathbf{x}_0$.

In Theorem 14, we show that under certain conditions, the two estimators $\mathcal{E}(\mathbf{x}_0, \mathbf{A}, \mathcal{S}, f(\cdot))$ and $\mathcal{E}(\mathbf{x}_0, \mathbf{G}, \mathcal{S}, f(\cdot))$ asymptotically exhibit the same phase transition behavior. Before stating our main result in Section 6.2, we discuss the assumptions needed for our universality to hold.

Assumptions

We show universality for a wide range of distributions on the measurement vector as well as a broad class of convex penalties. Here, we give the conditions needed for the measurement matrix,

Assumption 6 ***[The Measurement Vectors]*** *We say the measurement matrix* $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n]^\mathsf{T} \in \mathbb{R}^{n\times p}$ *satisfies Assumption 6 with parameters* $\mu \in \mathbb{R}^p$ *and* $\bar{\mathbf{\Sigma}} \in \mathbb{R}^{p\times p}$*, if the followings hold true.*

1. *[Sub-Exponential Tails] The vectors* $\mathbf{a}_i$*'s are independently drawn from a random sub-exponential distribution, with mean* μ *and covariance* $\bar{\mathbf{\Sigma}} \succ 0$*.*

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

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

Assumption 6 summarizes the technical conditions that are essential in the proof of our main theorem. The first assumption on the tail of the distribution enables us to exploit concentration inequalities for sub-exponential distributions. We allow the vector $\mathbf{a}_i$ to have a non-zero mean in Assumption 1.2. 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 $\mathbb{R}^p$, and not be biased towards a specific direction. Finally, Assumption 1.3 is meant to control the dependencies among the entries of $\mathbf{a}_i$ and is used to prove concentration of $\frac{1}{p}\mathbf{a}_i^\top \mathbf{M}\mathbf{a}_i$ around its mean, for a matrix $\mathbf{M}$ with bounded operator norm. For instance, for a Gaussian vector $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we have $\mathrm{Var}[\|\mathbf{g}\|^2] = 2p$ and $\mathbb{E}^2[\|\mathbf{g}\|^2] = p^2$. So Assumption 1.3 is satisfied with $c_2 = 2$ and $\tau_2 = 1$. We will examine these assumptions for the applications discussed in Section 6.3.
In addition, we need to enforce a few conditions on the penalty function $f(\cdot)$ as follows,

Assumption 7 ***[The Penalty Function]*** *We say the funtion $f(\cdot)$ satisfies Assumption 2, if the following holds true.*

1. *[Separablity] $f(\cdot)$ is continuous, convex and separable, where $f(\mathbf{x}) = \sum_{i=1}^{p} f_i(x_i)$.*

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 2.1, we only consider the special (yet popular) case of separable penalty functions. Common choices include $\|\mathbf{x}\|_{\ell_1}$ and $\|\mathbf{x}\|_{\ell_2}^2$ for vectors, and $\|\mathbf{X}\|_{\ell_1}$, $\|\mathbf{X}\|_F$ and $\mathrm{Tr}(\mathbf{X})$ (which is equivalent to the nuclear norm of $\mathbf{X}$ when $\mathbf{X} \in \mathbb{S}_+$) for matrices. We can also apply our theorem for ℓ_p-norm. This is due to

the fact that replacing $\|\cdot\|_{\ell_p}$ with $\|\cdot\|_{\ell_p}^p$ does not change our estimate, and the latter is a separable function.

6.2 Main Result

In this section, we 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. So we can replace them with the Gaussian random vectors with the same mean and covariance. Next, using CGMT framework [165, 171], we analyze the phase transition in the case with Gaussian measurements, in Corollary 7. Later, we will apply this result to some well-known problems in Section 6.3.

Universality Theorem

Theorem 14 ***[non-Gaussian=Gaussian]*** *Consider the problem of recovering $\mathbf{x}_0 \in \mathcal{S} \subseteq \mathbb{R}^p$ from the measurements $\mathbf{y} = \mathbf{A}\mathbf{x}_0 \in \mathbb{R}^n$, using a convex penalty function $f(\cdot)$ in the estimator $\mathcal{E}\{\mathbf{x}_0, \mathbf{A}, \mathcal{S}, f(\cdot)\}$ in* (6.3)*. Assume $\mathcal{S}$ is a convex set and p and n are growing to infinity at a fixed rate $n = \theta(p)$. Also assume that*

1. *$f : \mathbb{R}^p \to \mathbb{R}$ is a convex function that satisfies Assumption 7.*
2. *The measurement matrix $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n]^\mathsf{T}$ satisfies Assumption 6, with $\mu := \mathbb{E}[\mathbf{a}_i]$ and $\bar{\mathbf{\Sigma}} := Cov[\mathbf{a}_i]$ for all $i = 1, \ldots, n$.*
3. *$\mathbf{G} = [\mathbf{g}_1, \ldots, \mathbf{g}_n]^\mathsf{T} \in \mathbb{R}^{n \times p}$ is a random Gaussian matrix with independent rows drawn from Gaussian distribution $\mathcal{N}(\mu, \bar{\mathbf{\Sigma}})$.*

Then the estimator $\mathcal{E}\{\mathbf{x}_0, \mathbf{A}, \mathcal{S}, f(\cdot)\}$ (introduced in Definition 3) succeeds in recovering $\mathbf{x}_0$ with probability approaching one (as p 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 14 shows that only the mean and covariance of the measurement vectors $\mathbf{a}_i$ affect the required number of measurements for perfect recovery in (6.3). Although Theorem 14 holds for n and p growing to infinity, the result of our numerical simulations in Section 6.2, indicates the validity of universality for values of p and n ranging in the order of hundreds.

Analysis of the Gaussian Estimator

Theorem 14 shows the equivalence of the convex estimator $\mathcal{E}\{\mathbf{x}_0, \mathbf{A}, \mathcal{S}, f(\cdot)\}$ and the Gaussian estimator $\mathcal{E}\{\mathbf{x}_0, \mathbf{G}, \mathcal{S}, f(\cdot)\}$. We can utilize the CGMT framework to

analyze the perfect recovery conditions for $\mathcal{E}\{\mathbf{x}_0, \mathbf{G}, \mathcal{S}, f(\cdot)\}$. Before doing so, we need the definition of the *descent cone*,

Definition 4 *[Descent Cone] The descent cone of a convex function $f(\cdot)$ at point $\mathbf{x}_0$ is defined as*

$$\mathcal{D}_f(\mathbf{x}_0) = Cone\left(\{\mathbf{y} \ : \ f(\mathbf{y}) \leq f(\mathbf{x}_0)\}\right) \ , \tag{6.4}$$

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

Corollary 7 *Consider the problem of recovering the vector $\mathbf{x}_0 \in \mathcal{S}$, given the observations $\mathbf{y} = \mathbf{G}\mathbf{x}_0 \in \mathbb{R}^n$, 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 μ and covariance $\bar{\mathbf{\Sigma}} = \mathbf{M}\mathbf{M}^\mathsf{T}$. Let $\delta := n/p$ and the set $\mathcal{S}$ and the penalty function $f(\cdot)$ be convex. $\mathcal{E}\{\mathbf{x}_0, \mathbf{G}, \mathcal{S}, f(\cdot)\}$ succeed in recovering $\mathbf{x}_0$ with probability approaching one (as p and n grow to infinity), if and only if*

$$\sqrt{\delta} > \sqrt{\delta^\star} = \mathbb{E}\left[\max_{\substack{\mathbf{w}\in(\mathcal{S}-\mathbf{x}_0)\cap D_f(\mathbf{x}_0) \\ \frac{1}{\sqrt{p}}\mathbf{M}^\mathsf{T}\mathbf{w}\in S_{p-1}}} \frac{\mathbf{w}^\mathsf{T}\mathbf{g}}{p\sqrt{1+\frac{1}{p}(\mathbf{w}^\mathsf{T}\mu)^2}}\right] \tag{6.5}$$

where S_{p-1} is the p-dimensional unit sphere, and the expected value is over the Gaussian vector $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \bar{\mathbf{\Sigma}})$.

["Pseudo Gaussian Width"] When $\mu = \mathbf{0}$ and $\bar{\mathbf{\Sigma}} = \mathbf{I}$, the expected value in (6.5) resembles the definition of the *Gaussian width* [139]. It has been shown that when the measurements are i.i.d. Gaussian, the square of the Gaussian width indicates the phase transition for linear inverse problems [6, 40, 156]. The Gaussian width has been computed for several interesting examples, such as sparse recovery, and low-rank matrix recovery. Using our universality result in Theorem 14, we can state that the square of the Gaussian width indicates the phase transition in the non-Gaussian setting as well.

Numerical Results

To validate the result of Theorem 14, we performed numerical simulations under various distributions for the measurement vectors. For our simulations in Figure 6.1, we use the estimator $\mathcal{E}\{\mathbf{x}_0, \mathbf{A}, \mathbb{R}^p, \|\cdot\|_{\ell_1}\}$ to recover a k-sparse signal $\mathbf{x}_0$ under three random ensembles for the measurement vectors $\{\mathbf{a}_i\}_{i=1}^n$. In each of the three plots,

we computed the norm of the estimation error $\mathcal{E}\{\mathbf{x}_0, \mathbf{A}, \mathbb{R}^p, \|\cdot\|_{\ell_1}\}$, for different over sampling ratios $\delta = n/p$ and multiple sparsity factors $s = k/p$. We generated the measurement vectors $\{\mathbf{a}_i\}_{i=1}^n$ for each figure, as follows,

- For each trial, we generate a random matrix $\mathbf{M} \in \mathbb{R}^{p\times p}$, with i.i.d. standard Gaussian random variables. $\bar{\mathbf{\Sigma}} = \mathbf{M}\mathbf{M}^\mathsf{T}$ will play the role of the covariance matrix of the measurement vectors.
- For Figure 6.1a, $\{\mathbf{a}_i\}_{i=1}^n$ are drawn independently from the Gaussian distribution $\mathcal{N}(\mathbf{0}, \bar{\mathbf{\Sigma}})$.
- For the measurement vectors of the Figure 6.1b, we first generate i.i.d centered bernouli vectors Ber(.8), and multiply each vector by $\mathbf{M}$.
- For the measurement vectors of the Figure 6.1c, we first generate i.i.d centered χ_1 vectors, and multiply each vector by $\mathbf{M}$.

The blue line in the figures shows the theoretical phase transition derived as a result of Corollary 7. It can be observed that the phase transition for all the three random schemes is the same, as predicted by Theorem 14. It also matches the theoretical phase transition derived from Corollary 7.

Next, to illustrate the applicability and the implications of the results, we present some examples where our universality theorem can be applied.

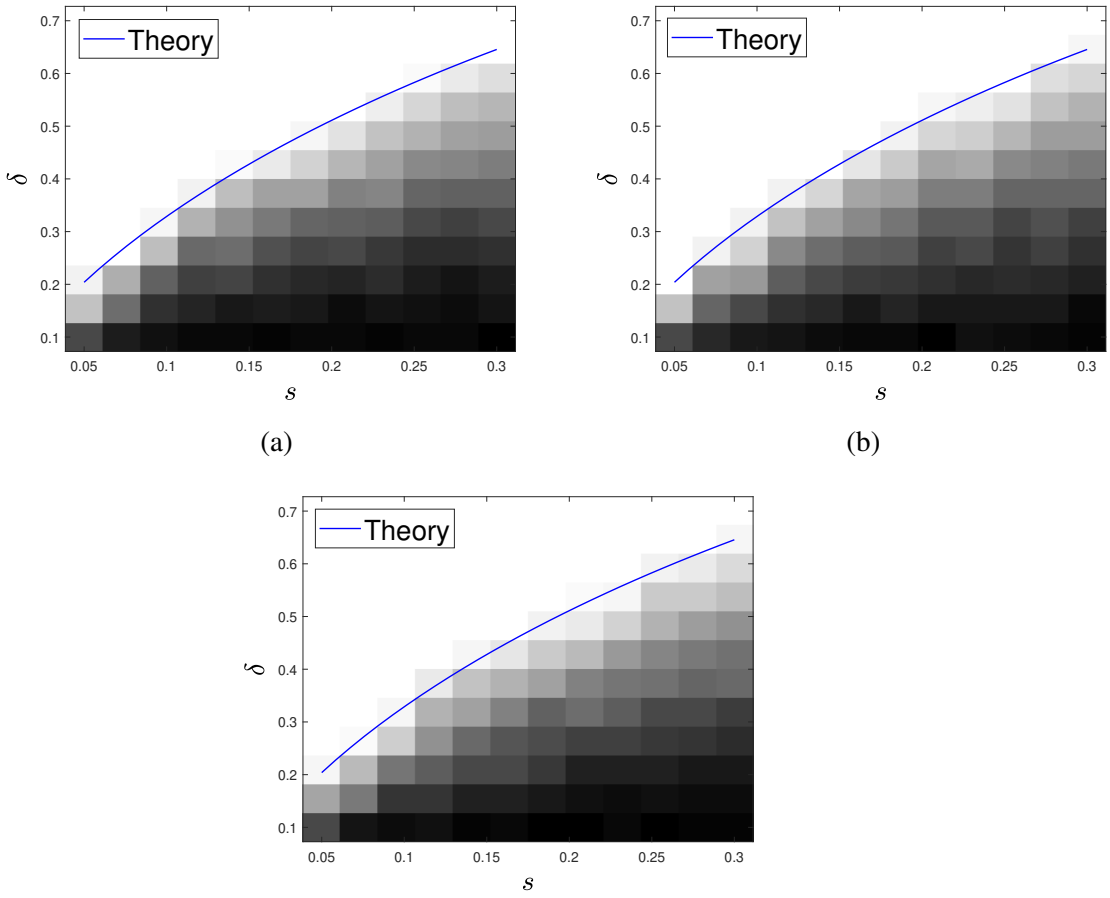
6.3 Applications: Quadratic Measurements

In this section we consider the problem of recovering a matrix from (so-called) *quadratic measurements*. The goal is to reconstruct a symmetric matrix $\mathbf{X}_0 \in \mathbb{R}^{p\times p}$ in a convex set $\mathcal{S}$, given n measurements of the form,

$$y_i = \mathbf{a}_i^\mathsf{T}\mathbf{X}_0\mathbf{a}_i = \mathrm{Tr}\left(\mathbf{X}_0 \cdot (\mathbf{a}_i\mathbf{a}_i^\mathsf{T})\right), \quad i = 1, \ldots, n\ . \tag{6.6}$$

Depending on the application, the matrix $\mathbf{X}_0$ may exhibit various structures. Similar to (6.3), we use the convex penalty function $f : \mathbb{R}^{p\times p} \to \mathbb{R}$, to enforce this structure via the following convex estimator,

$$\begin{aligned} \hat{\mathbf{X}} = \arg\min_{\mathbf{X}\in\mathcal{S}}\ & f(\mathbf{X}) \\ \text{subject to:}\quad & \mathbf{a}_i^\mathsf{T}\mathbf{X}\mathbf{a}_i = \mathbf{a}_i^\mathsf{T}\mathbf{X}_0\mathbf{a}_i, \quad i = 1, \ldots, n\ . \end{aligned} \tag{6.7}$$



(a)

(b)

(c)

Figure 6.1: Phase transition regimes for the estimator $\mathcal{E}\{\mathbf{x}_0, \mathbf{A}, \mathbb{R}^p, \|\cdot\|_{\ell_1}\}$, in terms of the oversampling ratio $\delta = \frac{n}{p}$ and $s = \frac{\|\mathbf{x}_0\|_0}{p}$, for the cases of (a) Gaussian measurements and (b) Bernoulli measurements and (c) χ^2 measurements. The blue lines indicate the theoretical estimate for the phase transition derived from Corollary 7. In the simulations we used vectors of size $p = 256$. The data is averaged over 10 independent realization of the measurements.

Note that the measurements in (6.6) are linear with respect to the matrix $\mathbf{X}_0$, yet quadratic with respect to the measurement vectors $\mathbf{a}_i$. We can define $\tilde{\mathbf{x}}_0 := \text{Vec}(\mathbf{X}_0) \in \mathbb{R}^{p^2}$ and $\tilde{\mathbf{a}}_i := \text{Vec}(\mathbf{a}_i\mathbf{a}_i^\top) \in \mathbb{R}^{p^2}$, such that the measurements take the familiar form, $y_i = \tilde{\mathbf{a}}_i^\top \tilde{\mathbf{x}}_0$. In order to apply the result of Theorem 14, one should check if the vectors $\{\tilde{\mathbf{a}}_i\}_{i=1}^n$ satisfy Assumption 6.
It can be shown that if the vectors $\{\mathbf{a}_i\}_{i=1}^n$ satisfy the following conditions, then Assumption 6 holds true for $\{\tilde{\mathbf{a}}_i = \text{Vec}(\mathbf{a}_i\mathbf{a}_i^\top)\}_{i=1}^n$.

Assumption 8 *We say vectors $\{\mathbf{a}_i\}_{i=1}^n$ satisfy Assumption 3, 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. Therefore, when Assumption 8 holds, we can apply Theorem 14 to show that the required number of measurements for perfect recovery in (6.7) is equal to the required number of measurements for the success of the following estimator,

$$\begin{aligned} \hat{\mathbf{X}} = \arg\min_{\mathbf{X}\in\mathcal{S}} \; & f(\mathbf{X}) \\ \text{subject to:} \quad & \text{Tr}\left((\mathbf{H}_i + \mathbf{I})\mathbf{X}\right) = \text{Tr}\left((\mathbf{H}_i + \mathbf{I})\mathbf{X}_0\right), \quad i = 1, \ldots, n \,, \end{aligned} \tag{6.8}$$

where $\mathbf{I}$ is the $p \times p$ identity matrix and $\mathbf{H}_i$'s are independent Gaussian Wigner matrices (defined in Section 6.1). Corollary 8 presents a formal statement.

Corollary 8 *Consider the problem of recovering the matrix $\mathbf{X}_0 \in \mathcal{S} \subseteq \mathbb{R}^{p\times p}$, from n quadratic measurements of the form* (6.6)*, using the estimator* (6.7)*. Let $\mathcal{S}$ and $f(\cdot)$ be convex set and function satisying Assumption 7. Assume,*

- *The measurement vectors $\{\mathbf{a}_i\}_{i=1}^n$ satisfy Assumption 8, and,*
- *$\{\mathbf{H}_i \in \mathbb{R}^{p\times p}\}_{i=1}^n$ is a set of independent Gaussian Wigner matrices.*

Then, as n and p grow to infinity at a fixed rate $n = \theta(p)$, the estimator (6.7) *perfectly recovers $\mathbf{X}_0$ with probability approaching one if and only if the estimator* (6.8) *perfectly recovers $\mathbf{X}_0$ with probability approaching one.*

Therefore, in order to find the phase transition, it is sufficient to analyze the equivalent optimization (6.8) which is possible via the CGMT framework. Proceeding onward, we exploit the CGMT framework along with Corollary 7 to find the required number of measurements for the recovery of $\mathbf{X}_0$ in two specific applications.

Low-rank Matrix Recovery

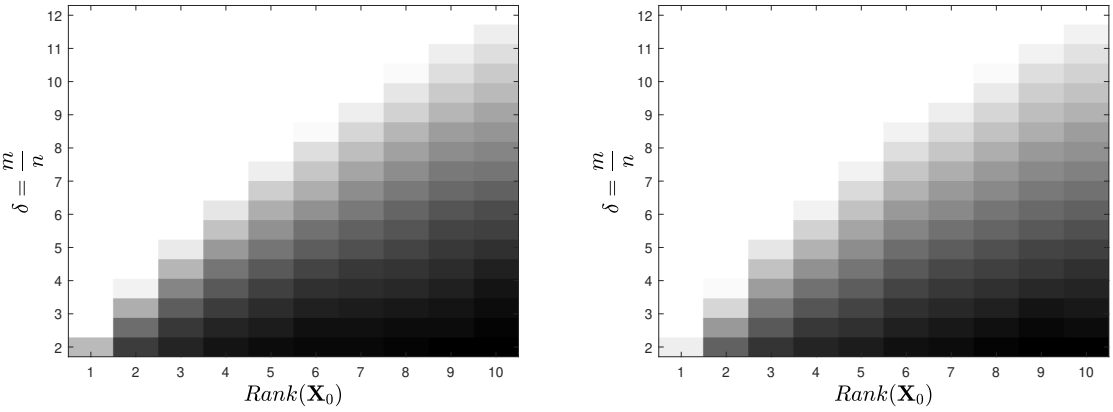
Assume the unknown matrix $\mathbf{X}_0 \succeq \mathbf{0}$ has rank r, where r is a constant (i.e., r does not grow with problem dimensions n, p.) Such matrices appear in many applications such as traffic data monitoring, array signal processing and phase retrieval. The nuclear norm, $||\cdot||_\star$, is often used as the convex surrogate for low-rank matrix recovery [135]. Hence, we are interested in analyzing the optimization (6.7), with the choice of $f(\mathbf{X}) = \|\mathbf{X}\|_\star$, where the optimization is over the set of PSD matrices. Note that $\mathrm{Tr}(\cdot) = ||\cdot||_\star$ within this set, which satisfies Assumption 7.
According to Corollary 8, the perfect recovery in (6.7) is equivalent to perfect recovery in (6.8), where the same choice of $f(\mathbf{X}) = \mathrm{Tr}(\mathbf{X})$. The analysis of the later through CGMT yields the following corollary.

Corollary 9 *Consider the optimization program* (6.7)*, where the matrix* $\mathbf{X}_0 \succeq 0$ *has rank* r*,* $f(\mathbf{X}) = Tr(\mathbf{X})$*, the set* $\mathcal{S}$ *is the PSD cone and the measurement vectors* $\{\mathbf{a}_i\}_{i=1}^n$ *satisfy Assumption 8. Assume* $p, n \to \infty$ *at the proportional rate* $\delta := \frac{n}{p} \in (0, +\infty)$*. The estimator perfectly recovers* $\mathbf{X}_0$ *if* $\delta > 3r$*.*

Corollary 9 indicates that $3rp$ measurements is needed to perfectly recover a rank-r PSD matrix $\mathbf{X}_0$, from quadratic measurements. Although, the error of estimation gets extremely small, much before the threshold $n = 3pr$. To the extent of our knowledge, this is the first work that precisely computes the phase transition of low-rank matrix recovery from quadratic measurements. Figure6.2 depicts the result of numerical simulations. For different values of r and δ, the Frobenius norm of the error of the estimators (6.7) and (6.8) has been computed, which shows the same phase transition in both cases.

Phase Transition of PhaseLift in Phase Retrieval

An important application for the result of Corollary 9, is when the underlying matrix $\mathbf{X}_0$ is of rank 1. This appears in the problem of phase retrieval, where $\mathbf{X}_0 = \mathbf{x}_0\mathbf{x}_0^\mathsf{T}$ is the lifted version of the signal. The optimization program (6.7) with $f(\mathbf{X}) = \mathrm{Tr}(\mathbf{X})$ in this case, is known as PhaseLift [36]. Corollary 9 states that the phase transition of



(a) (b)

Figure 6.2: Phase transition regimes for both estimators 6.7 and (6.8), with $f(\mathbf{X}) = \mathrm{Tr}(\mathbf{X})$, in terms of the oversampling ratio $\delta = \frac{n}{p}$ and $r = \mathrm{Rank}(\mathbf{X}_0)$, for the cases of (a) estimator (6.7) with quadratic measurements and (b) estimator (6.8) with Gaussian measurements. In the simulations we used matrices of size $p = 40$. The data is averaged over 20 independent realization of the measurements.

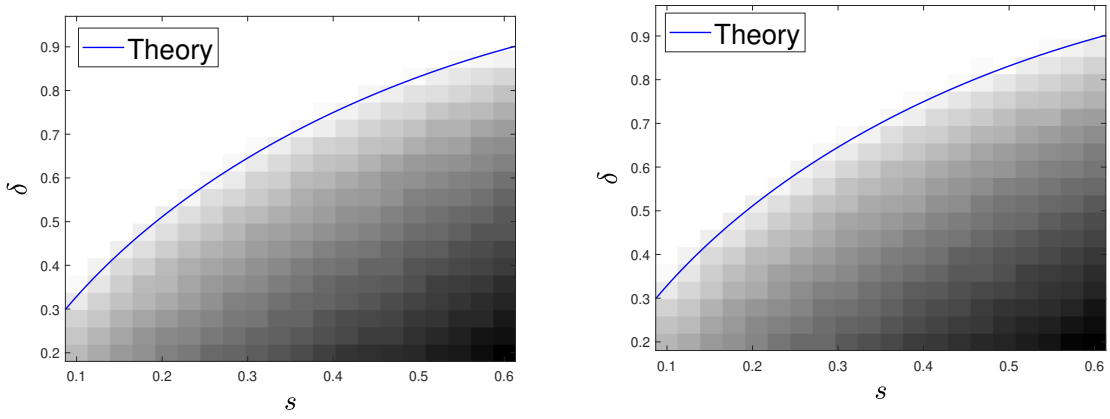
the PhaseLift algorithm happens at $\delta^\star = 3$, i.e., $n > 3p$ measurements is needed for the perfect signal reconstruction in PhaseLift. We should emphasize the significance of this result as establishing the exact phase transition of the PhaseLift algorithm was long an open problem.

Sparse Matrix Recovery

Let $\mathbf{X}_0 \succeq 0$ represent the covariance matrix of a set of random variables. In certain applications, the covariance matrix has many near-zero entries as the correlations are small for many pairs of random variables. Such matrices arise in applications in spectrum estimation, biology and finance [44, 67]. We are interested in analyzing estimator (6.7), where $f(\mathbf{X}) = \|\mathbf{X}\|_{\ell_1}$ promotes the sparsity in the optimization. As $\|\cdot\|_{\ell_1}$ satisfies Assumption 7, applying the result of Corollary 8, the perfect recovery in (6.7) is equivalent to the perfect recovery in the estimator (6.8), with the same penalty function. Analyzing the optimization (6.8) via CGMT leads to the following result:

Corollary 10 *Let $\delta := \frac{n}{p^2}$, $s := \frac{\|\mathbf{X}_0\|_0}{p^2}$. As $p \to \infty$, the optimization program* (6.7)*, with $f(\mathbf{X}) = \|\mathbf{X}\|_{\ell_1}$ can successfully recover the signal iff $\delta > \delta^\star$, where $\delta^\star$ is the unique solution to the following nonlinear equation,*

$$x \cdot Q^{-1}\left(\frac{2x-s}{2-2s}\right) = (1-s)\phi\left(Q^{-1}\left(\frac{2x-s}{2-2s}\right)\right), \tag{6.9}$$



(a) (b)

Figure 6.3: Phase transition regimes for both estimators (6.7) and (6.8), with $f(\mathbf{X}) = \|\mathbf{X}\|_{\ell_1}$, in terms of the oversampling ratio $\delta = \frac{n}{p^2}$ and $s = \frac{\|\mathbf{X}_0\|_0}{p^2}$, for the cases of (a) estimator (6.7) with quadratic measurements and (b) estimator (6.8) with Gaussian measurements. The blue lines indicate the theoretical estimate for the phase transition derived from equation (6.9). In the simulations we used matrices of size $p = 40$. The data is averaged over 20 independent realization of the measurements.

Model	Penalty function $f(\cdot)$	No. of required measurements
k sparse matrix	$\|\cdot\|_{\ell_1}$	$p^2\delta^\star$ defined in (6.9)
Rank-r PSD matrix	$\mathrm{Tr}(\cdot)$	3pr
S&L (k, r) matrix	$\mathrm{Tr}(\cdot) + \lambda\|\cdot\|_1$	$O(\min(k^2, rp))$

Table 6.1: Summary of the parameters that are discussed in this section. The last row is for a $p \times p$ rank-r matrix whose smallest sub-matrix with non-zero entries is k by k. The third column shows the number of required quadratic measurements for perfect recovery.

where $\phi(x) = \exp(-x^2/2)/\sqrt{2\pi}$ *and* $Q^{-1}(\cdot)$ *is inverse of the Q-function.*

Figure 6.3b compares the empirical result with the theoretical phase transition derived from Corollary 10 Each plot shows the norm of the error with respect to the sparsity of the matrix $\mathbf{X}_0$ and the ratio $\delta = \frac{n}{p^2}$. A comparison between the two plots indicates that the phase transitions of the two estimators (6.7) and (6.8) with $f(\mathbf{X}) = \|\mathbf{X}\|_{\ell_1}$ match.

6.4 Simultaneously Sparse and Low-rank Matrices

Another interesting example is where the unknown matrix $\mathbf{X}_0 \succeq 0$ is simultaneously sparse and low rank. To recover $\mathbf{X}_0$, we would like to simultaneously minimize the

penalty functions $f^{(1)}(\mathbf{X}) = \|\mathbf{X}\|_{\ell_1}$ and $f^{(2)}(\mathbf{X}) = \|\mathbf{X}\|_\star$, for all feasible matrices $\mathbf{X} \in \mathcal{S}$ that align our measurements in (6.6). Here, each function $f^{(i)}(\cdot)$ enforces one of the structures on $\mathbf{X}$. So, a natural choice for the regularizer function in (6.7) would be $f(\mathbf{X}) = f^{(1)}(\mathbf{X}) + \lambda f^{(2)}(\mathbf{X})$, where λ is a regularizing parameter. Oymak et al [128] studied phase transition for perfect recovery of simultaneously structured matrices. Their results are based on Gordon's comparison lemma which is only applicable to the cases of linear Gaussian measurements. We can use the result of Corollary 8 to extend their result to settings with quadratic measurements, as the phase transition regime is equivalent in both cases. Let $\mathbf{X}_0 \in \mathbb{R}^{p\times p}$ be a rank-r PSD matrix. Also assume that the largest sub-matrix in $\mathbf{X}_0$ that contains all non-zero entries is k by k. If we choose $f(\mathbf{X}) = \|\mathbf{X}\|_{\ell_1} + \lambda \mathrm{Tr}(\mathbf{X})$, they show that $O(\min(k^2, rp))$ measurements is required for perfect recovery.

Conclusion

We have investigated an estimation problem under linear observations. We aimed to characterize the minimum number of observations that are needed for perfect recovery of the unknown model. Our main result indicated that this phase transition, only depends on the first two statistics of the measurement vector. Therefore, it remains unchanged as we replace these vectors with the Gaussian one, with the same mean vector and covariance matrix. The later can be analyzed through existing frameworks such as CGMT. As one of the applications of this universality, we investigated the case of matrix recovery via the so called quadratic measurements, and derived the minimum number of observations required for the recovery of a structured matrix. Due to the space constraint, we moved the discussions regarding the case of simultaneously structured matrices to the appendix. Table 6.1, summarizes these results for the cases of three structures.

Proof of Theorem 14

Consider the following optimization

$$\Phi_1 = \min_{\mathbf{A}\mathbf{x}_0=\mathbf{A}\mathbf{x}} f(\mathbf{x}) \,, \tag{6.10}$$

Without loss of generality, assume that $f(0) = 0$. We change the variable to $\mathbf{w} = \mathbf{x} - \mathbf{x}_0$, which gives the following

$$\Phi_1 = \min_{\mathbf{A}\mathbf{w}=0} f(\mathbf{w} + \mathbf{x}_0) \,, \tag{6.11}$$

This optimization has perfect recovery, iff $\hat{\mathbf{w}} = 0$, or equivalently iff $\Phi_1 = 0$. We would like to show that if $\Phi_1 = 0$ with probability converging to 1, then the same

holds if we replace the measurements vectors $\mathbf{a}_i$, with another set of measurement vectors with the same mean and covariance. We rewrite this optimization in the form of this min-max optimization,

$$\begin{aligned}\Phi_1 &= \sup_{\lambda>0} \min_{\mathbf{w}} \frac{\lambda}{2}\|\mathbf{A}\mathbf{w}\|^2 + f(\mathbf{w}+\mathbf{x}_0) \\ &= \sup_{\lambda>0} \min_{\mu>0} \min_{\mathbf{w}} \frac{\lambda}{2}\|\mathbf{A}\mathbf{w}\|^2 + f(\mathbf{w}+\mathbf{x}_0) + \frac{1}{2\mu}\|\mathbf{w}\|^2 \\ &= \sup_{\lambda>0} \lambda \cdot \min_{\mu>0} \min_{\mathbf{w}} \frac{1}{2}\|\mathbf{A}\mathbf{w}\|^2 + \frac{1}{\lambda} f(\mathbf{w}+\mathbf{x}_0) + \frac{1}{2\lambda\mu}\|\mathbf{w}\|^2 \end{aligned} \tag{6.12}$$

Informally, we first show that for fixed values of λ and μ, the values of last minimization remains unchanged as we change the random measurement vectors inside it (as p and n grow to infinity). Next, we use Lemma 18 (See [165] Section A.4 and B.5) to switch the min-max over μ and λ, with the limit over p and n.

By fixing the values of λ and μ, from now on, we redefine the function $f(\cdot)$ to be $\frac{1}{\lambda} f(\mathbf{w}+\mathbf{x}_0) + \frac{1}{2\lambda\mu}\|\mathbf{w}\|^2$, which is strongly convex. Note that we would like the following assumptions holds for these two set of random measurement vectors.

Assumption 1: Assume $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n]^\mathsf{T} \in \mathbb{R}^{n\times p}$ and $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_n]^\mathsf{T} \in \mathbb{R}^{n\times p}$ are two random matrices, such that

$$\begin{aligned} &\mathbf{e} = \mathbb{E}\left[\mathbf{a}_i\right] = \mathbb{E}\left[\mathbf{b}_i\right] \quad \forall i \\ &\mathbf{\Sigma} = \mathbb{E}\left[\mathbf{a}_i\mathbf{a}_i^\mathsf{T}\right] = \mathbb{E}\left[\mathbf{b}_i\mathbf{b}_i^\mathsf{T}\right] \quad \forall i \\ &\lim_{p\to\infty} \frac{\|\mathbf{e}\|^2}{p^2} = 0, \end{aligned} \tag{6.13}$$

Besides, there exists $\tau > 0$ such that for any matrix $\mathbf{M} \in \mathbb{R}^{p\times p}$ such that $\|\mathbf{M}\|_2 \le \kappa$, there exists some c that only depends on κ that

$$\begin{aligned} &\frac{1}{p^2}\mathrm{Var}\left(\mathbf{a}_i^\mathsf{T}\mathbf{M}\mathbf{a}_i\right) \le c\cdot p^{-\tau} \quad \text{and,} \\ &\frac{1}{p^2}\mathrm{Var}\left(\mathbf{b}_i^\mathsf{T}\mathbf{M}\mathbf{b}_i\right) \le c\cdot p^{-\tau}\ . \end{aligned} \tag{6.14}$$

Now we want to investigate equivalence of the following two optimizations. Let $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n]$ and $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_n]$ be n by p measurement matrices and

$$\begin{aligned} \Phi_{\mathbf{B}} &= \min_{\mathbf{w}} \frac{1}{2n}\sum_{i=1}^{n}\left(z_i - \mathbf{w}^\mathsf{T}\mathbf{a}_i\right)^2 + f\left(\mathbf{w}+\mathbf{x}_0\right)\ , \\ \Phi_{\mathbf{A}} &= \min_{\mathbf{w}} \frac{1}{2n}\sum_{i=1}^{n}\left(z_i - \mathbf{w}^\mathsf{T}\mathbf{b}_i\right)^2 + f\left(\mathbf{w}+\mathbf{x}_0\right)\ . \end{aligned} \tag{6.15}$$

Theorem 15 *Consider the optimizations in* (6.15)*. If*

$$\lim_{n,p\to\infty} |\mathbb{E}\left[\Phi_{\mathbf{B}} - \Phi_{\mathbf{A}}\right]| = 0\,, \tag{6.16}$$

and if for constants C and $\delta > 0$,

$$Pr\left(|\Phi_{\mathbf{A}} - C| > \delta\right) \xrightarrow{P} 0\,, \tag{6.17}$$

as $n, p \to \infty$. Then,

$$Pr\left(|\Phi_{\mathbf{B}} - C| > 3\delta\right) \xrightarrow{P} 0\,, \tag{6.18}$$

Proof 7 *We first define the function $g : \mathbb{R} \to \mathbb{R}$ as follows.*

$$g(x) = \begin{cases} 0 & \text{if} \quad |x| \leq 1, \\ (|x| - 1)^2 & \text{if} \quad 1 < |x| \leq 2, \\ 2 - (|x| - 3)^2 & \text{if} \quad 2 < |x| \leq 3, \\ 2 & \text{if} \quad |x| > 3\,. \end{cases} \tag{6.19}$$

Note that $g(.)$ is continuously differentiable with its first derivative bounded by 2. Now,

$$\begin{aligned} Pr\left\{|\Phi_{\mathbf{B}} - C| > 3\delta\right\} &= Pr\left\{g\left(\frac{\Phi_{\mathbf{B}} - C}{\delta}\right) > 2\right\} \leq \frac{1}{2}\mathbb{E}\left[g\left(\frac{\Phi_{\mathbf{B}} - C}{\delta}\right)\right] \\ &\leq \frac{1}{2}\mathbb{E}\left[g\left(\frac{\Phi_{\mathbf{A}} - C}{\delta}\right)\right] + \frac{1}{2}\left|\mathbb{E}\left[g\left(\frac{\Phi_{\mathbf{A}} - C}{\delta}\right) - g\left(\frac{\Phi_{\mathbf{B}} - C}{\delta}\right)\right]\right| \\ &\leq Pr\left\{|\Phi_{\mathbf{A}} - C| > \delta\right\} + \frac{1}{2}\left|\mathbb{E}\left[g'(\zeta) \cdot \left(\frac{\Phi_{\mathbf{A}} - C}{\delta} - \frac{\Phi_{\mathbf{B}} - C}{\delta}\right)\right]\right| \\ &\leq Pr\left\{|\Phi_{\mathbf{A}} - C| > \delta\right\} + \frac{1}{\delta}\left|\mathbb{E}\left[\Phi_{\mathbf{A}} - \Phi_{\mathbf{B}}\right]\right| \xrightarrow{n,p\to\infty} 0 \end{aligned} \tag{6.20}$$

Theorem 16 *Consider the optimizations in* (6.15)*. If* $\mathbf{A}$*,* $\mathbf{B}$ *and $f(.)$ satisfy Assumption 6 and 7, respectively, then*

$$\lim_{n,p\to\infty} |\mathbb{E}\left[\Phi_{\mathbf{A}}\right] - \mathbb{E}\left[\Phi_{\mathbf{B}}\right]| \to 0\,. \tag{6.21}$$

Proof 8 *For $k = 0, \ldots, n$, we define*

$$\Phi_k := \min_{\mathbf{w}} \frac{1}{2n}\sum_{i=1}^{k}\left(z_i - \mathbf{a}_i^{\mathsf{T}}\mathbf{w}\right)^2 + \frac{1}{2n}\sum_{i=k+1}^{n}\left(z_i - \mathbf{b}_i^{\mathsf{T}}\mathbf{w}\right)^2 + f\left(\mathbf{w} + \mathbf{x}_0\right)\,. \tag{6.22}$$

We have

$$|\mathbb{E}\left[\Phi_{\mathbf{A}} - \Phi_{\mathbf{B}}\right]| = |\mathbb{E}\left[\Phi_n - \Phi_0\right]| \leq \sum_{k=1}^{n} |\mathbb{E}\left[\Phi_k - \Phi_{k-1}\right]| \ . \tag{6.23}$$

Now it suffices to show that there exists a constant c, such that for any k,

$$|\mathbb{E}\left[\Phi_k - \Phi_{k-1}\right]| \leq c\ n^{-(1+\tau/2)} \ , \tag{6.24}$$

for some positive constant τ*. Since, then combining* (6.24) *and* (6.23) *yields,*

$$|\mathbb{E}\left[\Phi_{\mathbf{A}} - \Phi_{\mathbf{B}}\right]| \leq \sum_{k=1}^{n} |\mathbb{E}\left[\Phi_k - \Phi_{k-1}\right]| \leq c\ n^{-\tau/2} \to 0 \ . \tag{6.25}$$

Let

$$\begin{aligned} &\mathbf{M}_k = [\mathbf{a}_1, \ldots, \mathbf{a}_{k-1}, \mathbf{b}_{k+1}, \ldots, \mathbf{b}_n]^\mathsf{T} \in \mathbb{R}^{(n-1)\times p}, \quad \textit{and,} \\ &\mathbf{z}_k = [z_1, \ldots, z_{k-1}, z_{k+1}, \ldots, z_n]^\mathsf{T} \in \mathbb{R}^{n-1} \ . \end{aligned} \tag{6.26}$$

This helps us rewrite Φ_k *and* Φ_{k-1} *as*

$$\begin{aligned} \Phi_k &= \min_{\mathbf{w}} \ \frac{1}{2n}\|\mathbf{z}_k - \mathbf{M}_k\mathbf{w}\|^2 + \frac{1}{2n}\left(z_k - \mathbf{a}_k^\mathsf{T}\mathbf{w}\right)^2 + f\left(\mathbf{w} + \mathbf{x}_0\right) \ , \\ \Phi_{k-1} &= \min_{\mathbf{w}} \ \frac{1}{2n}\|\mathbf{z}_k - \mathbf{M}_k\mathbf{w}\|^2 + \frac{1}{2n}\left(z_k - \mathbf{b}_k^\mathsf{T}\mathbf{w}\right)^2 + f\left(\mathbf{w} + \mathbf{x}_0\right) \ . \end{aligned} \tag{6.27}$$

As of this point, we fix k and drop the subscript k from z_k*,* $\mathbf{z}_k$*,* $\mathbf{M}_k$*,* $\mathbf{a}_k$ *and* $\mathbf{b}_k$ *for simplicity. The expectation in* (6.24) *is over the randomness in z,* $\mathbf{z}$*,* $\mathbf{M}$*,* $\mathbf{a}$ *and* $\mathbf{b}$*, which can be written as*

$$\begin{aligned} |\mathbb{E}\left[\Phi_k - \Phi_{k-1}\right]| &= \left|\mathbb{E}_{\{\mathbf{M},\mathbf{z}\}}\left[\mathbb{E}_{\{z,\mathbf{a},\mathbf{b}\}}\left[\Phi_k - \Phi_{k-1}\middle|\{\mathbf{M},\mathbf{z}\}\right]\right]\right| \\ &\leq \mathbb{E}_{\{\mathbf{M},\mathbf{z}\}}\left[\left|\mathbb{E}_{\{z,\mathbf{a},\mathbf{b}\}\middle|\{\mathbf{M},\mathbf{z}\}}\left[\Phi_k - \Phi_{k-1}\right]\right|\right] \ . \end{aligned} \tag{6.28}$$

We first fix $\mathbf{M}$ *and* $\mathbf{z}$*, and bound the inner expectation in* (6.28)*. Now let,*

$$\begin{aligned} &\phi(\mathbf{a}, z, \mathbf{w}) = \frac{1}{2n}\|\mathbf{z} - \mathbf{M}\mathbf{w}\|^2 + \frac{1}{2n}\left(z - \mathbf{a}^\mathsf{T}\mathbf{w}\right)^2 + f\left(\mathbf{w} + \mathbf{x}_0\right), \\ &\Phi(\mathbf{a}, z) = \min_{\mathbf{w}} \ \phi(\mathbf{a}, \mathbf{w}) \ , \\ &\bar{\Phi} = \Phi(\mathbf{0}, 0), \ \textit{and,} \quad \bar{\mathbf{w}} = \arg\min \ \phi(\mathbf{0}, 0, \mathbf{w}) \ . \end{aligned} \tag{6.29}$$

With these new definitions, we have $\Phi_k = \Phi(\mathbf{a}, z)$ *and* $\Phi_{k-1} = \Phi(\mathbf{b}, z)$ *and thus,*

$$\begin{aligned}\left|\mathbb{E}_{\{z,\mathbf{a},\mathbf{b}\}}\left[\Phi_k - \Phi_{k-1}\right]\right| &= \left|\mathbb{E}_{\{z,\mathbf{a},\mathbf{b}\}}\left[\Phi(\mathbf{a}, z) - \Phi(\mathbf{b}, z)\right]\right| \\ &\leq \left|\mathbb{E}_{\{z,\mathbf{a}\}}\left[\Phi(\mathbf{a}, z) - \bar{\Phi} - \frac{\sigma^2 + \frac{\|\bar{\mathbf{w}}\|^2}{n}}{2n(1 + \mathbb{E}[\mathbf{b}^\mathsf{T}\boldsymbol{\Omega}\mathbf{b}])}\right]\right| \\ &+ \left|\mathbb{E}_{\{z,\mathbf{b}\}}\left[\Phi(\mathbf{b}, z) - \bar{\Phi} - \frac{\sigma^2 + \frac{\|\bar{\mathbf{w}}\|^2}{n}}{2n(1 + \mathbb{E}[\mathbf{b}^\mathsf{T}\boldsymbol{\Omega}\mathbf{b}])}\right]\right| \end{aligned} \tag{6.30}$$

So since $\mathbb{E}[\mathbf{b}^\mathsf{T}\boldsymbol{\Omega}\mathbf{b}] = \mathbb{E}[\mathbf{a}^\mathsf{T}\boldsymbol{\Omega}\mathbf{a}]$*, it remains to show that for positive constants* c *and* τ*,*

$$\begin{aligned}&\left|\mathbb{E}_{\{z,\mathbf{a}\}}\left[\Phi(\mathbf{a}, z) - \bar{\Phi} - \frac{\sigma^2 + \frac{\|\bar{\mathbf{w}}\|^2}{n}}{2n(1 + \mathbb{E}[\mathbf{a}^\mathsf{T}\boldsymbol{\Omega}\mathbf{a}])}\right]\right| \leq c\, n^{-(1+\tau/2)}\,, \quad \textit{and,} \\ &\left|\mathbb{E}_{\{z,\mathbf{b}\}}\left[\Phi(\mathbf{b}, z) - \bar{\Phi} - \frac{\sigma^2 + \frac{\|\bar{\mathbf{w}}\|^2}{n}}{2n(1 + \mathbb{E}[\mathbf{b}^\mathsf{T}\boldsymbol{\Omega}\mathbf{b}])}\right]\right| \leq c\, n^{-(1+\tau/2)}\,. \end{aligned} \tag{6.31}$$

We show the later, and the proof of the first is similar. Define $\mathbf{v} = \frac{\partial f(\bar{\mathbf{w}}+\mathbf{x}_0)}{\partial \mathbf{w}}$ *and* $\mathbf{V} = \frac{\partial^2 f(\bar{\mathbf{w}}+\mathbf{x}_0)}{\partial \mathbf{w}^2}$ *and*

$$\begin{aligned}\psi(\mathbf{b}, z, \mathbf{w}) &= \frac{1}{2n}\|\mathbf{z} - \mathbf{M}\mathbf{w}\|^2 + \frac{1}{2n}\left(z - \mathbf{b}^\mathsf{T}\mathbf{w}\right)^2 + f\left(\bar{\mathbf{w}} + \mathbf{x}_0\right) + \mathbf{v}^\mathsf{T}(\mathbf{w} - \bar{\mathbf{w}}) \\ &\quad + \frac{1}{2}(\mathbf{w} - \bar{\mathbf{w}})^\mathsf{T}\mathbf{V}(\,\mathbf{w} - \bar{\mathbf{w}})\,, \\ \Psi(\mathbf{b}, z) &= \min_{\mathbf{w}}\ \psi(\mathbf{b}, z, \mathbf{w})\,, \quad \textit{and,} \quad \tilde{\mathbf{w}} = \arg\min\ \psi(\mathbf{b}, z, \mathbf{w})\,. \end{aligned} \tag{6.32}$$

Note that by writing the optimality conditions, it is easy to show that $\Psi(\mathbf{0}, 0) = \Phi(\mathbf{0}, 0) = \bar{\Phi}$*. Thus,*

$$\begin{aligned}\mathbb{E}_{\{z,\mathbf{b}\}}\left|\left[\Phi(\mathbf{b}, z) - \bar{\Phi} - \frac{\sigma^2 + \frac{\|\bar{\mathbf{w}}\|^2}{n}}{2n(1 + \mathbb{E}[\mathbf{b}^\mathsf{T}\boldsymbol{\Omega}\mathbf{b}])}\right]\right| &\leq \mathbb{E}_{\{z,\mathbf{b}\}}\left[|\Phi(\mathbf{b}, z) - \Psi(\mathbf{b}, z)|\right] \\ &+ \left|\mathbb{E}_{\{z,\mathbf{b}\}}\left[\Psi(\mathbf{b}, z) - \Psi(\mathbf{0}, 0) - \frac{\sigma^2 + \frac{\|\bar{\mathbf{w}}\|^2}{n}}{2n(1 + \mathbb{E}[\mathbf{b}^\mathsf{T}\boldsymbol{\Omega}\mathbf{b}])}\right]\right|. \end{aligned} \tag{6.33}$$

So we have to bound the two terms on the right hand side of (6.33)*. We start with bounding* $\mathbb{E}_{\{z,\mathbf{b}\}}\left[|\Phi(\mathbf{b}, z) - \Psi(\mathbf{b}, z)|\right]$*. Note that for any* $\mathbf{w}$ *we have*

$$|\psi(\mathbf{b}, z, \mathbf{w}) - \phi(\mathbf{b}, z, \mathbf{w})| \leq \frac{C_f}{n}\|\mathbf{w} - \bar{\mathbf{w}}\|_3^3\,. \tag{6.34}$$

Besides, due to strong convexity of $\bar{f}(.)$ *we have,*

$$|\psi(\mathbf{b}, z, \mathbf{w}) - \Psi(\mathbf{b}, z)| \geq \frac{\epsilon}{n}\|\mathbf{w} - \tilde{\mathbf{w}}\|_2^2 \,. \tag{6.35}$$

We have two cases.

First if $\|\tilde{\mathbf{w}} - \bar{\mathbf{w}}\|_3 \leq \frac{\epsilon}{9\,C_f}$. *Consider the set* $\mathcal{S} = \{\mathbf{w} \,:\, \|\mathbf{w} - \tilde{\mathbf{w}}\|_3 = \|\tilde{\mathbf{w}} - \bar{\mathbf{w}}\|_3\}$. *For any* $\mathbf{w}$ *in the set* $\mathcal{S}$ *we have*

$$\begin{aligned}
\phi(\mathbf{b}, z, \mathbf{w}) - \phi(\mathbf{b}, z, \tilde{\mathbf{w}}) &\geq \psi(\mathbf{b}, z, \mathbf{w}) - \psi(\mathbf{b}, z, \tilde{\mathbf{w}}) - \frac{C_f}{n}\left(\|\mathbf{w} - \bar{\mathbf{w}}\|_3^3 + \|\tilde{\mathbf{w}} - \bar{\mathbf{w}}\|_3^3\right) \\
&\geq \frac{\epsilon}{n}\|\mathbf{w} - \tilde{\mathbf{w}}\|_2^2 - \frac{C_f}{n}\left(\|\mathbf{w} - \bar{\mathbf{w}}\|_3^3 + \|\tilde{\mathbf{w}} - \bar{\mathbf{w}}\|_3^3\right) \\
&\geq \frac{\epsilon}{n}\|\mathbf{w} - \tilde{\mathbf{w}}\|_3^2 - \frac{C_f}{n}\left(4\,\|\mathbf{w} - \bar{\mathbf{w}}\|_3^3 + 5\,\|\tilde{\mathbf{w}} - \bar{\mathbf{w}}\|_3^3\right) \\
&= \frac{9\,C_f}{n}\|\tilde{\mathbf{w}} - \bar{\mathbf{w}}\|_3^2\left(\frac{\epsilon}{9\,C_f} - \|\tilde{\mathbf{w}} - \bar{\mathbf{w}}\|_3\right) \geq 0 \,.
\end{aligned} \tag{6.36}$$

This means that the optimal value of $\phi(\mathbf{b}, z, \mathbf{w})$ *lies within* $\mathcal{S}$. *Now if* $\mathbf{w}_\phi = \arg\min \phi(\mathbf{b}, z, \mathbf{w})$,

$$\begin{aligned}
\Psi(\mathbf{b}, z) - \Phi(\mathbf{b}, z) &= \left(\psi(\mathbf{b}, z, \tilde{\mathbf{w}}) - \psi(\mathbf{b}, z, \mathbf{w}_\phi)\right) + \left(\psi(\mathbf{b}, z, \mathbf{w}_\phi) - \phi(\mathbf{b}, z, \mathbf{w}_\phi)\right) \\
&\leq \left(\psi(\mathbf{b}, z, \mathbf{w}_\phi) - \phi(\mathbf{b}, z, \mathbf{w}_\phi)\right) \leq \frac{C_f}{n}\|\mathbf{w}_\phi - \bar{\mathbf{w}}\|_3^3 \\
&\leq \frac{4\,C_f}{n}\left(\|\mathbf{w}_\phi - \tilde{\mathbf{w}}\|_3^3 + \|\tilde{\mathbf{w}} - \bar{\mathbf{w}}\|_3^3\right) \leq \frac{8\,C_f}{n}\|\tilde{\mathbf{w}} - \bar{\mathbf{w}}\|_3^3 \,.
\end{aligned} \tag{6.37}$$

And,

$$\begin{aligned}
\Phi(\mathbf{b}, z) - \Psi(\mathbf{b}, z) &= \left(\phi(\mathbf{b}, z, \mathbf{w}_\phi) - \phi(\mathbf{b}, z, \tilde{\mathbf{w}})\right) + \left(\phi(\mathbf{b}, z, \tilde{\mathbf{w}}) - \psi(\mathbf{b}, z, \tilde{\mathbf{w}})\right) \\
&\leq \left(\phi(\mathbf{b}, z, \tilde{\mathbf{w}}) - \psi(\mathbf{b}, z, \tilde{\mathbf{w}})\right) \leq \frac{C_f}{n}\|\tilde{\mathbf{w}} - \bar{\mathbf{w}}\|_3^3 \,.
\end{aligned} \tag{6.38}$$

Thus, (6.37) *and* (6.37) *implies that*

$$|\Phi(\mathbf{b}, z) - \Psi(\mathbf{b}, z)| \leq \frac{8\,C_f}{n}\|\tilde{\mathbf{w}} - \bar{\mathbf{w}}\|_3^3 \,. \tag{6.39}$$

Case 2 if $\|\tilde{\mathbf{w}} - \bar{\mathbf{w}}\|_3 \geq \frac{\epsilon}{9\,C_f}$.

$$\begin{aligned}
\Phi(\mathbf{b}, z) - \Psi(\mathbf{b}, z) &= \left(\phi(\mathbf{b}, z, \mathbf{w}_\phi) - \phi(\mathbf{b}, z, \bar{\mathbf{w}})\right) + \left(\phi(\mathbf{b}, z, \bar{\mathbf{w}}) - \phi(\mathbf{0}, 0, \bar{\mathbf{w}})\right) \\
&\quad + \left(\psi(\mathbf{0}, 0, \bar{\mathbf{w}}) - \psi(\mathbf{b}, z, \bar{\mathbf{w}})\right) + \left(\psi(\mathbf{b}, z, \bar{\mathbf{w}}) - \psi(\mathbf{b}, z, \tilde{\mathbf{w}})\right) \\
&\leq \psi(\mathbf{b}, z, \bar{\mathbf{w}}) - \psi(\mathbf{b}, z, \tilde{\mathbf{w}}) \leq \frac{1}{2n}\left(z - \mathbf{b}^\mathsf{T}\bar{\mathbf{w}}\right)^2 \,.
\end{aligned} \tag{6.40}$$

$$\Psi(\mathbf{b}, z) - \Phi(\mathbf{b}, z) \leq (\psi(\mathbf{b}, z, \tilde{\mathbf{w}}) - \psi(\mathbf{b}, z, \bar{\mathbf{w}})) + (\psi(\mathbf{b}, z, \bar{\mathbf{w}}) - \psi(\mathbf{0}, 0, \bar{\mathbf{w}})) + (\phi(\mathbf{0}, 0, \bar{\mathbf{w}}) - \phi(\mathbf{0}, 0, \bar{\mathbf{w}})) \leq \frac{1}{2n}\left(z - \mathbf{b}^\mathsf{T}\bar{\mathbf{w}}\right)^2 . \tag{6.41}$$

So finally,

$$|\Psi(\mathbf{b}, z) - \Phi(\mathbf{b}, z)| \leq \frac{1}{2n}\left(z - \mathbf{b}^\mathsf{T}\bar{\mathbf{w}}\right)^2 . \tag{6.42}$$

So by combining the two cases, we get

$$|\Phi(\mathbf{b}, z) - \Psi(\mathbf{b}, z)| \leq \mathbb{1}_{\|\tilde{\mathbf{w}}-\bar{\mathbf{w}}\|_3 \leq \frac{\epsilon}{9\,C_f}} \left(\frac{8\,C_f}{n}\|\tilde{\mathbf{w}} - \bar{\mathbf{w}}\|_3^3\right) + \mathbb{1}_{\|\tilde{\mathbf{w}}-\bar{\mathbf{w}}\|_3 > \frac{\epsilon}{9\,C_f}} \left(\frac{1}{2n}\left(z - \mathbf{b}^\mathsf{T}\bar{\mathbf{w}}\right)^2\right) . \tag{6.43}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left[|\Phi(\mathbf{b}, z) - \Psi(\mathbf{b}, z)|\right] \\
&\leq \mathbb{E}\left[\mathbb{1}_{\|\tilde{\mathbf{w}}-\bar{\mathbf{w}}\|_3 \leq \frac{\epsilon}{9\,C_f}} \left(\frac{8\,C_f}{n}\|\tilde{\mathbf{w}} - \bar{\mathbf{w}}\|_3^3\right)\right] + \mathbb{E}\left[\mathbb{1}_{\|\tilde{\mathbf{w}}-\bar{\mathbf{w}}\|_3 > \frac{\epsilon}{9\,C_f}} \left(\frac{1}{2n}\left(z - \mathbf{b}^\mathsf{T}\bar{\mathbf{w}}\right)^2\right)\right] \\
&\leq \frac{8C_f}{n}\mathbb{E}\left[\|\tilde{\mathbf{w}} - \bar{\mathbf{w}}\|_3^3\right] + \frac{1}{2n}\sqrt{Pr\left\{\|\tilde{\mathbf{w}} - \bar{\mathbf{w}}\|_3 \geq \frac{\epsilon}{9\,C_f}\right\}\,\mathbb{E}[\left(z - \mathbf{b}^\mathsf{T}\bar{\mathbf{w}}\right)^4]} \\
&\leq \frac{8C_f}{n}\mathbb{E}\left[\|\tilde{\mathbf{w}} - \bar{\mathbf{w}}\|_3^3\right] + \frac{1}{2n}\sqrt{\frac{\mathbb{E}\left[\|\tilde{\mathbf{w}} - \bar{\mathbf{w}}\|_3^3\right]}{(\frac{\epsilon}{9C_f})^3}\,\mathbb{E}[\left(z - \mathbf{b}^\mathsf{T}\bar{\mathbf{w}}\right)^4]} \\
&\leq \frac{C}{n^{5/4}}
\end{aligned} \tag{6.44}$$

On the other hand, it is easy to see that

$$\Psi(\mathbf{b}, z) - \Psi(\mathbf{0}, 0) = \frac{(z - \mathbf{b}^\mathsf{T}\bar{\mathbf{w}})^2}{2n(1 + \mathbf{b}^\mathsf{T}\boldsymbol{\Omega}^{-1}\mathbf{b})} , \tag{6.45}$$

where $\boldsymbol{\Omega} = \mathbf{V} + \mathbf{M}^\mathsf{T}\mathbf{M}$*. Note that*

$$\begin{aligned}
&\left|\mathbb{E}\left[\Psi(\mathbf{b}, z) - \Psi(\mathbf{0}, 0) - \frac{\sigma^2 + \frac{\|\bar{\mathbf{w}}\|^2}{n}}{2n(1 + \mathbb{E}[\mathbf{b}^\mathsf{T}\boldsymbol{\Omega}\mathbf{b}])}\right]\right| \\
&= \left|\mathbb{E}\left[\frac{(z - \mathbf{b}^\mathsf{T}\bar{\mathbf{w}})^2}{2n(1 + \mathbf{b}^\mathsf{T}\boldsymbol{\Omega}^{-1}\mathbf{b})} - \frac{\sigma^2 + \frac{\|\bar{\mathbf{w}}\|^2}{n}}{2n(1 + \mathbb{E}[\mathbf{b}^\mathsf{T}\boldsymbol{\Omega}\mathbf{b}])}\right]\right| \\
&\leq \frac{1}{2n}\mathbb{E}\left[(z - \mathbf{b}^\mathsf{T}\bar{\mathbf{w}})^2\ \left|\mathbf{b}^\mathsf{T}\boldsymbol{\Omega}\mathbf{b} - \mathbb{E}[\mathbf{b}^\mathsf{T}\boldsymbol{\Omega}\mathbf{b}]\right|\right] \\
&\leq \frac{1}{2n}\sqrt{\mathbb{E}\left[(z - \mathbf{b}^\mathsf{T}\bar{\mathbf{w}})^4\right]\ \mathbb{E}\left[\left(\mathbf{b}^\mathsf{T}\boldsymbol{\Omega}\mathbf{b} - \mathbb{E}[\mathbf{b}^\mathsf{T}\boldsymbol{\Omega}\mathbf{b}]\right)^2\right]} \\
&\leq \frac{C}{n^{1+\tau/2}} .
\end{aligned} \tag{6.46}$$

Now putting (6.44) *and* (6.46) *in* (6.33)*, results in*

$$\left|\mathbb{E}_{\{z,\mathbf{b}\}}\left[\Phi(\mathbf{b},z)-\bar{\Phi}-\frac{\sigma^2+\frac{\|\bar{\mathbf{w}}\|^2}{n}}{2n(1+\mathbb{E}[\mathbf{b}^\mathsf{T}\mathbf{\Omega}\mathbf{b}])}\right]\right|$$

$$\leq \frac{8C_f}{n}\mathbb{E}\left[\|\tilde{\mathbf{w}}-\bar{\mathbf{w}}\|_3^3\right]+\frac{27C_f^{3/2}}{2n\epsilon^{3/2}}\sqrt{\mathbb{E}\left[\|\tilde{\mathbf{w}}-\bar{\mathbf{w}}\|_3^3\right]\ \mathbb{E}[\left(z-\mathbf{b}^\mathsf{T}\bar{\mathbf{w}}\right)^4]}$$

$$+\frac{1}{2n}\sqrt{\mathbb{E}\left[(z-\mathbf{b}^\mathsf{T}\bar{\mathbf{w}})^4\right]\ \mathbb{E}\left[\left(\mathbf{b}^\mathsf{T}\mathbf{\Omega}\mathbf{b}-\mathbb{E}[\mathbf{b}^\mathsf{T}\mathbf{\Omega}\mathbf{b}]\right)^2\right]} \tag{6.47}$$

$$c\ n^{-(1+\tau/2)} \tag{6.48}$$

It remains to bound $\mathbb{E}\left[(z-\mathbf{b}^\mathsf{T}\bar{\mathbf{w}})^4\right]$ *and* $\mathbb{E}\left[\|\tilde{\mathbf{w}}-\bar{\mathbf{w}}\|_3^3\right]$*. For the first one, let* $\frac{1}{p}\mathbf{e}=\mathbb{E}[\mathbf{b}]$ *and* $\tilde{\mathbf{b}}=\mathbf{b}-\frac{1}{p}\mathbf{e}$*. Then,*

$$\begin{aligned}\mathbb{E}\left[(z-\mathbf{b}^\mathsf{T}\bar{\mathbf{w}})^4\right]&=\mathbb{E}[z^4]+6\mathbb{E}[z^2]\ \mathbb{E}[(\mathbf{b}^\mathsf{T}\bar{\mathbf{w}})^2]+\mathbb{E}[(\mathbf{b}^\mathsf{T}\bar{\mathbf{w}})^4]\\&=\mathbb{E}[z^4]+\frac{6\mathbb{E}[z^2]}{p}(\mathbb{E}[(\tilde{\mathbf{b}}^\mathsf{T}\bar{\mathbf{w}})^2]+(\mathbf{e}^\mathsf{T}\bar{\mathbf{w}})^2)+\mathbb{E}[(\tilde{\mathbf{b}}^\mathsf{T}\bar{\mathbf{w}})^4]\\&\quad+6\mathbb{E}[(\tilde{\mathbf{b}}^\mathsf{T}\bar{\mathbf{w}})^2](\mathbf{e}^\mathsf{T}\bar{\mathbf{w}})^2+(\mathbf{e}^\mathsf{T}\bar{\mathbf{w}})^4\\&\leq C_1+C_2\|\bar{\mathbf{w}}\|^2+C_3\|\bar{\mathbf{w}}\|^4\ .\end{aligned} \tag{6.49}$$

On the other hand, let $\mathbf{\Omega}^{-1}=[\omega_1\ldots,\omega_p]^\mathsf{T}$*. Since* $\mathbf{\Omega}^{-1}\leq 1/\epsilon$*,*

$$\begin{aligned}\mathbb{E}\left[\|\tilde{\mathbf{w}}-\bar{\mathbf{w}}\|_3^3\right]&=\mathbb{E}\left[\left\|\frac{(z-\mathbf{b}^\mathsf{T}\bar{\mathbf{w}})}{(1+\mathbf{b}^\mathsf{T}\mathbf{\Omega}^{-1}\mathbf{b})}\ \mathbf{\Omega}^{-1}\mathbf{b}\right\|_3^3\right]\leq\mathbb{E}\left[\left\|(z-\mathbf{b}^\mathsf{T}\bar{\mathbf{w}})\ \mathbf{\Omega}^{-1}\mathbf{b}\right\|_3^3\right]\\&\leq 4\,\mathbb{E}\left[\left\|(z-\mathbf{b}^\mathsf{T}\bar{\mathbf{w}})\ \mathbf{\Omega}^{-1}\tilde{\mathbf{b}}\right\|_3^3\right]+\frac{4}{p^3}\mathbb{E}\left[\left\|(z-\mathbf{b}^\mathsf{T}\bar{\mathbf{w}})\ \mathbf{\Omega}^{-1}\mathbf{e}\right\|_3^3\right]\\&\leq 4\sqrt{\mathbb{E}\left[(z-\mathbf{b}^\mathsf{T}\bar{\mathbf{w}})^6\right]\ \mathbb{E}\left[\left\|\mathbf{\Omega}^{-1}\tilde{\mathbf{b}}\right\|_3^6\right]}+\frac{4}{p^3}\left\|\mathbf{\Omega}^{-1}\mathbf{e}\right\|_3^3\sqrt{\mathbb{E}\left[(z-\mathbf{b}^\mathsf{T}\bar{\mathbf{w}})^6\right]}\\&\leq 4\sqrt{\mathbb{E}\left[(z-\mathbf{b}^\mathsf{T}\bar{\mathbf{w}})^6\right]\ \mathbb{E}\left[\sum_k|\omega_k^\mathsf{T}\tilde{\mathbf{b}}|^3\right]^2}+\frac{4}{p^3}\left\|\mathbf{\Omega}^{-1}\mathbf{e}\right\|_2^3\sqrt{\mathbb{E}\left[(z-\mathbf{b}^\mathsf{T}\bar{\mathbf{w}})^6\right]}\\&\leq(\frac{C}{p\epsilon^3}+\frac{4\|\mathbf{e}\|_2^3}{\epsilon^2p^3})\sqrt{\mathbb{E}\left[(z-\mathbf{b}^\mathsf{T}\bar{\mathbf{w}})^6\right]}\end{aligned} \tag{6.50}$$

which concludes the proof.

Theorem 17 *let* $\mathbf{W_A}$ *and* $\mathbf{W_B}$ *bt the optimal solutions to* (6.15)*. If for any function* $f(.)$*, that satisfies our conditions,*

$$\Phi_\mathbf{A}-\Phi_\mathbf{B}\rightarrow 0\ , \tag{6.51}$$

then,

$$\frac{1}{p^2}\|\mathbf{W_A}\|_F^2 - \frac{1}{p^2}\|\mathbf{W_B}\|_F^2 \to 0\ . \tag{6.52}$$

Proof 9 *Assume that* $\frac{1}{p^2}\|\mathbf{W_A}\|_F^2$ *and* $\frac{1}{p^2}\|\mathbf{W_B}\|_F^2$ *converge to difference values of* $C_\mathbf{A}$ *and* $C_\mathbf{B}$*. Choose* $C = (C_\mathbf{B} + C_\mathbf{A})/2$ *and consider the following optimization,*

$$\begin{aligned}
\bar{\Phi}_\mathbf{A} &= \min_{\substack{\frac{1}{p^2}\|\mathbf{W}\|_F^2 \le C \\ \mathbf{W}\in\mathbf{H}^p}} \frac{1}{2n}\sum_{i=1}^{n}\left(z_i - Tr(\mathbf{A}_i\cdot\mathbf{W})\right)^2 + f\left(\mathbf{W}\right)\ , \\
\bar{\Phi}_\mathbf{B} &= \min_{\substack{\frac{1}{p^2}\|\mathbf{W}\|_F^2 \le C \\ \mathbf{W}\in\mathbf{H}^p}} \frac{1}{2n}\sum_{i=1}^{n}\left(z_i - Tr(\mathbf{B}_i\cdot\mathbf{W})\right)^2 + f\left(\mathbf{W}\right)\ .
\end{aligned} \tag{6.53}$$

We show that the two should converge to the same value, which is a contradiction since $f(.)$ *is strongly convex and one should converge to* $\Phi_\mathbf{A}$ *and the other should be larger that* $\Phi_\mathbf{B}$*. Using min-max theorem, they can be rewritten as*

$$\begin{aligned}
\bar{\Phi}_\mathbf{A} &= \sup_{\lambda>0}\ -\lambda\, C + \min_{\mathbf{W}\in\mathbf{H}^p}\ \frac{1}{2n}\sum_{i=1}^{n}\left(z_i - Tr(\mathbf{A}_i\cdot\mathbf{W})\right)^2 + f\left(\mathbf{W}\right) + \frac{\lambda}{p^2}\|\mathbf{W}\|_F^2\ , \\
\bar{\Phi}_\mathbf{B} &= \sup_{\lambda>0}\ -\lambda\, C + \min_{\mathbf{W}\in\mathbf{H}^p}\ \frac{1}{2n}\sum_{i=1}^{n}\left(z_i - Tr(\mathbf{B}_i\cdot\mathbf{W})\right)^2 + f\left(\mathbf{W}\right) + \frac{\lambda}{p^2}\|\mathbf{W}\|_F^2\ .
\end{aligned} \tag{6.54}$$

Due to the assumption of the theorem, the two inside converge to the same value for any fixed λ*. So the concave version of Lemma 18 shows that* $\bar{\Phi}_\mathbf{A}$ *and* $\bar{\Phi}_\mathbf{B}$ *also converge to the same value which is a contradiction.*

Lemma 18 *Consider a series of convex functions* $f_p : \mathbb{R}^{>0} \to \mathbb{R}$ *that converges point-wise to the function* $f : \mathbb{R}^{>0} \to \mathbb{R}$*. Besides, there exists* $M > 0$ *such that for any* $x > M$*, we have* $f(x) > \inf_{s>0}\ f(s)$*. Then* $f(.)$ *is also convex and* $\inf_{s>0}\ f_p(s) \xrightarrow{p} \inf_{s>0}\ f(s)$*.*

Lemma 19 *Let* $\bar{\mathbf{w}}$ *be the optimal solution to the optimization*

$$\min_{\mathbf{w}}\ \frac{1}{2}\|\mathbf{z} - \mathbf{A}\mathbf{w}\|^2 + f(\mathbf{w} + \mathbf{x}_0)\ , \tag{6.55}$$

where $f(.)$ *is strongly convex with constant* ϵ*. Then*

$$\|\bar{\mathbf{w}}\| \le \frac{2}{\epsilon}(\|\mathbf{A}^\top\mathbf{z}\| + \|\nabla f(\mathbf{x}_0)\|) \tag{6.56}$$

Proof 10 *let*

$$\phi(\mathbf{A}, \mathbf{w}) = \frac{1}{2}\|\mathbf{z} - \mathbf{A}\mathbf{w}\|^2 + f(\mathbf{w} + \mathbf{x}_0) \ . \tag{6.57}$$

We have

$$0 > \phi(\mathbf{A}, \bar{\mathbf{w}}) - \phi(\mathbf{A}, 0) \geq \bar{\mathbf{w}}^\mathsf{T} \left(-\mathbf{A}^\mathsf{T}\mathbf{z} + \nabla f(\mathbf{x}_0)\right) + \frac{\epsilon}{2}\|\bar{\mathbf{w}}\|^2 \ .$$

Therefore,

$$\frac{\epsilon}{2}\|\bar{\mathbf{w}}\|^2 \leq \left|\bar{\mathbf{w}}^\mathsf{T} \left(-\mathbf{A}^\mathsf{T}\mathbf{z} + \nabla f(\mathbf{x}_0)\right)\right| \leq \|\bar{\mathbf{w}}\| \ \left(\|\mathbf{A}^\mathsf{T}\mathbf{z}\| + \|\nabla f(\mathbf{x}_0)\|\right) \ , \tag{6.58}$$

which concludes the proof. Now let $\bar{\mathbf{w}}$ *be the optimizer of* $\phi(\mathbf{A}, \mathbf{w})$ *and* $\mathbb{E}[\mathbf{A}] = \mathbf{1}\mathbf{e}^\mathsf{T}$. *Due to optimality we have,*

$$0 = \mathbf{A}^\mathsf{T}(\mathbf{A}\bar{\mathbf{w}} - \mathbf{z}) + \nabla f(\mathbf{x}_0 + \bar{\mathbf{w}}) \tag{6.59}$$

Lemma 20 *Let* $\bar{\mathbf{w}}$ *be the optimal solution to the optimization*

$$\min_{\mathbf{w}} \ \frac{1}{2}\|\mathbf{z} - \mathbf{A}\mathbf{w}\|^2 + f(\mathbf{w} + \mathbf{x}_0) \ , \tag{6.60}$$

where $f(.)$ *is strongly convex with constant* ϵ *and* $\mathbf{A} \in \mathbb{R}^{n \times p}$ *is a random value with* $\mathbb{E}[\mathbf{A}] = \mathbf{1}\mathbf{e}^t$ *and* $\mathbf{B} = \mathbf{A} - \mathbf{1}\mathbf{e}^t$. *Then*

$$\|\bar{\mathbf{w}}\| \leq \frac{2}{\epsilon}(\|\mathbf{A}^\mathsf{T}\mathbf{z}\| + \|\nabla f(\mathbf{x}_0)\|) \tag{6.61}$$

Proof 11 *We have,*

$$\begin{aligned} \phi(\mathbf{A}, 0) \geq \phi(\mathbf{A}, \bar{\mathbf{w}}) &= \frac{1}{2}\|\mathbf{z} - \mathbf{B}\bar{\mathbf{w}} - \mathbf{1}\mathbf{e}^\mathsf{T}\bar{\mathbf{w}}\|^2 + f(\bar{\mathbf{w}} + \mathbf{x}_0) \\ &\geq \frac{n}{2}(\mathbf{e}^\mathsf{T}\bar{\mathbf{w}})^2 + (\mathbf{e}^\mathsf{T}\bar{\mathbf{w}}) \cdot \mathbf{1}^\mathsf{T}(\mathbf{B}\bar{\mathbf{w}} - \mathbf{z}) \end{aligned} \tag{6.62}$$

Therefore,

$$(\mathbf{e}^\mathsf{T}\bar{\mathbf{w}})^2 + \frac{2}{n}(\mathbf{e}^\mathsf{T}\bar{\mathbf{w}}) \cdot \mathbf{1}^\mathsf{T}(\mathbf{B}\bar{\mathbf{w}} - \mathbf{z}) - \frac{2}{p}\phi(\mathbf{A}, 0) \leq 0 \tag{6.63}$$

This results in

$$\begin{aligned} |\mathbf{e}^\mathsf{T}\bar{\mathbf{w}}| &\leq \frac{2}{n}\left|\mathbf{1}^\mathsf{T}(\mathbf{B}\bar{\mathbf{w}} - \mathbf{z})\right| + \frac{2}{n}\phi(\mathbf{A}, 0) \leq \frac{2}{n}|\mathbf{1}^\mathsf{T}\mathbf{z}| + \frac{2}{n}\|\bar{\mathbf{w}}\| \cdot \|\mathbf{B}^\mathsf{T}\mathbf{1}\| + \frac{2}{n}\|\mathbf{z}\|^2 + f(\mathbf{x}_0) \\ &\leq \frac{2}{n}|\mathbf{1}^\mathsf{T}\mathbf{z}| + \frac{4}{n\epsilon}\left(\|\mathbf{A}^\mathsf{T}\mathbf{z}\| + \|\nabla f(\mathbf{x}_0)\|\right) \cdot \|\mathbf{B}^\mathsf{T}\mathbf{1}\| + \frac{2}{n}\|\mathbf{z}\|^2 + f(\mathbf{x}_0) \end{aligned} \tag{6.64}$$

Some notes on the Corollary 9

Out goal is to analyze the following optimization and see under what conditions it will succeed.

$$\min_{\frac{1}{p}\mathrm{Tr}\left((\mathbf{A}_i\sqrt{2}+\mathbb{I})\mathbf{X}_0\right)=\frac{1}{p}\mathrm{Tr}\left((\mathbf{A}_i\sqrt{2}+\mathbb{I})\mathbf{X}\right),\ \forall i} \mathrm{Tr}(\mathbf{X}-\mathbf{X}_0)\ , \quad \text{s.t.} \quad \mathbf{X}\succeq 0\ . \tag{6.65}$$

Here, we initially assume that $\mathbf{X}_0 = \mathbf{x}_0\mathbf{x}_0^\mathsf{T}$ is a rank-one matrix where $\frac{1}{p}\|\mathbf{x}_0\|^2 = 1$. This optimization succeeds in recovering $\mathbf{X}_0$, if and only if the optimal value of objective function is zero (and it fails if the optimal value is negative).
Now, assuming an iid standard Gaussian distribution for the entries of the matrices $\mathbf{A}_i$, we can use the proof of Theorem 1 and CGMT framework, and get the following optimization.

$$\begin{aligned}\min_{\substack{\alpha\geq 0\\ t\in\mathbb{R}}}\ \max_{\substack{\beta,\tau\geq 0\\ \gamma\in\mathbb{R}}}\ &\sqrt{\beta}\sqrt{\delta}\sqrt{2\alpha^2+t^2}-\frac{\alpha}{2\tau}+\gamma(t+1)+\frac{1}{2\alpha\tau}-1\\ &-\frac{1}{2\alpha\tau}\left\|\left(\frac{1}{p}\mathbf{x}_0\mathbf{x}_0^\mathsf{T}+\frac{\alpha\tau\beta}{\sqrt{p}}\mathbf{H}+\alpha\tau(\gamma-1)\mathbb{I}\right)_+\right\|_F^2\ ,\end{aligned} \tag{6.66}$$

where $\mathbf{H}$ is a Wigner random matrix, and $(\cdot)_+$ is the positive definite part of a matrix. We will have perfect recovery if and only if this optimization is non-negative. First step of understanding this optimization, would be to analyze the high dimensional behavior of the positive part of the matrix

$$\frac{1}{p}\mathbf{x}_0\mathbf{x}_0^\mathsf{T}+\frac{\alpha\tau\beta}{\sqrt{p}}\mathbf{H}+\alpha\tau(\gamma-1)\mathbb{I}\ . \tag{6.67}$$

We know that the eigenvalue distribution of matrix $\mathbf{H}$ follows the semi-circular law and has eigenvalues from -2 to 2. Besides, adding a rank-one matrix $\mathbf{x}_0\mathbf{x}_0^\mathsf{T}/p$, affects this eigen-value distribution only if the coefficient of this rank one distortion is more than the coefficient of $\mathbf{H}/\sqrt{p}$. Meaning that if $\alpha\beta\tau < 1$ we will have an extra eigenvalue at $1+\alpha^2\tau^2\beta^2$. Otherwise, the eigenvalues of $\frac{1}{p}\mathbf{x}_0\mathbf{x}_0^\mathsf{T}+\frac{\alpha\tau\beta}{\sqrt{p}}\mathbf{H}$ will have the same distribution of the eigenvalues of $\frac{\alpha\tau\beta}{\sqrt{p}}\mathbf{H}$ which is semi-circular. Besides, in optimization (6.66), we should always have $\alpha\tau(\gamma-1+2\beta) < 0$. Otherwise, the positive part of the matrix in (6.67) will have an infinitely large Frobenius norm, which the minimization will avoid. Therefore, we sould always have $\gamma-1-2\beta < 0$. Considering this constraint, if we solve this optimization over γ and the optimization

becomes,

$$\min_{\substack{\alpha\geq 0\\ t\in\mathbb{R}}} \max_{\beta\geq 0} \begin{cases} \beta(\sqrt{\delta}\sqrt{2\alpha^2+t^2} - \sqrt{2(\alpha^2-t^2)(1+t)}) + t & \text{if } t^2 < \alpha^2 < t^2 + 2(t+1)(1-\sqrt{1+t})^2 \,, \\ \beta(\sqrt{\delta}\sqrt{2\alpha^2+t^2} - \alpha\phi_1(\alpha,t) - 2t) + t & \text{if } t^2 + 2(t+1)(1-\sqrt{1+t})^2 < \alpha^2 < 1 \,, \\ \beta(\sqrt{\delta}\sqrt{2\alpha^2+t^2} - 2t - 2) + t & \text{if } \alpha^2 > 1 \,, \\ \infty & \text{if } \alpha^2 \leq t^2 \,. \end{cases} \tag{6.68}$$

where the function $\phi(\cdot,\cdot)$ is defined as

$$\phi_1(\alpha,t) = \min_{\frac{1-\sqrt{1+t}}{\alpha}\leq\tau\leq\frac{1}{\alpha}} \frac{1}{2\tau} + 3\tau - 2\alpha\tau^2 + \frac{\alpha^2\tau^3}{2} \,. \tag{6.69}$$

So we would like to see under what conditions, the optimization (6.68) will be negative, which means the initial optimization (6.65) will fail in recovering $\mathbf{X}_0$. The optimization (6.68) will be negative, if there exists a negative $t < 0$ that makes the coefficient of β negative.

It's not hard to see that the minimization over α and t in (6.68) happens when both α and t go to zero. Thus, it's the ratio of α/t will define the result of this optimization. A closer look at this optimization shows that it will be minimized when $\alpha^2 > \frac{3}{2}t^2$, especially, when $t/\alpha- \to 0$. In this scenario, the second case in the objective function of (6.68) holds, and the objective function will converge to $\beta\alpha(\sqrt{2\delta} = \sqrt{6})$ which is always positive if $\delta > 3$. In the other words, this optimization is zero if $\delta \geq 3$, otherwise, there exists negative values of t that makes it negative. Thus we have perfect recovery iff $\delta > 3$.

This proof was for the case that rank of the unknown matrix $\mathbf{X}_0$ was one. For the case of a rank r matrix, the proof will be the same, expect that $\mathbf{X}_0 = \sum_{i=1}^{r} \mathbf{x}_i\mathbf{x}_i^\mathsf{T}$. So in the matrix (6.67), we will have a rank r distortion of a Wigner matrix, and following the same steps, we can show that if $\delta > 3r$, the objective function always remains non-negative.

Chapter 7

PERFORMANCE ANALYSIS OF CONVEX DATA DETECTION IN MIMO

7.1 Introduction

We consider the problem of recovering a transmit signal[1], $\beta_0 \in \mathcal{D}^p$, form n (noisy) linear observations of the form $\mathbf{y} = \mathbf{X}\beta_0 + \mathbf{z}$, where $\mathcal{D} \subset \mathbb{C}$ denotes the discrete transmit constellation, and $\mathbf{z} \in \mathbb{C}^n$ is the noise vector. This problem has a pivotal role in signal detection in multiple-input, multiple-output (MIMO) communication systems [90, 91, 126], where $\mathbf{X} \in \mathbb{C}^{n\times p}$, often referred to as the channel state information, is a known matrix. In such settings, n and p correspond to the number of transmit and receive antennas, respectively.

The Maximum Likelihood(ML) estimator is the desirable theoretical solution for this problem. There has been numerous studies to investigate algorithms that can generate exact or approximate solutions for this problem. Due to the combinatorial nature [185] of the problem, exact algorithms (e.g. sphere decoding [83]) are computationally prohibitive, especially in a very large system (e.g. massive MIMO) [140]. Therefore, various heuristics have been proposed and used in practice [71, 81] to approximate the ML solution. Despite tractable computational complexity, the precise performance analysis of such methods are often challenging.

Due to the practical advantages of convex algorithms, one conventional approach to solve this problem is to relax the discrete set Dc to a continuous convex set $\mathbf{S}$ and utilize convex programming to search over $\mathbf{S}$ instead of $\mathcal{D}$ [108, 162, 198]. The performance of this method for data recovery has been investigated in the works of [9, 91, 166] for the real valued constellations, specifically, BPSK and PAM when the channel matrix is Gaussian. To do so, Thrampoulidis et. al. [166] utilized a framework that they had developed, known as the CGMT framework [4, 165]. The CGMT framework has been successfully applied to analyze the performance in a number of other applications including analysis of regularized M- estimators [165], and PhaseMax in phase retrieval [53, 141, 142]. Unfortunately, The CGMT framework can not be readily extended to the complex settings (which indeed is the desirable case in many practical applications).

[1] This chapter is mainly based on the work in [1]

The major result of this section is to introduce a new comparison lemma for complex Gaussian processes to study the convex detection problem for complex constellations. In particular, we *precisely* characterize the symbol error rate performance of the convex method, for a general constellation $\mathcal{D}$ and a convex relaxation $\mathbf{S}$. Our theorem also allows us to derive the necessary and sufficient number of antennas, n, required for data recovery in the high-SNR regime which enables us to precisely characterize the phase-transition regions. Through our analysis, we can further observe the relationship between the choice of the convex relaxation with its corresponding phase transition. As an example, we analyze the loss in performance when choosing a relaxation that is easier to implement in a convex program for the case of PSK modulation.

7.2 Problem Setup

Notations We gather here the basic notations that are used throughout this section. We reserve the letter j for the complex unit. For a complex scalar $x \in \mathbb{C}$, x_{Re} and x_{Im} correspond to the real are imaginary parts of x, respectively and $|x| = \sqrt{x_{\mathrm{Re}}^2 + x_{\mathrm{Im}}^2}$. $\mathcal{N}(\mu, \sigma^2)$ denotes real Gaussian distribution with mean μ and variance σ^2. Similarly, $\mathcal{N}_{\mathbb{C}}(\mu, \sigma^2)$ refers to a *complex* Gaussian distribution with real and imaginary parts drawn independently from $\mathcal{N}(\mu_{\mathrm{Re}}, \sigma^2/2)$ and $\mathcal{N}(\mu_{\mathrm{Im}}, \sigma^2/2)$, respectively. $X \sim p_X$ implies that the random variable X has a density p_X. We reserve the letters G and H to denote (scalar) standard normal random variables. Similarly, $H_{\mathbb{C}}$ is reserved to denote a *complex* $\mathcal{N}_{\mathbb{C}}(0, 2)$ random variable. The bold lower letters are reserved for vectors and for a vector $\mathbf{v}$, $\mathbf{v}_i$ denotes its i^{th} entry. Finally, for a *convex* set $\mathbf{S} \subset \mathbb{C}$, the projection and distance functions with respect to $\mathcal{D}$ are defined as

$$\begin{aligned} \mathcal{P}_{\mathbf{S}}(\mathbf{x}) &:= \arg\min_{\mathbf{y}\in\mathbf{S}} \|\mathbf{x} - \mathbf{y}\| \\ \mathcal{P}_{\mathbf{S}}(\mathbf{x}) &:= \min_{\mathbf{y}\in\mathbf{S}} \|\mathbf{x} - \mathbf{y}\|. \end{aligned} \tag{7.1}$$

Setup Our goal is to recover an p-dimensional vector $\beta_0 \in \mathbb{C}^p$ where the entries of β_0 are independenty drawn from the discrete set $\mathcal{D} \subset \mathbb{C}$ with distribution $\mathbf{x}_{0,i} \sim p_X$. The set $\mathcal{D}$ defines the modulation used for data transmission (e.g. QAM, PSK, etc.). For this purpose, we are given the noisy multiple-input multiple-output (MIMO) relation of the form

$$\mathbf{y} = \mathbf{X}\beta_0 + \mathbf{z} \in \mathbb{C}^n, \tag{7.2}$$

where $\mathbf{X} \in \mathbb{C}^{n\times p}$ is the known MIMO channel matrix with i.i.d. entries drawn from $\mathcal{N}_{\mathbb{C}}(0, \frac{1}{p})$ and $\mathbf{z} \in \mathbb{C}^n$ is the unknown noise vector with i.i.d. random complex Gaussian $\mathcal{N}_{\mathbb{C}}(0, \sigma^2)$ entries.

Estimator The ML estimator of $\mathbf{x}_0$ in this scenario is

$$\hat{\beta} = \arg\min_{\beta\in\mathcal{D}^p} \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\beta\|^2. \tag{7.3}$$

Since solving (7.3) is computationally intractable, a variety of heuristic methods, such as zero-forcing, MMSE, decision-feedback, have been proposed. In this section, we make use of convex programming to estimate β_0. In the first step, we relax $\mathcal{D}$ to a convex set $\mathbf{S}$ and minimize the objective function of (7.3) over this relaxed convex set,

$$\tilde{\beta} = \arg\min_{\beta\in\mathbf{S}^p} \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\beta\|^2. \tag{7.4}$$

Next, we map each entry of $\tilde{\beta}$ to the closest point in $\mathcal{D}$ to build our final estimation of β_0,

$$\hat{\beta}_i = \arg\min_{\beta\in\mathcal{D}} |\beta - \tilde{\beta}_i|, \quad i = 1, \ldots, p. \tag{7.5}$$

We refer to this method as the Convex Decoder Algorithm (CDA). In this section, we will precisely analyze the performance of the CDA as a function of the problem parameters such as σ, n/p, $\mathcal{D}$ and $\mathbf{S}$. Note that the performance of CDA depends on the constellation $\mathcal{D}$ and the way we relax it to the convex set $\mathbf{S}$. Later in Section 7.3, we observe the impact of choosing two different relaxations on the performance of CDA and its phase-transition regions with the help of our main theorem.

Symbol error probability We characterize the performance of CDA in terms of the symbol error probability, defined as the expected value of the Symbol Error Rate (SER) where,

$$\begin{aligned} SER &:= \frac{1}{p}\sum_{i=1}^{p} \mathbb{1}_{\hat{\beta}_i \neq \beta_{0,i}}, \\ P_e &:= \mathbb{E}[SER] = \frac{1}{p}\sum_{i=1}^{p} \mathbb{P}\left(\hat{\beta}_i \neq (\beta_0)_i\right). \end{aligned} \tag{7.6}$$

Here $\hat{\beta}$ is the output of CDA in (7.5), $\mathbb{1}_{\mathcal{E}}$ is the indicator of the event $\mathcal{E}$ and the probability $\mathbb{P}(\cdot)$ is over the randomness of $\mathbf{X}$, $\mathbf{z}$ and β_0. We introduce the notation $\mathbf{S}_x$ for $x \in \mathcal{D}$, as the set of all points in $\mathbf{S}$ that will be mapped to β in (7.5). Equivalently,

$$\mathbf{S}_\beta := \{\beta' \in \mathcal{S} : \forall \mathbf{y} \in \mathcal{D},\ |beta' - beta| < |beta' - y|\}. \tag{7.7}$$

This notation helps us interpret our main theorem more clearly. Using this notation, we can rewrite the symbol error probability defined in (7.6) as

$$P_e = \frac{1}{p} \sum_{i=1}^{p} \mathbb{P}\left(\tilde{\beta}_i \notin \mathbf{S}_{(\beta_0)_i}\right), \tag{7.8}$$

where $\tilde{\beta}$ is the minimizer of (7.4).

Assumptions We impose two mild assumptions on the problem. First, we assume that the entries of β_0 are i.i.d. random variables with $\beta_{0,i} \sim p_\beta$ and also $\mathbb{P}_\beta(r_1 + jr_2) = \mathbb{P}_\beta(r_2 + jr_1),\ \forall r_1, r_2 \in \mathbb{R}$. Second, we want the convex set to be *symmetric* in the sense that if $(r_1 + jr_2) \in \mathbf{S}$, then also $(r_2 + jr_1) \in \mathbf{S}$.

Modulations

Using our main theorem, we can precisely analyze the SER of the CDA in terms of the SER for a constellation $\mathcal{D}$ which we relax to an arbitrary convex set $\mathbf{S}$. For a better understanding of the theorem and to show how to apply it to different schemes, we will work with two conventional modulations; Phase-Shift Keying (PSK) and Quadrature Amplitude Modulation (QAM).

N-PSK Constellation: In the N-PSK constellation, each entry of β_0 is randomly drawn from $\mathcal{D} = \left\{e^{\frac{j2\pi}{N}i} : i = 0, \ldots, N-1\right\}$. The entries of $\mathcal{D}$ are distributed over the unit circle in the complex space and therefore the Signal to Noise Ratio (SNR) will be $1/\sigma^2$. Next, we need an appropriate convex relaxation of $\mathcal{D}$ for CDA. We suggest two candidates for this purpose and compare their performances later in Section 7.3. In one which we will refer to as the Circular Relaxation (CR), we choose the set $\mathbf{S}^{(\mathrm{CR})} = \{\beta \in \mathbb{C} : |\beta| \leq 1\}$ as the convex set in (7.4). The simple structure of $\mathbf{S}^{(\mathrm{CR})}$ makes its implementation easier in the convex program (7.4). In another scenario, we consider the convex hull of $\mathcal{D}$ as the relaxed set $\mathbf{S}$ and refer to it as Convex Hull Relaxation (CHR). Thus, $\mathbf{S}^{(\mathrm{CHR})} = \mathrm{Conv}(\mathcal{D})$ will be used in (7.4) which might be harder to implement compared to $\mathbf{S}^{(\mathrm{CR})}$. But we will show that since (CHR) is a tighter relaxation, its corresponding CDA performs better in

terms of SER.

N^2-QAM Constellation: We also briefly talk about the N^2-QAM modulation where

$$\mathcal{D} = \left\{ (r + j\ s) - \frac{N-1}{2}(1+j) : r, s \in \{0, \dots, N-1\} \right\}.$$

Under this constellation the SNR will be $\frac{N(N^2-1)}{6\sigma^2}$. The relaxation that is often used for this modulation is known as the Box Relaxation (BR) [166] which is

$$\mathbf{S} = \left\{ (x + jy) \in \mathbb{C} : |x| \leq \frac{N-1}{2}, |y| \leq \frac{N-1}{2} \right\} \tag{7.9}$$

Using our main theorem, we can calculate the SER of CDA under box relaxation and rederive the results of [91, 166].

7.3 Main Result

Our main result explicitly characterizes the limiting behavior of the symbol error rate of the convex decoder algorithm, under the high dimensional regime where $n, p \to \infty$ with a constant ratio $\delta := n/p$.

Theorem 18 *(SER analysis of CDA) Let SER denote the symbol error rate of the Convex Decoder Algorithm (CDA), for random signal $\beta_0 \in \mathcal{D}$ with entries drawn independently from the distribution p_X. Let $\mathbf{S}$ be a convex relaxation of $\mathcal{D}$ and $\mathbf{S}$ and p_X satisfy the assumptions in Section 7.2. Fix SNR and $\delta = n/p$ and consider the optimization*

$$\min_{\tau > 0} \frac{\delta - 1}{2\tau\delta} + \frac{\sigma^2 \tau}{4} + \frac{\tau}{4} \mathbb{E}[Dist^2_{\mathbf{S}}(X + \frac{H_{\mathbb{C}}}{\tau\sqrt{\delta}})]. \tag{7.10}$$

If (7.10) *has a unique answer τ^*, then in the limit of $p, n \to \infty$*

$$\lim_{p,n \to \infty} P_e = \mathbb{P}\left(\mathcal{P}_{\mathbf{S}}(X + \frac{H_{\mathbb{C}}}{\tau^*\sqrt{\delta}}) \notin \mathbf{S}_X \right). \tag{7.11}$$

The expected value and probability in (7.10) *and* (7.11) *are over $X \sim p_X$ and $H_{\mathbb{C}} \sim \mathcal{N}_{\mathbb{C}}(0, 2)$, respectively.*

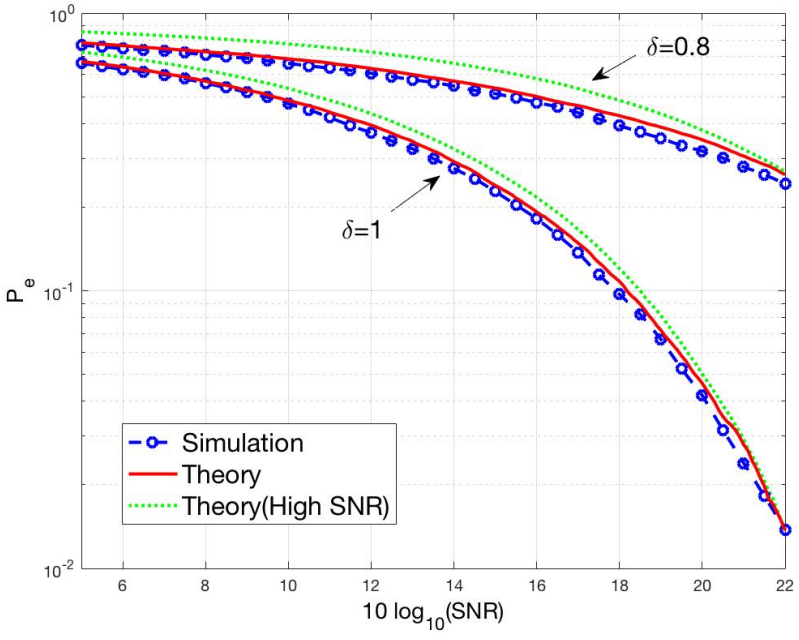


Figure 7.1: SER Performance of the Circular Relaxation (CR) for 16-PSK: P_e as a function of SNR for the two cases where $\delta = .8$ and $\delta = 1$. The theoretical prediction follows from Theorem 18 and the high-SNR analysis comes from Section 7.3. For the simulation, we used signals of size $p = 128$ with each entry chosen randomly uniform from the set $\mathbf{S}_{PSK} = \left\{ e^{\frac{j\pi}{8} i} : i = 0, \ldots, 15 \right\}$. The data are averages over 30 independent realizations of the channel matrix and the noise vector.

Theorem 18 provides a formula to calculate the SER of the convex decoder, under a general constellation in the high dimensional regime.

Theorem 18 provides a formula to calculate the SER of the convex decoder, under a general constellation in the high dimensional regime.

Remark 27 *(Computing τ^*) The objective function in* (7.10) *is convex and only involves one scalar variable. Thus, τ^* can, in principle, be efficiently numerically computed. It can be shown that τ^* is the minimizer of* (7.10) *if and only if it is the answer to the corresponding first-order optimality condition,*

$$\frac{1}{\tau^{*2}} = \frac{1}{2}\left(\sigma^2 + \mathbb{E}\left[\left|X - \mathcal{P}_{\mathbf{S}}(X + \frac{H_{\mathbb{C}}}{\tau^*\sqrt{\delta}})\right|^2\right]\right) . \tag{7.12}$$

Although, this does not provide us with a closed form formula to calculate τ^*, in all our simulations a fixed-point iterative method converges to τ^* over a handful of

iterations. It can be also shown that (7.12) *has a unique solution if* $\delta > \delta^*$ *for some* $\delta^* \in (0, 1)$ *which depends on* $\mathbf{S}$.

Next, we apply Theorem 18 to the N-PSK and N^2-QAM modulations introduced in Section 7.2 to calculate their corresponding symbol error probabilities and phase-transition thresholds in the high-SNR regime. Figures 1 and 2 verify the accuracy of the prediction of Theorem 18 for 16-PSK and 16-QAM modulations, respectively. Note that although the theorem requires $p \to \infty$, the prediction is already accurate for $p = 128$. In these figures, we have also plotted the high-SNR expressions for SER that we derive in the next Section for both modulations. Interestingly, we observe that this high-SNR expression gives us a good enough approximation of the exact value of SER, even for small practical values of SNR.

N-PSK Constellation

Under the N-PSK setup, the set $\mathcal{D}$ is defined in Section 7.2. We investigate the error performance of the convex decoder algorithm for two different convex relaxations for the set $\mathcal{D}$; Circular Relaxation (CR) and Convex-Hull Relaxation (CHR). The effect of using different relaxations shows up in the projection function in equations (7.12) and (7.11). Define $\mathbf{S}^{(\mathrm{SR})} = \{c \in \mathbb{C} : |c| \leq 1\}$ as the circular relaxation of $\mathcal{D}$. The projection function on this set has the following form,

$$\mathcal{P}_{\mathbf{S}^{(\mathrm{CR})}}(\beta) = \begin{cases} \beta & \text{if } |\beta| \leq 1 \\ \beta/|\beta| & \text{otherwise.} \end{cases} \tag{7.13}$$

Therefore, τ^* can be efficiently calculated using a fixed-point iterative method to solve (7.12). Furthermore, due to the symmetric nature of the N-PSK constellation, the probability of error for each symbol in $\mathcal{D}$ can be derived in the following closed form,

$$P_e = \mathbb{P}\left(|G| > \tan(\frac{\pi}{N})(H + \tau^*\sqrt{\delta})\right), \tag{7.14}$$

where G and H are i.i.d. $\mathcal{N}(0, 1)$.

High-SNR Analysis Let $\mathbf{S}^{(\mathrm{CR})}$ and $\mathbf{S}^{(\mathrm{CHR})}$ denote the circular relaxation and convex-hull relaxation of the set $\mathbf{S}$. It can be shown that for SNR $\gg 1$, τ^* grows large proportional to $\sqrt{\mathrm{SNR}}$. As a consequence, the last term in (7.10) can be approximated by $\frac{1}{8\tau\delta}$ and $\frac{N+4}{8N\tau\delta}$ for the cases of (CR) and (CHR), respectively. This results

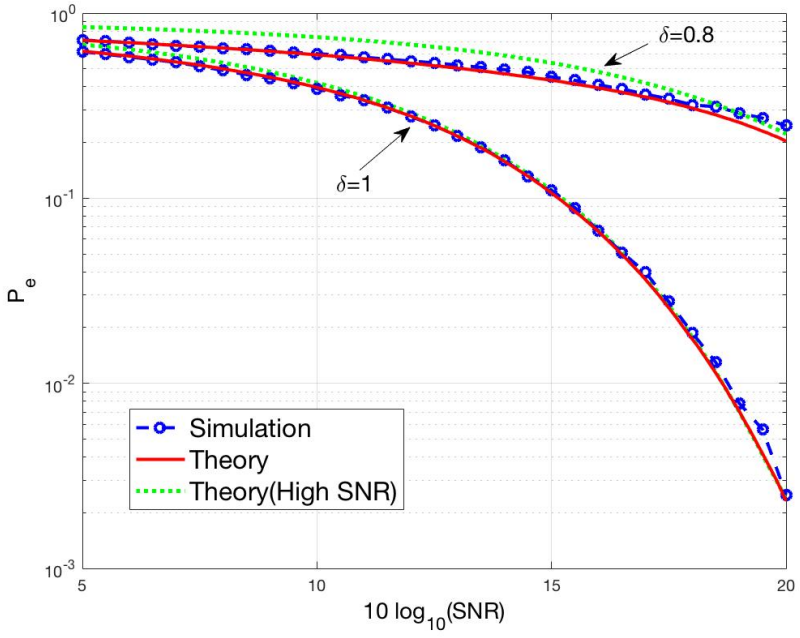


Figure 7.2: SER Performance of the Box Relaxation for 16-QAM: P_e as a function of SNR for the two cases where $\delta = .8$ and $\delta = 1$. The theoretical prediction follows from Theorem 18 and the high-SNR analysis comes from Section 7.3. For the simulation, we used signals of size $p = 128$ with each entry chosen uniformly at random in the set $\mathbf{S}_{QAM} = \{\pm 1, \pm 3\}^2$. The data are averages over 30 independent realizations of the channel matrix and the noise vector.

in $\tau^* = \sqrt{\frac{2\text{SNR}(\delta - 3/4)}{\delta}}$ for (CR) and $\tau^* = \sqrt{\frac{2\text{SNR}(\delta - 3/4 + 1/N)}{\delta}}$ for (CHR). Putting these values for τ^* in (7.14) yields their corresponding high-SNR symbol error probabilities,

$$P_e^{(\text{CR})} = \mathbb{P}\left(|G| > \tan(\frac{\pi}{N})(H + \sqrt{2\text{SNR} \cdot (\delta - 3/4)})\right), \tag{7.15}$$

$$P_e^{(\text{CHR})} = \mathbb{P}\left(|G| > \tan(\frac{\pi}{N})(H + \sqrt{2\text{SNR} \cdot (\delta - 3/4 + 1/N)})\right). \tag{7.16}$$

The difference between phase-transitions of these two cases can be observed from equations (7.15) and (7.16). While for (CR) we need $\delta > 3/4$ for consistent data recovery, this threshold changes to $\delta > (3/4 - 1/N)$ for (CHR). This essentially means that p/N additional MIMO receivers is required at the expense of having a simpler convex set. This verifies the fact that while the optimization over $\mathbf{S}^{(\text{CR})}$ might be done faster over the Circular Relaxation due to its simple structure, we need more measurements (or higher SNR) to get the same performance for (CR)

compared to (CHR). In other words, the performance of (CR) is $10\log_{10}(\frac{\delta-3/4+1/p}{\delta-3/4})$ off that of (CHR).

Comparison to the matched filter bound. The matched filter is the ideal impractical case where we assume to have the first $p-1$ entries of β_0 and we want to recover the last entry. We compare the symbol error probability of this scenario, referred to as the Matched Filter Bound (MFB), with the P_e of the convex decoder that can be derived from Theorem 18. The matched filter bound corresponds to the probability of error in detecting $X \in \mathcal{D}$ from $\tilde{\mathbf{y}} = X\mathbf{x} + \mathbf{z}$, where $\mathbf{x} \in \mathbb{C}^p$ with Gaussian entries drawn from $\mathcal{N}_{\mathbb{C}}(0, \frac{1}{\sqrt{p}})$, and $\mathbf{z}$ is the noise vector with entries $\mathcal{N}_{\mathbb{C}}(0, \sigma^2)$. Then, the probability of error of the ML estimator of X in N-PSK will be

$$\mathbb{P}\left(|G| > \tan(\frac{\pi}{N})(H + \sqrt{2\mathrm{SNR}\cdot\delta})\right) \ . \tag{7.17}$$

Comparison of (7.15) with (7.17) shows that in the high-SNR regime the performance of (CR) is $10\log_{10}(\frac{\delta}{\delta-3/4})$dB off from the (MFB). In particular, in the square case ($\delta = 1$), where the number of receive and transmit antennas are the same, the (CR) is 6dB off the (MFB). Besides, as $\delta \to \infty$ (meaning that the number of antennas grows largecompared to users), the performance of (CR) and (CHR) approaches (MFB).

N^2-QAM Constellation

In N^2-QAM, each entry of β_0 is randomly chosen from the set $\mathcal{D}$ defined in Section 7.2 with distribution p_X. The conventional relaxation for this constellation is the Box Relaxation (BR) [108, 162, 198] defined in (7.9). Similar to the previous section, In order to use Theorem 18, we need to form the projection function to $\mathbf{S}$ in (7.1) which is straightforward for a box set. Once τ^* is obtained using equation (7.12) (or recruiting other methods to solve (7.10)), we shall use (7.11) to calculate P_e of N^2-QAM constellation. Here, unlike the N-PSK case, the probability of error in the recovery is not the same for different symbols in $\mathcal{D}$.

Using the same set of arguments in Section 7.3, it can be shown that in the high-SNR regime, the last term in the objective function of (7.10) approaches $\frac{1}{2\tau\delta N}$. Therefore the answer to the minimization problem will be $\tau^* = \sqrt{\frac{2\mathrm{SNR}(\delta-(N-1)/N)}{\delta}}$. This implies that $\delta^\star = \frac{N-1}{N}$ is the recovery threshold for the Box Relaxation of the set $\mathcal{D}$. It can also be shown that for $\delta > \delta^\star$, the problem (7.10) is strictly convex

and therefore has a unique solution. This is consistent with the result of [91] which proves the same phase-transition region for the Box Relaxation.

7.4 Proof Outline

In this section we introduce the main ideas used in the proof of Theorem 18. The goal is to analyze the performance of the following optimization problem:

$$\hat{\beta} = \arg\min_{\beta \in \mathbf{S}^p} \frac{1}{2n}||\mathbf{y} - \mathbf{X}\beta||^2 \tag{7.18}$$

We rewrite (7.18) by changing variable to vector $\mathbf{w} = \beta - \beta_0$

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbf{S}^p - \beta_0} \frac{1}{2n}||\mathbf{z} - \mathbf{X}\mathbf{w}|| \tag{7.19}$$

Now let $\tilde{\mathbf{X}} = [\mathbf{X}_R, \mathbf{X}_I; -\mathbf{X}_I, \mathbf{X}_R] \in \mathbb{R}^{2n \times 2p}$ and $\mathbf{z}' = [\mathbf{z}_R; \mathbf{z}_I] \in \mathbb{R}^{2n}$, where $\mathbf{X}_R$ and $\mathbf{z}_R$ ($\mathbf{X}_I$ and $\mathbf{z}_I$) are the real (imaginary) parts of $\mathbf{X}$ and $\mathbf{z}$, respectively. Now (7.19) can be written as

$$\mathbf{w}^\star = \arg \min_{\substack{\mathbf{w} \in \mathbb{R} b b^{2p} \\ \mathbf{w}_i + j\mathbf{w}_{n+i} \in \mathbf{S} - (\beta_0)_i}} \frac{1}{4n}||\mathbf{z}' - \frac{1}{\sqrt{2p}}\tilde{\mathbf{x}} \cdot \mathbf{w}||^2 \ . \tag{7.20}$$

This optimization is difficult to analyze and current methods for asymptotic analysis of such optimizations fail here, because of the dependence between the entries of $\tilde{\mathbf{X}}$. The main step of our proof is to show that in the asymptotic regime when $p, n =\!\!\rightarrow \infty$ with $n/p = \delta$, the SER in the optimization (7.19) converges to the one in the following.

$$\mathbf{w}^\star = \arg \min_{\substack{\mathbf{w} \in \mathbb{R}^{2p} \\ \mathbf{w}_i + j\mathbf{w}_{p+i} \in \mathbf{S} - (\beta_0)_i}} \frac{1}{4n}||\mathbf{z}' - \frac{1}{\sqrt{2p}}\mathbf{B} \cdot \mathbf{w}||^2 \ . \tag{7.21}$$

Here, $\mathbf{z}' \in \mathbb{R}^{2n}$ is a vector with i.i.d. $\mathcal{N}(0, \sigma^2/2)$ entries and $\mathbf{B} \in \mathbb{R}^{(2n)\times(2p)}$ is a matrix whose entries are independently drawn from $\mathcal{N}(0, 1)$. To do so, we first show this in the case that both the objective functions have an extra strongly convex term $\epsilon\|\mathbf{w}\|^2/2$. Under this scenario, we can utilize the Lindeberg method as in [3, 130]. In equations (6.15) followed by Theorems 15 and 16 we prove that these two optimizations converge to the same value (with the difference that for our optimizations we should change two dependent rows of $\mathbf{B}$ with two dependent rows of $\tilde{\mathbf{X}}$).

The idea is to replace the rows of $\tilde{\mathbf{X}}$ to $\mathbf{B}$ in n steps. In each step, we replace the

rows i and $p+i$ in the $\tilde{\mathbf{X}}$ (that are independent from the rest of $\tilde{\mathbf{X}}$) with the the rows i and $i+p$ in the $\mathbf{B}$. We can show that each step changes the SER on the order of $O(p^{5/4})$. So as $n \to \infty$, the SER doesn't change. Next, we use the RIP condition for Gaussian matrices to show that removing the extra $\epsilon\|\mathbf{w}\|^2/2$ term in the optimization does not affect the SER for small enough ϵ (See Sections 3.1 and 3.3.2 in the appendix of [131] for more details regarding these two steps). Then, we just need to analyze performance of (7.21) instead of (7.19). For the rest of the proof, we apply the CGMT framework and the same tools as in [165](Section 5.3). The idea is to rewrite (7.21) as the following min-max problem,

$$\min_{\mathbf{w}_i+j\mathbf{w}_{p+i}\in\mathbf{S}-(\beta_0)_i} \max_{\mathbf{u}\in\mathbb{R}^{2n}} \frac{1}{\sqrt{2n}}\mathbf{u}^t\mathbf{z}' - \frac{1}{2\sqrt{pn}}\mathbf{u}^t\mathbf{B}\mathbf{w} - \frac{1}{2}||\mathbf{u}||^2 , \tag{7.22}$$

This enables us to apply the CGMT which associates with (7.22), the following simplified optimization whose analysis provides us with the desired properties of the initial optimization.

$$\min_{\mathbf{w}_i+j\mathbf{w}_{p+i}\in\mathbf{S}-(\beta_0)_i} \max_{\mathbf{u}} \frac{1}{\sqrt{2n}}\mathbf{u}^t\mathbf{z}' - \frac{1}{2}||\mathbf{u}||^2 + \frac{1}{2\sqrt{pn}}(\mathbf{g}^t\mathbf{u}||\mathbf{w}|| + \mathbf{h}^t\mathbf{w}_0||\mathbf{u}||) ,$$

where $\mathbf{g} \in \mathbb{R}^{2n}$ and $\mathbf{h} \in \mathbb{R}^{2p}$ have i.i.d. standard Gaussian entries. It can be shown that the optimization over $\mathbf{u}$ results in

$$\min_{\mathbf{w}_i+j\mathbf{w}_{p+i}\in\mathbf{S}-(\beta_0)_i} \frac{1}{\sqrt{2n}}||\mathbf{z}' + \frac{||\mathbf{w}||}{\sqrt{2p}}\mathbf{g}|| + \frac{1}{2\sqrt{pn}}\mathbf{h}^t\mathbf{w} . \tag{7.23}$$

Using $\sqrt{x} = \min_{\tau>0} \frac{1}{2\tau} + \frac{\tau x}{2}$, optimization (7.23) can be written as

$$\min_{\tau>0} \frac{1}{2\tau} + \frac{\tau\|\mathbf{z}'\|^2}{4n} + \min_{\mathbf{w}_i+j\mathbf{w}_{p+i}\in\mathbf{S}-(\beta_0)_i} \frac{\tau||\mathbf{w}||^2\|\mathbf{g}\|^2}{8np} + \frac{1}{2\sqrt{pn}}\mathbf{h}^t\mathbf{w}. \tag{7.24}$$

Using dimension reduction techniques, we can show that from the following deterministic optimization, we can tightly infer the properties of (7.24).

$$\min_{\tau>0} \frac{1}{2\tau} + \frac{\tau\sigma^2}{4} + \min_{\mathbf{w}_i+j\mathbf{w}_{p+i}\in\mathbf{S}-(\beta_0)_i} \frac{\tau||\mathbf{w}||^2}{4p} + \frac{1}{2\sqrt{pn}}\mathbf{h}^t\mathbf{w}. \tag{7.25}$$

A completion of squares in the minimization over $\mathbf{w}$, the weak law of large numbers and convex techniques (See Section A.3 in [165] to see how WLLN can be applied here) results in the final deterministic optimization

$$\min_{\tau>0} \frac{\delta-1}{2\tau\delta} + \frac{\sigma^2\tau}{4} + \frac{\tau}{4}\mathbb{E}[\mathrm{Dist}^2_{\mathbf{S}}(X + \frac{H_c}{\tau\sqrt{\delta}})]. \tag{7.26}$$

Besides, the optimal $\mathbf{w}$ can be obtained by putting the optimizer of (7.26) in the minimization over $\mathbf{w}$ in the last term of the (7.25). Similar to the proof of [166](section 3), SER of $\mathbf{w}^*$ derived here is equal to the one from (7.19) which is

$$P_e \rightarrow \mathbb{P}\left(\mathcal{P}_{\mathbf{S}}(X + \frac{H_{\mathbb{C}}}{\tau^* \sqrt{\delta}}) \notin \mathbf{S}_X\right) \tag{7.27}$$

Chapter 8

ACHIEVING NEAR MAXIMUM-LIKELIHOOD PERFORMANCE IN MASSIVE MIMO

8.1 Introduction

The last several decades has seen a sustained exponential growth rate in wireless traffic (known as Cooper's law). Due to increasing applications of wireless networks, such as real-time video, wireless gaming, virtual reality, the emerging internet-of-things, etc., this exponential growth is forecast to continue unabatedly into the future. The goal of the upcoming 5G standard is to enable reliable communication at these higher data rates (often 100 Mbits/sec and beyond). One of the most promising technologies suggested for 5G is Massive MIMO, where each base station is equipped with a very large number of antennas (many tens to even hundreds). This increases the capacity of the network many-fold (ideally, by the number of base station antennas) without adding new base stations or increasing the frequency spectrum.

in this section, we address the signal processing challenge in Massive MIMO [126, 185], where the base station is confronted with a deluge of data, coming from potentially several hundred independent streams, and is required to reliably and efficiently recover the data. The maximum-likelihood (ML) estimator, which is the desireable theoretical solution for this problem, can not be computed exactly due to its combinatorial nature [83, 185]. Various heuristics, with tractable computational complexities, have been proposed to approximate the ML solution often have performance quite distant from ML.

Due to practical advantages, convex-optimization-based methods have gained significant attenstions in the recent years. The performance of convex relaxation for signal recovery in MIMO communication systems have been analyzed in [9, 166, 174] for real-valued constellations (e.g. BPSK, PAM), and very recently, in [1], it has been extended to the complex-valued constellations (e.g. PSK, QAM).

Here, we develop an algorithm that employs a combination of convex and non-convex techniques. Our proposed method essentially exploits the solution of a convex optimization as the initial point for a local search method. We provide the-

oretical bounds on the performance of *symbol update* as the post-processing local search method. Combining this with the theoretical guarantees on the performance of the convex optimization, we will find the regimes of SNR under which our proposed method has a performance very close to the ML estimate. Our numerical simulations validates this result by showing that the proposed method has a performance very close to the matched filter bound (which itself is a lowerbound on the performance of the ML estimate.)

8.2 Problem Formulation

Notations We gather here the notations that are used throughout the paper. The bold lower letters are reserved for vectors and upper letters are used for matrices. For a vector $\mathbf{v}$, $\mathbf{v}_i$ represent its i^{th} entry, and for a matrix $\mathbf{M}$, $\mathbf{M}_k$ represent its k^{th} column. We reserve the letter j for the complex unit. For a complex scalar $x \in \mathbb{C}$, $\text{Re}\{x\}$, and $\text{Im}\{x\}$ correspond to its real and imaginary part, and $|x|$ is its absolute value. $\mathcal{N}_{\mathbb{C}}(\mu, \sigma^2)$ denotes the complex Gaussian distribution, with real and imaginary parts drawn independently from $\mathcal{N}(\mu_{\text{Re}}, \sigma^2/2)$, and $\mathcal{N}(\mu_{\text{Img}}, \sigma^2/2)$. $Q(\cdot)$ denotes the tail distribution of standrad normal distribution, and $\xrightarrow{P}$ indicates convergence in probability. $X \sim p_X$ implies that the random variable X has a density p_X. For a set $\mathcal{S}$, $\text{conv}(\mathcal{S})$ denotes its convex hull.

Problem Setup Consider the problem of recovering a transmitted data vector $\beta_0 \in \mathcal{D}^p$, where p is the number of data streams and $\mathcal{D} \subset \mathbb{C}$ defines the transmit constellation (QAM, PSK, etc.) for the streams. In our setting, p can be considered to be the number of mobile users operating in a cell at a certain frequency band. Assuming that the base station is employing n receive antennas, we seek to recover $\beta_0 \in \mathcal{D}^p$ from the noisy linear relation $\mathbf{y} = \sqrt{\frac{\text{SNR}}{p}}\mathbf{X}\beta_0 + \mathbf{z} \in \mathbb{C}^n$, where, $\mathbf{X} \in \mathbb{C}^{n\times p}$ is the known MIMO channel matrix and $\mathbf{z} \in \mathbb{C}^n$ is the unknown noise vector. We consider a random setup where the entries of $\mathbf{X}$ and $\mathbf{z}$ are both random. The noise vector $\mathbf{z}$ has i.i.d. zero-mean unit-variance circularly-symmetric complex Gaussian ($\mathcal{N}_{\mathbb{C}}(0, 1)$) entries. In our analysis, we will also assume that the entries of $\mathbf{X}$ are i.i.d. $\mathcal{N}_{\mathbb{C}}(0, 1)$, in which case SNR is, in fact, the signal-to-noise-ratio. This is a reasonable assumption when the environment is rich-scattering and there is sufficient separation between the receive antennas.

When the symbols of β_0 are chosen uniformly at random, the estimator that minimizes the block probability of error is the maximum-likelihood (ML) estimator

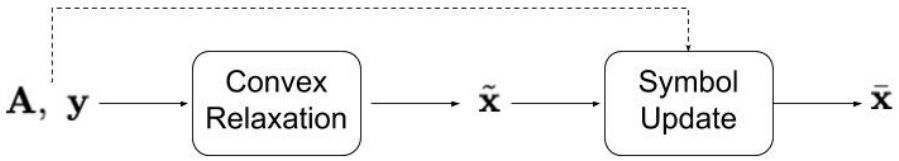


Figure 8.1: The proposed two-pronged algorithm.

which is given by

$$\hat{\beta}_{\mathrm{ML}} = \arg \min_{\beta \in \mathcal{D}^p} \frac{1}{2n} \left\| \mathbf{y} - \sqrt{\frac{\mathrm{SNR}}{p}} \mathbf{X}\beta \right\|^2 . \tag{8.1}$$

This problem is computationally intractable (in fact, known to be NP hard) when p and n are large. As a result, many computationally-efficient heuristics have been proposed to approximately solve (8.1), including zero-forcing, mmse, and decision-feedback-equalization. Unfortunately, these methods underperform ML by quite a bit (especially when p and n are large). Furthermore, they often lack rigorous performance analysis, especially in the case of decision-feedback-equalization. An exact method for solving (8.1) is the sphere decoder [70]. While it has reasonable complexity when p and n are small, it cannot be scaled to the dimensions encountered in massive MIMO [83].

We propose a two-pronged attack to develop methods that can efficiently achieve near ML performance. However, before doing so, let us examine more closely the performance of (8.1). This will give us a benchmark against which we can compare our methods. As mentioned, computing the ML solution of (8.1) is practically impossible. Furthermore, a rigorous formula for the SER (symbol-error-rate) corresponding to (8.1) is not known. Using the heuristic replica method from statistical physics, Tanaka [163] gave a formula for the SER of (8.1) when the modulation is BPSK, i.e., $\mathcal{D} = \{\pm 1\}$ and all signals are real. Since we cannot be certain of the validity of that formula, we use a rigorous *lower bound* on the SER of the ML estimator which is called the matched filter bound (MFB). This bound is obtained by assuming that a genie provides the receiver with the correct value of all symbols in β_0 except for the first one, and we use the maximume-likelihood estimator for just the first symbol (instead of the whole block). This is a lower bound on the SER of any estimator and we shall use it for comparison purposes.

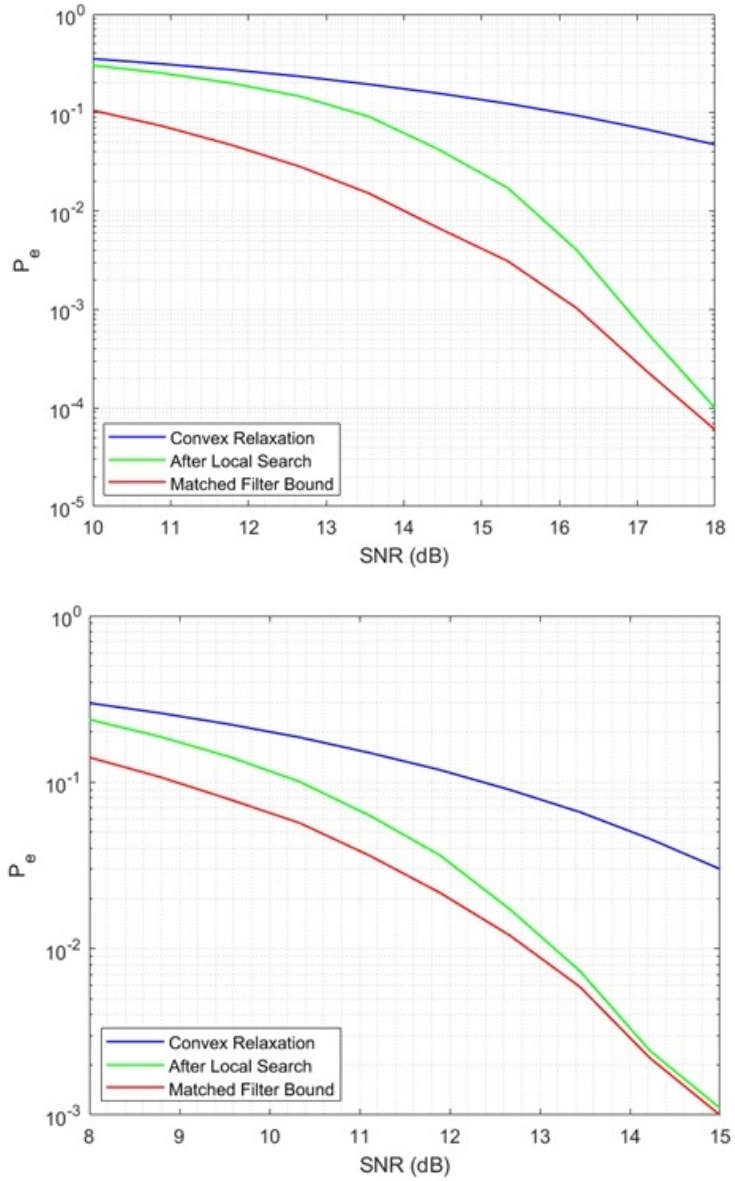


Figure 8.2: SER Performance of the convex relaxation (Blue line), After doing the local search (Green line) and the Matched Filter Bound (Red line) for 8-PSK: SER as a function of SNR for the two cases. For the simulation, we used signals of size n = 128 with each entry chosen randomly uniform from 8-PSK constellation. The data are averages over 100 independent realizations of the channel matrix and the noise vector. The left figure corresponds to $\delta = .9$ and for the right figure $\delta = 1.1$.

Two-step Algorithm

Our two-step algorithm computes the solution to a convex optimization and then perform a local search to further improves the performance. We mathematically present our proposed method in Algorithm 2. The convex optimization (8.2) inimizes the loss function in (8.1) over the set C, where C is a convex set that contains

all the constellation points. After finding the solution to the optimization program, we map it to closest point in the constellation. This step can be done entrywise and therefore efficiently.

In the second step of the algorithm, we perform a local search method. For our analysis, we use a simple symbol update algorithm, which iteratively checks if changing a certain entry can lower the cost function. Note that the optimization program (8.3) has only $|\mathcal{D}|$ feasible points.

Figure 8.2 shows the performance of our two step algorithm, in terms of the symbol-error-rate (SER), with respect to SNR, for PSK and QAM constellations. As depicted in the figure, the second step makes considerable contribution to the performance, especially for larger values of SNR. Besides, we see that the performance of the two-step algorithm gets stupendously close to MFB for larger SNR values.

Algorithm 2 Two-pronged Algorithm

Step 1: Convex Relaxation

- ⋄ Choose a convex set C, such that $\mathcal{D}^p \subseteq C$,
- ⋄ Solve the following convex optimization problem,

$$\hat{\beta} = \arg\min_{\beta \in C} \frac{1}{2n} ||\mathbf{y} - \sqrt{\frac{\mathrm{SNR}}{p}} \mathbf{X}\beta||^2, \tag{8.2}$$

- ⋄ Map the result to the closest point in constellation, i.e., for $i = 1, 2, \ldots, p$, i.e. $\tilde{\beta}_i := \arg\min_{\beta \in \mathcal{D}} |\beta - \hat{\beta}_i|$.

Step 2: Local Search (Symbol Update)

- ⋄ For $i = 1, 2, \ldots, p$:

$$\begin{aligned} \bar{\beta}_i = \arg\min_{\beta \in \mathcal{D}^p} \; & ||\mathbf{y} - \sqrt{\frac{\mathrm{SNR}}{p}} \mathbf{X}\beta|| \\ \text{s.t.} \quad & \beta_j = \tilde{\beta}_j \;,\; \forall j \neq i. \end{aligned} \tag{8.3}$$

Analysis of the convex relaxation

The convex-optimization-based methods [108, 162, 198] have gained significant attentions in recent years mainly due to the availability of fast convex solvers as well as the theoretical guarantees that can be provided on their performance. The difficulty in solving (8.1) is the nonconvexity of the constraint set $\mathcal{D}^p$ (the set is, in

fact, discrete). One approach to approximating (8.1) is to relax the discrete set to a convex one. If we remove the constraint entirely and write

$$\tilde{\beta}_{\mathrm{ZF}} = q\left(\arg\min_{\beta} \frac{1}{2n}\left\|\mathbf{y} - \sqrt{\frac{\mathrm{SNR}}{p}}\mathbf{X}\beta\right\|^2\right), \tag{8.4}$$

where q simply quantizes each entry of its argument to the closest constellation point in $\mathcal{D}$, we obtain the zero-forcing equalizer. Similarly, if we relax $\mathcal{D}^p$ to the ball $\|\beta\| \leq c$, for an appropriate constant c, then the MMSE equalizer can be written as

$$\tilde{\beta}_{\mathrm{MMSE}} = q\left(\arg\min_{\|\beta\|\leq c} \frac{1}{2n}\left\|\mathbf{y} - \sqrt{\frac{\mathrm{SNR}}{p}}\mathbf{X}\beta\right\|^2\right). \tag{8.5}$$

For example, for BPSK we should take $c = \sqrt{p}$.
Clearly, neither of the above relaxations are the best convex relaxations that one can use. The best would be to relax $\mathcal{D}$ to its convex hull, $\mathbf{S} = \mathrm{conv}(\mathcal{D})$, in which case the estimator becomes,

$$\tilde{\beta}_{\mathrm{conv}} = q\left(\arg\min_{\beta\in\mathbf{S}^p} \frac{1}{2n}\left\|\mathbf{y} - \sqrt{\frac{\mathrm{SNR}}{p}}\mathbf{X}\beta\right\|^2\right). \tag{8.6}$$

Many efficient methods for solving the above convex optimization problem exist [23, 28].

While these relaxations have been extensively used, analyzing their performance was, for a long time, an open problem. In the past 4-5 years, the performance of the convex relaxation methods have been analyzed for *real-valued* constellation in the works of [9, 91, 166, 174]. More specifically, Thrampoulidis et. al. [174] exploited a framework known as the **C**onvex **G**aussian **M**in-max **T**heorem (**CGMT**) [165, 171] to compute the performance of the box relaxation method for BPSK and PAM modulations. More recently, in [1], the authors of this paper have extended the previous results and provided a precise performance analysis for a *complex-valued* constellation, $\mathcal{D} \subset \mathbb{C}$, a general convex set C such that $\mathcal{D}^p \subset C$.
In order to compare performance of different convex relaxations, let us consider

8-PSK constellation as an example, where $\mathcal{D}_{\text{8-PSK}} = \left\{ e^{j\frac{2\pi}{8}i} | i = 0, \ldots, 7 \right\}$. Using Theorem 3.1 in [1], the symbol error rate for the aforementioned relaxations can be bounded as follows,

$$\begin{aligned}
\text{SER}_{\text{ZF}} &\leq 16 \cdot Q\left(\sqrt{\text{SNR} \cdot (\delta - 1)}\right) . \\
\text{SER}_{\text{Conv}} &\leq 16 \cdot Q\left(\sqrt{\text{SNR} \cdot (\delta - \frac{5}{8})}\right) \quad \text{and,} \\
\text{SER}_{\text{MFB}} &\leq 16 \cdot Q\left(\sqrt{\text{SNR} \cdot \delta}\right) . \qquad (8.7)
\end{aligned}$$

Here, SER_{ZF}, SER_{Conv} and SER_{MFB} correspond to the symbol error rate of zero-forcing, convex hull relaxation and matched filter bound, respectively.
More importantly, this result allows one to say something about the *distribution of errors in a single block of transmission*. It turns out that the number of errors in each block of data, i.e., the number of errors in $\tilde{\beta}$, concentrates around $p\text{SER}$ [165]. This is a remarkable result and tells us that we have a very good idea about the number of errors that appear in the output of the convex optimization step. This knowledge will be critical in developing the second step of our approach.

Analysis of the post processing (symbol update)
In this section we analyze the performance of the post-processing local search method. The goal of this step is to reduce the number of erronous symbols in the result of the first step. As depicted in Figure 8.1, the symbol update method receives an input $\tilde{\mathbf{x}} \in \mathcal{D}^p$, which is the result of the first step of the algorithm. Using the measurement matrix, $\mathbf{X}$, and the measurement vector $\mathbf{y}$, it updates the symbols one-by-one as in (8.3), to reduce the the objective value, $||\mathbf{y} - \mathbf{X}\beta||$. We analyze this step to realize the number errors this step is able to correct, in terms of a few problem parameters that we will define next.
For a constellation set $\mathcal{D} \subset \mathbb{C}$, we define $d_{\min} = d_{\min}(\mathcal{D}) := \min_{s_1 \neq s_2 \in \mathcal{D}} |s_1 - s_2|$, and $d_{\max} = d_{\max}(\mathcal{D}) := \max_{s_1, s_2 \in \mathcal{D}} |s_1 - s_2|$. We assume the entries of $\tilde{\beta}$ are independently distributed such that, $(\beta_i^*, \tilde{\beta}_i) \sim p_d$, for $i = 1, 2, \ldots, p$. Here, p_d denotes a probability mass function (PMF) over $\mathcal{D}^2$. We use p_d to define the *average error distance*, as follows,

Definition 5 *Let $P : \mathcal{D}^2 \to [0, 1]$ be a probability mass function over $\mathcal{D}^2$. The*

average error distance $\bar{D}$ is defined as,

$$\bar{D} := \bar{D}(P) := \sqrt{\mathbb{E}_{(x,y)\sim P}\left[|x-y|^2\right]} \tag{8.8}$$

Theorem 19 provides an upper bound on the SER of the output of the symbol update algorithm, in terms of the parameters $\bar{D}$ and $d_{\min}$.

Theorem 19 *Consider the symbol update algorithm as in* (8.3)*. Let* $\mathbf{X}$ *has i.i.d. complex normal entries, and for* $i = 1, 2, \ldots, p$, $(\beta_i^*, \tilde{\beta}_i) \overset{i.i.d.}{\sim} p_d$*, where* p_d *is a probability mass function over* $\mathcal{D}^2$*. Then, as* $p, n \to \infty$ *with the fixed ratio* $\delta := \frac{n}{p}$*, the symbol-error-rate of the output,* $\bar{\beta}$*, can be bounded almost surely as,*

$$SER \leq (|\mathcal{D}|-1)Q\Big(\frac{d_{min}}{\sqrt{2}}\sqrt{\frac{\delta \cdot SNR}{1+SNR\cdot \bar{D}^2}}\Big)\,. \tag{8.9}$$

Before stating the proof, the following remark is in place.

Remark 28 *It is worth nothing that the probability of error in the input,* $\tilde{\beta}$ *, lies whitin the PMF* p_d *and consequently* $\bar{D}$*. Parameter* $\bar{D}$ *is only defined for a tighter analysis of this step and can be simply bounded as* $\bar{D} \leq SER_{\tilde{\beta}} \cdot d_{max}$*, where* $SER_{\tilde{\beta}}$ *is the symbol-error-rate of the input signal* $\tilde{\beta}$*. Therefore, we have the following,*

$$SER \leq (|\mathcal{D}|-1)Q\Big(\frac{d_{min}}{\sqrt{2}}\sqrt{\frac{\delta \cdot SNR}{1+SNR\cdot d_{max}\cdot SER_{\tilde{\beta}}}}\Big)\,. \tag{8.10}$$

Proof 12 *(outline) As* $p \to \infty$*, the WLLN asserts,*

$$SER = \frac{1}{p}\sum_{i=1}^{p} \mathbf{1}_{\bar{\beta}_i \neq \beta_i^*} \xrightarrow{P} \mathbb{P}(\bar{\beta}_1 \neq \beta_1^*). \tag{8.11}$$

Recall that $\bar{\beta}_1$ *is the output of the symbol update which is computed via solving the following optimization,*

$$\bar{\tilde{\beta}}_1 = \arg\min_{s\in\mathcal{D}} ||\mathbf{y} - \sqrt{SNR}\big(\mathbf{X}\tilde{\beta} + (s-\tilde{\beta}_1)\mathbf{X}_1\big)||\,. \tag{8.12}$$

Replacing $\mathbf{y} = \sqrt{\frac{SNR}{p}}\mathbf{X}\beta^* + \mathbf{z}$ *would result,*

$$\bar{\beta}_1 = \arg\min_{s\in\mathcal{D}} ||\sqrt{\frac{SNR}{p}}(s-\beta_1^*)\mathbf{X}_1 + \mathbf{v}||, \tag{8.13}$$

where $\mathbf{v} \in \mathbb{C}^n$ *is defined as,*

$$\mathbf{v} := \sqrt{\frac{SNR}{p}} \sum_{i=2}^{p} (\tilde{\beta}_i - \beta_i^*)\mathbf{X}_i - \mathbf{z} \ . \tag{8.14}$$

Exploiting the assumption that the entries of $\mathbf{X}_i$*'s and* $\mathbf{z}$ *are independently drawn from* $\mathcal{N}_{\mathbb{C}}(0,1)$ *distribution and* $(\beta_i^*, \tilde{\beta}_i) \overset{i.i.d.}{\sim} p_d$*, we have the followig via WLLN,*

$$||\mathbf{v}|| \to \sqrt{n(1 + SNR \cdot \bar{D}^2)} \ . \tag{8.15}$$

Next, choose $s \in \mathcal{D} - \{\beta_1^*\}$*. We have,*

$$\begin{aligned} \mathbb{P}(\bar{\beta}_1 = s) \leq \mathbb{P}(&\sqrt{\frac{SNR}{p}}|s - \beta_1^*|^2 ||\mathbf{X}_1||^2 \\ &+2Re\{(s - \beta_1^*)\mathbf{v}^\star \mathbf{X}_1\} < 0) \ . \end{aligned} \tag{8.16}$$

It can be shown that the expression in the right-hand-side of (8.16) *converges in probability to,*

$$RHS \to Q\big(\frac{|s - \beta_1^*|}{\sqrt{2}} \sqrt{\frac{\delta \cdot SNR}{1 + SNR \cdot \bar{D}^2}}\ \big) \ . \tag{8.17}$$

Since $d_{min} \leq |s - \beta_1^*|$*, we can get an upper bound by replacing* $|s - \beta_1^*|$ *with* d_{min}*. The result (upper bound on SER) is then derived by taking a union bound over all* $s \in \mathcal{D} - \{\beta_1^*\}$*.*

The following proposition can be proved by combining the results of Theorem 2.1, with theoretical guarantees on the first step that are derived in [1].

Proposition 1 *Consider the assumptions of Theorem 19. Also assume that SNR goes to infinity as* $p \to \infty$ *(with any order larger than constant). Then the relative gap between the performance of the MFB (*SER_{MFB}*) and our two step algorithm (*SER_{Alg}*) goes to zero, i.e.*

$$\frac{SER_{Alg} - SER_{MFB}}{SER_{MFB}} \to 0 \ . \tag{8.18}$$

This Proposition simply states that the proposed algorithm achives the matched filter bound (MFB) when the SNR grows at a rate greater than a constant. Due to lack of space, we defer the proof, which is a combination of Theorem 19 and performance

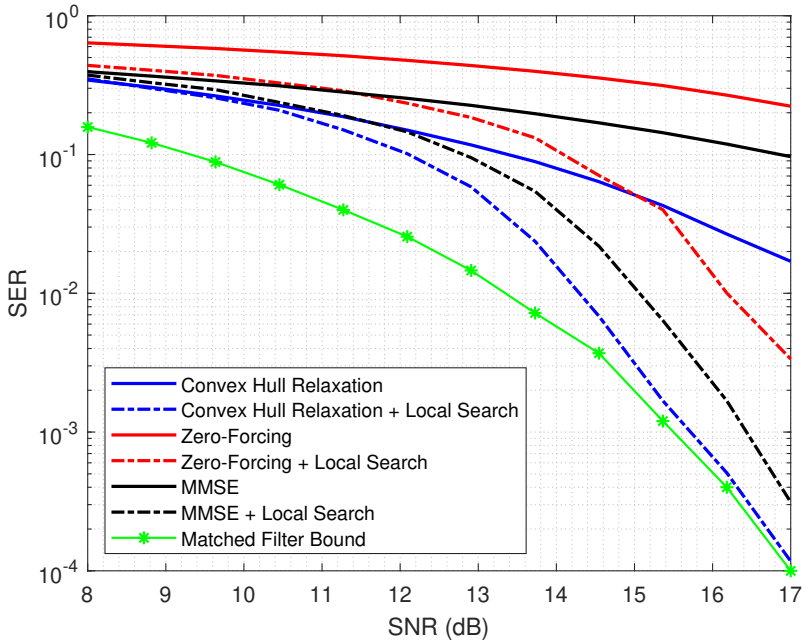


Figure 8.3: SER performance of our two-step algorithm with respect to SNR, for three choises of convex relaxation; Zero-Forcing (Red lines), MMSE (Black lines) and Convex Hull relaxation (Blue lines). The green curve corresponds to the performance of the Matched Filter Bound. Solid lines represent the performance of the first steps only, and the dashed lines are the final performance of the two-step algorithm. For our simulations we used $p = 128$ with $\frac{n}{p} = 1.1$. The results is averaged over 200 random independent realization of the channel matrix and noise vector.

analysis of the first step, to an extended version of this paper.

We conclude the section by presenting some numerical simulations that verify the validity of our proposition. Figure 8.3 shows the SER of our two-step algorithm with respect to SNR, for various convex relaxations in the first step. We used zero-forcing, MMSE, and Convex Hull relaxation for recovering an 8-PSK signal as the first step. As seen in the figure, applying the symbol update considerably improves the SER of the convex relaxation in the large SNR regime. We also reach the MFB performance as the SNR gets reasonably large. Besides, the best performance corresponds to the tightest relaxation. The convex hull relaxation performs better than MMSE which itself outperforms the Zero-Forcing.

Chapter 9

A PRECISE ANALYSIS OF PHASEMAX IN PHASE RETRIEVAL

9.1 Introduction

The[1] 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 physics such as astronomical imaging [69], X-ray crystallography [116], medical imaging [55], and optics [191]. 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.

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 [86]. 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* [11, 13, 36, 77, 87, 88, 128, 144, 190]. 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 [12, 78], *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 this method achieves perfect recovery for a nearly optimal number of random measurements. The analysis

[1] This chapter is mainly based on the work in [142]

in [12, 78, 82] 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 [52, 53]. However, for the practical case of complex signals, previous results could only provide an upper bound on C.

In this paper, we characterize the phase transition regimes for the perfect signal recovery in the PhaseMax algorithm. Our result is asymptotic and assumes that the measurement vectors are derived independently from Gaussian distribution. To the extent of our 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 developed Convex Gaussian Min-max Theorem (CGMT) [165] 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 [4, 165], massive MIMO [1, 166] and etc. CGMT has been also used by Dhifallah et. al. [53] to analyze the real version of the PhaseMax. But unfortunately, the complex case does not directly fit into the framework of CGMT. Therefore, in this paper we introduce a secondary optimization that provably has the same phase transition bounds as PhaseMax and that also can be analyzed by CGMT.

The organization of the chapter is as follows. In section 9.2 we introduce the main notations and mathematically setup the problem. In section 9.3, we present our main result followed by discussions and the result of numerical simulations. Finally, section 9.4 includes an outline of the proof of the main theorem.

9.2 Problem Setup

Notations

We gather here the basic notations that are used throughout this paper. We reserve the letter j for the complex unit. For a complex scalar $x \in \mathbb{C}$, x_{Re} and x_{Im} correspond to the real and imaginary parts of x, respectively, and $|x| = \sqrt{x_{\text{Re}}^2 + x_{\text{Im}}^2}$. $\mathcal{N}(\mu, \sigma^2)$ denotes real Gaussian distribution with mean μ and variance σ^2. Similarly, $\mathcal{N}_{\mathbb{C}}(\mu, \sigma^2)$ refers to a *complex* Gaussian distribution with real and imaginary parts drawn independently from $\mathcal{N}(\mu_{\text{Re}}, \sigma^2/2)$ and $\mathcal{N}(\mu_{\text{Im}}, \sigma^2/2)$, respectively. $\mathcal{R}(2\sigma^2)$ denotes the Rayleigh distribution with second moment equal to $2\sigma^2$. $X \sim p_X$ implies that the

random variable X has a density p_X. Bold lower letters are reserved for vectors and upper letters are used for matrices. For a vector $\mathbf{v}$, v_i denotes its i^{th} entry and $||\mathbf{v}||$ is its l_2 norm. $(\cdot)^\star$ is used to denote the conjugate transpose. For a complex vector $\mathbf{v}$, $\mathbf{v}_{\text{Re}}$ and $\mathbf{v}_{\text{Im}}$ denotes its real and complex parts, respectively. Also, $\mathbf{v}(k:l)$ is a column vector consisting of entries with index from k to l of $\mathbf{v}$. We use caligraphy letters for sets. For set $\mathcal{S}$, cone$(\mathcal{S})$ is the closed conical hull of $\mathcal{S}$.

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,

$$b_i = |\mathbf{a}_i^\star \mathbf{x}_0|, \ i = 1, \dots, m. \tag{9.1}$$

Throughout this paper we assume that $\{\mathbf{a}_i \in \mathbb{C}^n\}_{i=1}^m$ is the set of known measurement vectors where the $\mathbf{a}_i$'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 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 (9.1) into *convex* inequality constraints. This results in the following convex optimization problem:

$$\begin{aligned} \hat{x} = \arg\max_{\mathbf{x} \in \mathbb{C}^n} \quad & \text{Re}\{\mathbf{x}_{\text{init}}{}^\star \, \mathbf{x}\} \\ \text{subject to:} \quad & |\mathbf{a}_i^\star \mathbf{x}| \leq b_i \ , \quad 1 \leq i \leq m. \end{aligned} \tag{9.2}$$

This optimization searches for a feasible vector that posses the most real correlation with $\mathbf{x}_{\text{init}}$. Note that because of the global phase ambiguity of the measurements in (9.1), we can estimate $\mathbf{x}_0$ up to a global phase. Therefore, we define the following performance measure for the PhaseMax method,

$$\text{Dist}(\hat{x}, \mathbf{x}_0) = \min_{\phi \in [-\pi, \pi]} \frac{||\hat{x} e^{j\phi} - \mathbf{x}_0||}{||\mathbf{x}_0||} \ . \tag{9.3}$$

Under this setting, a perfect recovery of $\mathbf{x}_0$ means $\text{Dist}(\hat{x}, \mathbf{x}_0) = 0$. In this paper we investigate the necessary and sufficient conditions under which the optimization program (9.2) perfectly recovers the true signal.

9.3 Main Result

In this section, we present the main result of the paper which provides us with the necessary and sufficient number of measurements for the perfect recovery of the PhaseMax method in (9.2) 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 20, 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:

$$\rho_{\text{init}} := \max_{0 \le \phi < 2\pi} \frac{\text{Re}(e^{j\phi}\, \mathbf{x}_{\text{init}}^{\star}\, \mathbf{x}_0)}{||\mathbf{x}_0||\, ||\mathbf{x}_{\text{init}}||}. \tag{9.4}$$

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 (9.1)). Therefore, for convenience we assume 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 $\rho_{\text{init}} = \mathbf{x}_{\text{init}}^{\star}\, \mathbf{x}_0$. 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 paper which characterizes the phase transition regimes of PhaseMax for perfect recovery, in terms of δ and ρ_{init}.

Theorem 20 *Consider the PhaseMax problem defined in section 9.2. For a fixed oversampling ratio $\delta = \frac{m}{n} > 4$, the optimization program* (9.2) *perfectly recovers the true signal (in the sense that $\lim_{n\to\infty} \mathbb{P}(Dist(\hat{x}, \mathbf{x}_0) > \epsilon) = 0$, for any fixed $\epsilon > 0$) if and only if,*

$$\delta > \delta_{rec} := \frac{4}{\cos^2\theta} = \frac{4}{\rho_{init}^2}\,, \tag{9.5}$$

where ρ_{init} is defined in (9.4).

Theorem 20 establishes a sharp phase transition behavior for the performance of PhaseMax. The inequality (9.5) can also be rewritten in terms of θ (or ρ_{init}) when the oversampling ratio, δ, is fixed,

$$\rho_{\text{init}} = \cos\theta > \sqrt{\frac{4}{\delta}}\,. \tag{9.6}$$

The proof of Theorem 20 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 (9.2). 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 9.4.

Remark 29 *The condition $\delta > 4$ is proven to be fundamentally necessary for the phase retrieval problem under generic measurements to have a unique solution [45]. This is consistent with Theorem 20 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 30 *It is shown in the work of Goldstein et. al. [78] 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 Fig. 9.1 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 that comes from Theorem 20, perfectly predicts phase transition boundary.*

9.4 Proof Outline

In this part we introduce the main ideas used in the proof of Theorem 20. As mentioned earlier in section 9.3, we assume $\mathbf{x}_0$ is a unit norm vector aligned with $\mathbf{x}_{\text{init}}$. Due to 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 (9.2) 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 (9.2) 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 24 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 zero 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 24, these conditions will guarantee the perfect recovery in the initial PhaseMax optimization (9.2).

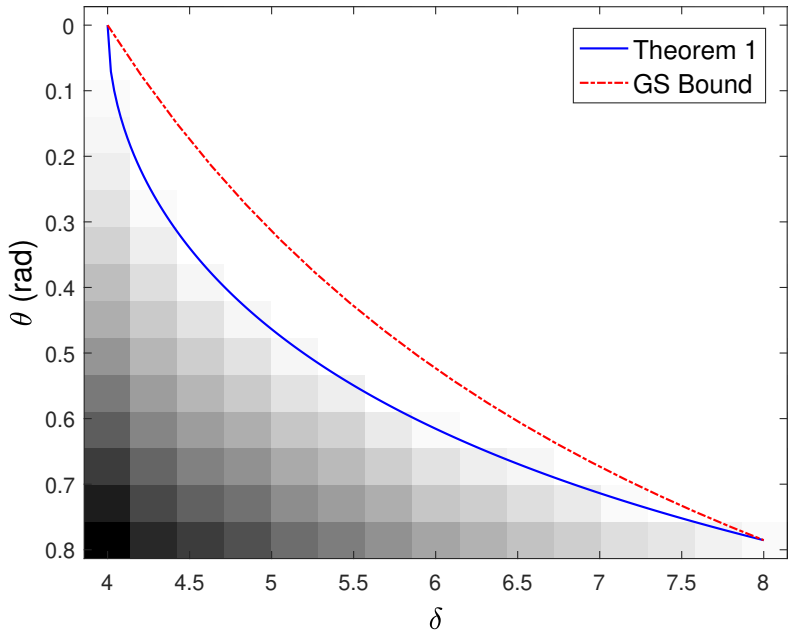


Figure 9.1: Phase transition regimes for the PhaseMax problem in terms of the oversampling ratio $\delta = m/n$ and θ, the angle between $\mathbf{x}_0$ and $\mathbf{x}_{\text{init}}$. For the empirical results, we used signals of size $n = 128$. The data is averaged over 10 independent realization of the measurement vectors. The blue line indicates the sharp phase transition bounds derived in Theorem 20 and the red line comes from the results of [78], which is referred to as the GS Bound.

Introducing the Real Optimization ERO

We define the error vector $\mathbf{w} := \mathbf{x} - \mathbf{x}_0$ and rewrite (9.2) in terms of $\mathbf{w}$,

$$\begin{aligned} \max_{\mathbf{w}\in\mathbb{C}^n} \quad & \text{Re}\{\mathbf{x}_{\text{init}}{}^{\star}\, \mathbf{w}\} \\ \text{subject to:} \quad & |\mathbf{a}_i^{\star}(\mathbf{e}_1 + \mathbf{w})| \leq b_i \ , \quad 1 \leq i \leq m. \end{aligned} \tag{9.7}$$

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

$$b_i = \tilde{\mathbf{a}}_i^{\star}\mathbf{x}_0 = (\tilde{\mathbf{a}}_i)_1, \quad \text{for} \ \ i = 1, 2, \ldots, m \ , \tag{9.8}$$

where $(\tilde{\mathbf{a}}_i)_1$ is the first entry of $\tilde{\mathbf{a}}_i$. Let $\mathcal{D} := \{\mathbf{w} \in \mathbb{C}^n : \text{Re}\{\mathbf{x}_{\text{init}}^{\star}\, \mathbf{w}\} \geq 0\}$ be the set of all vectors $\mathbf{w}$ with nonnegative objective value and $\mathcal{F} := \{\mathbf{w} \in \mathbb{C}^n : |\mathbf{a}_i^{\star}(\mathbf{e}_1 + \mathbf{w})| \leq b_i, \ \text{for} \ i = 1, 2, \ldots, m\}$ be the feasible set of the optimization problem (9.7). The

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

Lemma 21 $\mathbf{x}_0$ *is the unique optimal solution of* (9.2) *if and only if* $\mathcal{D} \bigcap \mathcal{F} = \{\mathbf{0}\}$.

Proof 13 *For* $\mathbf{w} \in \mathcal{D} \bigcap \mathcal{F}$, $\mathbf{x}_0 + \mathbf{w}$ *is a solution of* (9.2) *with an objective value greater than the value for* $\mathbf{x}_0$. *Therefore,* $\mathcal{D} \bigcap \mathcal{F} = \{\mathbf{0}\}$ *is equivalent to* $\mathbf{x}_0$ *be a local minimum of* (9.2) *which is also a global minimum due to convexity of* (9.2).

Lemma 22 $\mathcal{D} \bigcap \mathcal{F} = \{\mathbf{0}\}$ *if and only if* $\mathcal{D} \bigcap cone(\mathcal{F}) = \{\mathbf{0}\}$.

Proof 14 *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,*

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

Lemma 23 $cone(\mathcal{F}) = \bigcap_{i=1}^{m} \{\mathbf{w} \in \mathbb{C}^n : Re\{\tilde{\mathbf{a}}_i^\star \mathbf{w}\} \leq 0\}$.

Proof 15 *Let* $\mathbf{d} \in \mathcal{F}$,

$$|b_i + \tilde{\mathbf{a}}_i^\star \mathbf{d}| \leq b_i \ , \quad for \ \ i = 1, 2, \ldots, m. \tag{9.9}$$

Therefore,

$$\begin{aligned} Re\{\tilde{\mathbf{a}}_i^\star \mathbf{d}\} &= Re\{b_i + \tilde{\mathbf{a}}_i^\star \mathbf{d}\} - b_i \ , \\ &\leq |b_i + \tilde{\mathbf{a}}_i^\star \mathbf{d}| - b_i \ , \\ &\leq 0 \ . \end{aligned} \tag{9.10}$$

This shows that $cone(\mathcal{F}) \subseteq \bigcap_{i=1}^{m} \{\mathbf{w} \in \mathbb{C}^n : Re\{\tilde{\mathbf{a}}_i^\star \mathbf{w}\} \leq 0\}$. *To show the other direction, choose* $\mathbf{d} \in \mathbb{C}^n$ *such that:* $Re\{\tilde{\mathbf{a}}_i^\star \mathbf{d}\} < 0$, *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,* $\mathbf{d} \in cone(\mathcal{F})$. *This concludes the proof.*

We have the following corollary as a result of Lemma 21, Lemma 22, and Lemma 23.

Corollary 11 $\mathbf{x}_0$ *is the unique optimal solution of* (9.2) *if and only if,*

$$\{\mathbf{w} : Re\{\mathbf{x}_{init}^\star \mathbf{w}\} \geq 0 \ \ Re\{\tilde{\mathbf{a}}_i^\star \mathbf{w}\} \leq 0, \ \ for \ \ 1 \leq i \leq m\} = \{\mathbf{0}\}. \tag{9.11}$$

We are now ready to establish the equivalent real optimization ERO. We will show that the ERO has the exact phase transition bounds as PhaseMax in (9.2).

$$\begin{aligned} \max_{\mathbf{w}' \in \mathbb{R}^{2n-1}} \quad & \eta^T \mathbf{w}' \\ \text{subject to:} \quad & |\mathbf{a}_i'^T (\mathbf{e}_1 + \mathbf{w}')| \leq b_i \ , \quad 1 \leq i \leq m, \end{aligned} \tag{9.12}$$

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

$$\eta := \begin{bmatrix} \text{Re}\{\mathbf{x}_{\text{init}}\} \\ -\text{Im}\{\mathbf{x}_{\text{init}}(2:n)\} \end{bmatrix} \text{ and } \mathbf{a}_i' := \begin{bmatrix} \text{Re}\{\tilde{\mathbf{a}}_i\} \\ -\text{Im}\{\tilde{\mathbf{a}}_i(2:n)\} \end{bmatrix}, \ \forall i. \tag{9.13}$$

Here $\text{Im}\{\tilde{\mathbf{a}}_i(2:n)\}$ is the imaginary part of the last $n-1$ entries of $\tilde{\mathbf{a}}_i$. We conclude this step of the proof with the following lemma:

Lemma 24 $\mathbf{x}_0$ *is the unique optimal solution of the PhaseMax method if and only if* $\mathbf{w}' = 0$ *is the unique optimal solution of* (9.12)*.*

The proof of Lemma 24 is straightforward by defining

$$\mathbf{w}' = \begin{bmatrix} \text{Re}\{\mathbf{w}\} \\ \text{Im}\{\mathbf{w}(2:n)\} \end{bmatrix} \in \mathbb{R}^{2n-1} \ , \tag{9.14}$$

and then showing that the optimality conditions for $\mathbf{w}' = 0$ in (9.12) is equivalent to (9.11).

It is worth mentioning that the result of Lemma 24 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.

Convex Gaussian Min-Max Theorem

Our analysis is based on the recently developed Convex Gaussian Min-max Theorem (CGMT) [165]. 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:

$$\Phi(\mathbf{G}) := \min_{\mathbf{w} \in \mathcal{S}_{\mathbf{w}}} \max_{\mathbf{u} \in \mathcal{S}_{\mathbf{u}}} \mathbf{u}^T \mathbf{G} \mathbf{w} + \psi(\mathbf{u}, \mathbf{w}), \tag{9.15a}$$

$$\phi(\mathbf{g}, \mathbf{h}) := \min_{\mathbf{w} \in \mathcal{S}_{\mathbf{w}}} \max_{\mathbf{u} \in \mathcal{S}_{\mathbf{u}}} \|\mathbf{w}\| \mathbf{g}^T \mathbf{u} - \|\mathbf{u}\| \mathbf{h}^T \mathbf{w} + \psi(\mathbf{u}, \mathbf{w}), \tag{9.15b}$$

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 $\mathbf{w}_\phi := \mathbf{w}_\phi(\mathbf{g}, \mathbf{h})$ any optimal minimizers in (9.15a) and (9.15b), respectively. The following lemma is a result of CGMT [165].

Lemma 25 *Consider the two optimizations* (9.15a) *and* (9.15b). *Let* $\mathcal{S}_{\mathbf{w}}, \mathcal{S}_{\mathbf{u}}$ *be convex and compact sets,* ψ *be continuous and convex-concave on* $\mathcal{S}_{\mathbf{w}} \times \mathcal{S}_{\mathbf{u}}$, *and,* $\mathbf{G}, \mathbf{g}$ *and* $\mathbf{h}$ *all have entries iid standard normal. Suppose there exist* α *such that in the limit of* $n \to \infty$ *it holds in probability that* $\|\mathbf{w}_\phi(\mathbf{g}, \mathbf{h})\| \to \alpha$. *Then, the same holds for* $\mathbf{w}_\Phi(\mathbf{G})$ *and we have* $\|\mathbf{w}_\Phi(\mathbf{G})\| \to \alpha$.

In the next section, first we will rewrite the ERO in the form of the optimization (9.15a). This enables us to apply Lemma 25 to the ERO and derive an Auxiliary Optimization in the form of (9.15b). This lemma indicates that if $\|\mathbf{w}_\phi(\mathbf{g}, \mathbf{h})\| \to 0$ for the (AO), then $\|\mathbf{w}_\Phi(\mathbf{G})\| \to 0$ for the ERO and we have perfect recovery. (AO) can be analyzed using the 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 24 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 (9.13). Recall that $\mathbf{a}_i$'s are independently drawn from the complex Gaussian distribution with mean zero and covariance matrix $\mathbf{I}$. Therefore, the distribution of the entries of $\tilde{\mathbf{a}}_i$'s that were defined in section 9.4, is as follows:

1. 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.,

$$(\tilde{\mathbf{a}}_i)_1 \sim \mathcal{R}(1), \tag{9.16}$$

2. The remaining entries of $\tilde{\mathbf{a}}_i$ remain standard Gaussian random variables,

$$(\tilde{\mathbf{a}}_i)_k \sim \mathcal{N}_{\mathbb{C}}(0, 1), \quad \text{for } 2 \le k \le n\ , \tag{9.17}$$

3. 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 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 $\tilde{\mathbf{A}} \in \mathbb{R}^{m\times(2n-1)}$ be the remaining part (i.e., $\mathbf{A} = [\mathbf{A}_1 \ \ \tilde{\mathbf{A}}]$). $\mathbf{x}_0 = \mathbf{e}_1$ implies that $\mathbf{A}_1 = [b_1, b_2, \ldots, b_m]^T$, where b_i's are defined in (9.1). Using the Lagrange multipliers, we can reformulate (9.12) as the following minmax program,

$$\min_{\substack{w_1\in\mathbb{R}\\ \tilde{\mathbf{w}}\in\mathbb{R}^{2n-2}}} \max_{\lambda,\mu\in\mathbb{R}^m_+} \big(-\eta^T\mathbf{w} + (\lambda-\mu)^T\tilde{\mathbf{A}}\tilde{\mathbf{w}} - (\lambda+\mu)^T\mathbf{A}_1 + (\lambda-\mu)^T\mathbf{A}_1(1+w_1)\big), \tag{9.18}$$

where w_1 denotes the first entry of $\mathbf{w}$ and $\tilde{\mathbf{w}}$ is the remaining part. Define $\mathbf{v} := \lambda - \mu$. It can be shown that optimal values of (9.18) satisfy $\lambda + \mu = |\lambda - \mu|$. Here, $|\cdot|$ denotes the component-wise absolute value. Therefore, (9.18) 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:

$$\min_{\substack{w_1\in\mathbb{R}\\ \tilde{\mathbf{w}}\in\mathbb{R}^{2n-2}}} \max_{\mathbf{v}\in\mathbb{R}^m} -\eta^T\mathbf{w} + \mathbf{v}^T\tilde{\mathbf{A}}\tilde{\mathbf{w}} + \mathbf{v}^T\mathbf{A}_1(1+w_1) - |\mathbf{v}|^T\mathbf{A}_1. \tag{9.19}$$

Note that $\tilde{\mathbf{A}}$ has i.i.d. standard normal entries. One can check that (9.19) satisfies the condition of Lemma 25. Hence, we can form the (AO) as follows,

$$\min_{\substack{w_1\in\mathbb{R}\\ \tilde{\mathbf{w}}\in\mathbb{R}^{2n-2}}} \max_{\mathbf{v}\in\mathbb{R}^m} -\eta^T\mathbf{w} + \mathbf{v}^T\mathbf{g}||\tilde{\mathbf{w}}|| + ||\mathbf{v}||\mathbf{h}^T\tilde{\mathbf{w}} + \mathbf{v}^T\mathbf{A}_1(1+w_1) - |\mathbf{v}|^T\mathbf{A}_1, \tag{9.20}$$

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 (9.20) is similar to [53]. Due to lack of space, we defer technical details to the full version of the paper.

We conclude the paper with a theorem that characterizes the performance of the ERO. Let $\mathbf{w}^*$ be the optimizer of (9.20). Define $s^* := 1 + w_1^*$ and $t^* := ||\tilde{\mathbf{w}}^*||$.

Theorem 21 *In the asymptotic regime where $m, n \to \infty$, and $\delta := \frac{m}{n}$, s^* and t^* converges to the solution of the following deterministic optimization,*

$$\begin{aligned} &\max_{s\in[-1,1],\ t\ge 0} \rho_{init}\ s + \sqrt{1-{\rho_{init}}^2}\sqrt{t^2 - \frac{\delta}{2}p(t,s)} \\ &\textit{subject to: } p(t,s) \le \frac{2t^2}{\delta}. \end{aligned} \tag{9.21}$$

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

$$\begin{aligned} p(t,s) =& t^2 + (1+s)[1+s-\sqrt{t^2+(1+s)^2}] \\ &+ (1-s)[1-s-\sqrt{t^2+(1-s)^2}] \end{aligned} \tag{9.22}$$

It can be shown that $\rho_{\text{init}} > \frac{2}{\sqrt{\delta}}$ is the necessary and sufficient condition for $(t^*, s^*) = (0, 1)$ to be the unique solution of (9.21) which is equivalent to the perfect recovery in the ERO.

Chapter 10

CONCLUSION AND FUTURE WORK

We will conclude with some brief remarks n the results of this thesis, and mentioning a few related directions that worth further exploration.

The universality results we showed were for a few special, but useful and practical cases, and does not always hold. For instance, in the last chapter we observe that the phase transition we achieve for complex phase retrieval with phase-max, is not simply equivalent to the case where we replace the complex Gaussian matrix, with its corresponding real Gaussian matrix, as we did for massive MIMO. Also in Chapter 6, we have observed that the assumptions we mentioned for the regularizer are necessary. For instance, universality will not hold by choosing a nuclear norm as the regularizer, without being constraint to PSD matrices. Note that if we are constrained to PSD matrices, the nuclear norm simply becomes the trace function, which satisfies the assumptions of the Chapter 6.

The precise analysis we proposed in Chapter 2, paved the way for a few possible future directions. Calculating the best regularizer parameter λ and choosing the right loss function and regularizer based on the problem parameters are two interesting directions that we would like to consider as future work. Besides, the universality results enables us to go further, and analyze the best designs for the feature matrices $\mathbf{X}$. This can be helpful in the applications that we can manipulate the features matrix for the best performance.

Another area that has not been investigated yet, is the analysis of the count data models, such as Poisson regression. These models are very popular with tens of applications in telecommunications (number of arriving calls in a system), Biology (number of mutations on a DNA), Finance and insurance (number of losses or claims in a period of time), etc.. Using the results of Chapter 2, we can easily analyze their performance and derive results on consistency of count data models under various conditions.

Finally, new applications of generalized linear models and also non-linear models, always leads to new research directions where CGMT and such universality results framework are applicable. We hope that this thesis was a practical guide in using CGMT framework and corresponding universality results to those new directions.

BIBLIOGRAPHY

[1] Abbasi, E., Salehi, F., and Hassibi, B. (2019a). Performance analysis of convex data detection in mimo. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4554–4558. IEEE.

[2] Abbasi, E., Salehi, F., and Hassibi, B. (2019b). Sparse covariance estimation from quadratic measurements: A precise analysis. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 2074–2078. IEEE.

[3] Abbasi, E., Salehi, F., and Hassibi, B. (2019c). Universality in learning from linear measurements. In *Advances in Neural Information Processing Systems*, pages 12372–12382.

[4] Abbasi, E., Thrampoulidis, C., and Hassibi, B. (2016). General performance metrics for the lasso. In *Information Theory Workshop (ITW), 2016 IEEE*, pages 181–185. IEEE.

[5] Ait-Sahalia, Y., Fan, J., and Xiu, D. (2010). High-frequency covariance estimates with noisy and asynchronous financial data. *Journal of the American Statistical Association*, 105(492):1504–1517.

[6] Amelunxen, D., Lotz, M., McCoy, M. B., and Tropp, J. A. (2014). Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294.

[7] Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The annals of statistics*, pages 1100–1120.

[8] Arianandа, D. D. and Leus, G. (2012). Compressive wideband power spectrum estimation. *IEEE Transactions on signal processing*, 60(9):4775–4789.

[9] Atitallah, I. B., Thrampoulidis, C., Kammoun, A., Al-Naffouri, T. Y., Hassibi, B., and Alouini, M.-S. (2017). Ber analysis of regularized least squares for bpsk recovery. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 4262–4266. IEEE.

[10] Bach, F. R. (2010). Structured sparsity-inducing norms through submodular functions. In *Advances in Neural Information Processing Systems*, pages 118–126.

[11] Bahmani, S. and Romberg, J. (2015). Efficient compressive phase retrieval with constrained sensing vectors. In *Advances in Neural Information Processing Systems*, pages 523–531.

[12] Bahmani, S. and Romberg, J. (2016). Phase retrieval meets statistical learning theory: A flexible convex relaxation. *arXiv preprint arXiv:1610.04210*.

[13] Balan, R., Casazza, P., and Edidin, D. (2006). On signal reconstruction without phase. *Applied and Computational Harmonic Analysis*, 20(3):345–356.

[14] Banerjee, A., Chen, S., Fazayeli, F., and Sivakumar, V. (2014). Estimation with norm regularization. In *Advances in Neural Information Processing Systems*, pages 1556–1564.

[15] Baraniuk, R. G., Cevher, V., and Wakin, M. B. (2010). Low-dimensional models for dimensionality reduction and signal recovery: A geometric perspective. *Proceedings of the IEEE*, 98(6):959–971.

[16] Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2019). Benign overfitting in linear regression. *arXiv preprint arXiv:1906.11300*.

[17] Bayati, M., Lelarge, M., Montanari, A., et al. (2015). Universality in polytope phase transitions and message passing algorithms. *The Annals of Applied Probability*, 25(2):753–822.

[18] Bayati, M. and Montanari, A. (2012). The lasso risk for gaussian matrices. *Information Theory, IEEE Transactions on*, 58(4):1997–2017.

[19] Bean, D., Bickel, P. J., El Karoui, N., and Yu, B. (2013). Optimal m-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences*, 110(36):14563–14568.

[20] Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

[21] Belkin, M., Hsu, D. J., and Mitra, P. (2018). Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. In *Advances in neural information processing systems*, pages 2300–2311.

[22] Belloni, A., Chernozhukov, V., and Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806.

[23] Ben-Tal, A. and Nemirovski, A. (2001). *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*, volume 2. Siam.

[24] Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics*, pages 199–227.

[25] Billingsley, P. (1979). *Probability and measure*. N.Y.: Wiley.

[26] Boche, H., Calderbank, R., Kutyniok, G., and Vybíral, J. (2015). *Compressed sensing and its applications*. Springer.

[27] Boyd, C. R., Tolson, M. A., and Copes, W. S. (1987). Evaluating trauma care: the triss method. trauma score and the injury severity score. *The Journal of trauma*, 27(4):370–378.

[28] Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

[29] Bradic, J. and Chen, J. (2015). Robustness in sparse linear models: relative efficiency based on robust approximate message passing. *arXiv preprint arXiv:1507.08726*.

[30] Bunea, F. et al. (2008). Honest variable selection in linear and logistic regression models via l1 and l1+ l2 penalization. *Electronic Journal of Statistics*, 2:1153–1194.

[31] Cai, T. T., Zhang, A., et al. (2015). Rop: Matrix recovery via rank-one projections. *The Annals of Statistics*, 43(1):102–138.

[32] Cai, T. T., Zhang, C.-H., Zhou, H. H., et al. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144.

[33] Candes, E. J. (2008). The restricted isometry property and its implications for compressed sensing. *Comptes rendus mathematique*, 346(9-10):589–592.

[34] Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717.

[35] Candes, E. J., Romberg, J. K., and Tao, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223.

[36] Candes, E. J., Strohmer, T., and Voroninski, V. (2013). Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274.

[37] Candès, E. J. and Sur, P. (2018). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *arXiv preprint arXiv:1804.09753*.

[38] Candes, E. J. and Tao, T. (2006). Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE transactions on information theory*, 52(12):5406–5425.

[39] Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. (2010). Latent variable graphical model selection via convex optimization. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1610–1613. IEEE.

[40] Chandrasekaran, V., Recht, B., Parrilo, P. A., and Willsky, A. S. (2012). The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6):805–849.

[41] Charles, J., Ramina, G., Arian, M., and Christoph, S. (2015). Optimality of large mimo detection via approximate message passing. In *Information Theory (ISIT), 2015 IEEE International Symposium on*. IEEE.

[42] Chen, S. S., Donoho, D. L., and Saunders, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159.

[43] Chen, Y., Chi, Y., and Goldsmith, A. J. (2014). Estimation of simultaneously structured covariance matrices from quadratic measurements. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673. IEEE.

[44] Chen, Y., Chi, Y., and Goldsmith, A. J. (2015). Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory*, 61(7):4034–4059.

[45] Conca, A., Edidin, D., Hering, M., and Vinzant, C. (2015). An algebraic characterization of injectivity in phase retrieval. *Applied and Computational Harmonic Analysis*, 38(2):346–356.

[46] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

[47] Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers*, pages 326–334.

[48] Dasarathy, G., Shah, P., Bhaskar, B. N., and Nowak, R. D. (2015). Sketching sparse matrices, covariances, and graphs via tensor products. *IEEE Transactions on Information Theory*, 61(3):1373–1388.

[49] d'Aspremont, A., Banerjee, O., and El Ghaoui, L. (2008). First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 30(1):56–66.

[50] Davenport, M. A., Duarte, M. F., Eldar, Y. C., and Kutyniok, G. (2011). Introduction to compressed sensing. *Preprint*, 93:1–64.

[51] Deng, Z., Kammoun, A., and Thrampoulidis, C. (2019). A model of double descent for high-dimensional binary linear classification. *arXiv preprint arXiv:1911.05822*.

[52] Dhifallah, O. and Lu, Y. M. (2017). Fundamental limits of phasemax for phase retrieval: A replica analysis. *arXiv preprint arXiv:1708.03355*.

[53] Dhifallah, O., Thrampoulidis, C., and Lu, Y. M. (2017). Phase retrieval via linear programming: Fundamental limits and algorithmic improvements. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1071–1077. IEEE.

[54] Dhifallah, O., Thrampoulidis, C., and Lu, Y. M. (2018). Phase retrieval via polytope optimization: Geometry, phase transitions, and new algorithms. *arXiv preprint arXiv:1805.09555*.

[55] Dierolf, M., Menzel, A., Thibault, P., Schneider, P., Kewish, C. M., Wepf, R., Bunk, O., and Pfeiffer, F. (2010). Ptychographic x-ray computed tomography at the nanoscale. *Nature*, 467(7314):436–439.

[56] Donoho, D. and Montanari, A. (2016). High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3-4):935–969.

[57] Donoho, D. and Tanner, J. (2009a). Counting faces of randomly projected polytopes when the projection radically lowers dimension. *Journal of the American Mathematical Society*, 22(1):1–53.

[58] Donoho, D. and Tanner, J. (2009b). Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4273–4293.

[59] Donoho, D. L. (2006). High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension. *Discrete & Computational Geometry*, 35(4):617–652.

[60] Donoho, D. L. et al. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, pages 1–32.

[61] Donoho, D. L. et al. (2006). Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306.

[62] Donoho, D. L., Maleki, A., and Montanari, A. (2009). Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919.

[63] Donoho, D. L., Maleki, A., and Montanari, A. (2011). The noise-sensitivity phase transition in compressed sensing. *Information Theory, IEEE Transactions on*, 57(10):6920–6941.

[64] Donoho, D. L. and Montanari, A. (2015). Variance breakdown of huber (m)-estimators: n/p \rightarrow m\in (1,\infty). *arXiv preprint arXiv:1503.02106*.

[65] El Karoui, N. (2018). On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170(1-2):95–175.

[66] El Karoui, N., Bean, D., Bickel, P. J., Lim, C., and Yu, B. (2013). On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562.

[67] El Karoui, N. et al. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics*, 36(6):2717–2756.

[68] Eldar, Y. C. and Kutyniok, G. (2012). *Compressed sensing: theory and applications*. Cambridge University Press.

[69] Fienup, C. and Dainty, J. (1987). Phase retrieval and image reconstruction for astronomy. *Image Recovery: Theory and Application*, pages 231–275.

[70] Fincke, U. and Pohst, M. (1985). Improved methods for calculating vectors of short length in a lattice, including a complexity analysis. *Mathematics of computation*, 44(170):463–471.

[71] Foschini, G. J. (1996). Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas. *Bell labs technical journal*, 1(2):41–59.

[72] Foucart, S. and Rauhut, H. (2017). A mathematical introduction to compressive sensing. *Bull. Am. Math*, 54:151–165.

[73] Foygel, R. and Mackey, L. (2014). Corrupted sensing: Novel guarantees for separating structured signals. *IEEE Transactions on Information Theory*, 60(2):1223–1247.

[74] Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., et al. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

[75] Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

[76] Furrer, R. and Bengtsson, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in kalman filter variants. *Journal of Multivariate Analysis*, 98(2):227–255.

[77] Goemans, M. X. and Williamson, D. P. (1995). Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6):1115–1145.

[78] Goldstein, T. and Studer, C. (2016). Phasemax: Convex phase retrieval via basis pursuit. *arXiv preprint arXiv:1610.07531*.

[79] Gordon, Y. (1985). Some inequalities for gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289.

[80] Gordon, Y. (1988). On milman's inequality and random subspaces which escape through a mesh in r n. In *Geometric Aspects of Functional Analysis*, pages 84–106. Springer.

[81] Grötschel, M., Lovász, L., and Schrijver, A. (2012). *Geometric algorithms and combinatorial optimization*, volume 2. Springer Science & Business Media.

[82] Hand, P. and Voroninski, V. (2016). An elementary proof of convex phase retrieval in the natural parameter space via the linear program phasemax. *arXiv preprint arXiv:1611.03935*.

[83] Hassibi, B. and Vikalo, H. (2005). On the sphere-decoding algorithm i. expected complexity. *Signal Processing, IEEE Transactions on*, 53(8):2806–2818.

[84] Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2019). Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*.

[85] Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

[86] Jaganathan, K., Eldar, Y. C., and Hassibi, B. (2015). Phase retrieval: An overview of recent developments. *arXiv preprint arXiv:1510.07713*.

[87] Jaganathan, K., Oymak, S., and Hassibi, B. (2012). Recovery of sparse 1-d signals from the magnitudes of their fourier transform. In *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium On*, pages 1473–1477. IEEE.

[88] Jaganathan, K., Oymak, S., and Hassibi, B. (2013). Sparse phase retrieval: Convex algorithms and limitations. In *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, pages 1022–1026. IEEE.

[89] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

[90] Jeon, C., Ghods, R., Maleki, A., and Studer, C. (2015). Optimality of large mimo detection via approximate message passing. In *Information Theory (ISIT), 2015 IEEE International Symposium on*, pages 1227–1231. IEEE.

[91] Jeon, C., Maleki, A., and Studer, C. (2016). On the performance of mismatched data detection in large mimo systems. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pages 180–184. IEEE.

[92] Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, pages 295–327.

[93] Johnstone, I. M. and Lu, A. Y. (2004). Sparse principal components analysis. *Unpublished manuscript*, 7.

[94] Kakade, S., Shamir, O., Sindharan, K., and Tewari, A. (2010). Learning exponential families in high-dimensions: Strong convexity and sparsity. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 381–388.

[95] Karoui, N. E. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445*.

[96] King, G. and Zeng, L. (2001). Logistic regression in rare events data. *Political analysis*, 9(2):137–163.

[97] Kini, G. and Thrampoulidis, C. (2020). Analytic study of double descent in binary classification: The impact of loss. *arXiv preprint arXiv:2001.11572*.

[98] Koh, K., Kim, S.-J., and Boyd, S. (2007). An interior-point method for large-scale l1-regularized logistic regression. *Journal of Machine learning research*, 8(Jul):1519–1555.

[99] Krishnapuram, B., Carin, L., Figueiredo, M. A., and Hartemink, A. J. (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE transactions on pattern analysis and machine intelligence*, 27(6):957–968.

[100] Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of statistics*, 37(6B):4254.

[101] Lecué, G. and Mendelson, S. (2014). Sparse recovery under weak moment assumptions. *arXiv preprint arXiv:1401.2188*.

[102] Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411.

[103] Lehmann, E. L. and Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

[104] Li, X. and Voroninski, V. (2013). Sparse signal recovery from quadratic measurements via convex programming. *SIAM Journal on Mathematical Analysis*, 45(5):3019–3033.

[105] Li, Y., Sun, Y., and Chi, Y. (2016). Low-rank positive semidefinite matrix recovery from corrupted rank-one measurements. *IEEE Transactions on Signal Processing*, 65(2):397–408.

[106] Li, Y.-H., Hsieh, Y.-P., Zerbib, N., and Cevher, V. (2015). A geometric view on constrained m-estimators. *arXiv preprint arXiv:1506.08163*.

[107] Liese, F. and Miescke, K.-J. (2008). *Statistical decision theory: estimation, testing, and selection*. Springer Science & Business Media.

[108] Ma, W.-K., Davidson, T. N., Wong, K. M., Luo, Z.-Q., and Ching, P.-C. (2002). Quasi-maximum-likelihood multiuser detection using semi-definite relaxation with application to synchronous cdma. *IEEE transactions on signal processing*, 50(4):912–922.

[109] Madiman, M. and Barron, A. (2007). Generalized entropy power inequalities and monotonicity properties of information. *Information Theory, IEEE Transactions on*, 53(7):2317–2329.

[110] Mangasarian, O. L. and Recht, B. (2011). Probability of unique integer solution to a system of linear equations. *European Journal of Operational Research*, 214(1):27–30.

[111] Mazumder, R. and Hastie, T. (2012). The graphical lasso: New insights and alternatives. *Electronic journal of statistics*, 6:2125.

[112] McCoy, M. B., Cevher, V., Dinh, Q. T., Asaei, A., and Baldassarre, L. (2014). Convexity in source separation: Models, geometry, and algorithms. *Signal Processing Magazine, IEEE*, 31(3):87–95.

[113] McCoy, M. B. and Tropp, J. A. (2014). Sharp recovery bounds for convex demixing, with applications. *Foundations of Computational Mathematics*, 14(3):503–567.

[114] Mei, S. and Montanari, A. (2019). The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*.

[115] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, pages 1436–1462.

[116] Millane, R. P. (1990). Phase retrieval in crystallography and optics. *JOSA A*, 7(3):394–411.

[117] Miolane, L. and Montanari, A. (2018). The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *arXiv preprint arXiv:1811.01212*.

[118] Montanari, A., Ruan, F., Sohn, Y., and Yan, J. (2019). The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*.

[119] Mousavi, A., Maleki, A., and Baraniuk, R. G. (2013). Parameterless optimal approximate message passing. *arXiv preprint arXiv:1311.0035*.

[120] Muthukrishnan, S. et al. (2005). Data streams: Algorithms and applications. *Foundations and Trends® in Theoretical Computer Science*, 1(2):117–236.

[121] Narasimhan, T. L. and Chockalingam, A. (2014). Channel hardening-exploiting message passing (chemp) receiver in large-scale mimo systems. *Selected Topics in Signal Processing, IEEE Journal of*, 8(5):847–860.

[122] Needell, D. and Tropp, J. A. (2009). Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and computational harmonic analysis*, 26(3):301–321.

[123] Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557.

[124] Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.

[125] Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245.

[126] Ngo, H. Q., Larsson, E. G., and Marzetta, T. L. (2013). Energy and spectral efficiency of very large multiuser mimo systems. *Communications, IEEE Transactions on*, 61(4):1436–1449.

[127] Oymak, S. and Hassibi, B. (2010). New null space results and recovery thresholds for matrix rank minimization. *arXiv preprint arXiv:1011.6326*.

[128] Oymak, S., Jalali, A., Fazel, M., Eldar, Y. C., and Hassibi, B. (2015). Simultaneously structured models with application to sparse and low-rank matrices. *IEEE Transactions on Information Theory*, 61(5):2886–2908.

[129] Oymak, S., Thrampoulidis, C., and Hassibi, B. (2013). The squared-error of generalized lasso: A precise analysis. *arXiv preprint arXiv:1311.0830*.

[130] Oymak, S. and Tropp, J. A. (2017). Universality laws for randomized dimension reduction, with applications. *Information and Inference: A Journal of the IMA*, 7(3):337–446.

[131] Panahi, A. and Hassibi, B. (2017). A universal analysis of large-scale regularized least squares solutions. In *Advances in Neural Information Processing Systems*, pages 3381–3390.

[132] Parikh, N., Boyd, S., et al. (2014). Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239.

[133] Plan, Y. and Vershynin, R. (2015). The generalized lasso with non-linear observations. *arXiv preprint arXiv:1502.04071*.

[134] Ravikumar, P., Wainwright, M. J., Raskutti, G., Yu, B., et al. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980.

[135] Recht, B., Fazel, M., and Parrilo, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501.

[136] Rockafellar, R. T. (2015). *Convex analysis*. Princeton university press.

[137] Rockafellar, R. T. and Wets, R. J.-B. (2009). *Variational analysis*, volume 317. Springer Science & Business Media.

[138] Rothman, A. J., Bickel, P. J., Levina, E., Zhu, J., et al. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515.

[139] Rudelson, M. and Vershynin, R. (2006). Sparse reconstruction by convex relaxation: Fourier and gaussian measurements. In *2006 40th Annual Conference on Information Sciences and Systems*, pages 207–212. IEEE.

[140] Rusek, F., Persson, D., Lau, B. K., Larsson, E. G., Marzetta, T. L., Edfors, O., and Tufvesson, F. (2013). Scaling up mimo: Opportunities and challenges with very large arrays. *IEEE signal processing magazine*, 30(1):40–60.

[141] Salehi, F., Abbasi, E., and Hassibi, B. (2018a). Learning without the phase: Regularized phasemax achieves optimal sample complexity. In *Advances in Neural Information Processing Systems*, pages 8655–8666.

[142] Salehi, F., Abbasi, E., and Hassibi, B. (2018b). A precise analysis of phasemax in phase retrieval. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 976–980. IEEE.

[143] Salehi, F., Abbasi, E., and Hassibi, B. (2019). The impact of regularization on high-dimensional logistic regression. In *Advances in Neural Information Processing Systems*, pages 11982–11992.

[144] Salehi, F., Jaganathan, K., and Hassibi, B. (2017). Multiple illumination phaseless super-resolution (mips) with applications to phaseless super doa estimation and diffraction imaging. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 3949–3953. IEEE.

[145] Scharf, L. L. and Friedlander, B. (1994). Matched subspace detectors. *IEEE Transactions on signal processing*, 42(8):2146–2157.

[146] Serdobolskii, V. (2013). *Multivariate statistical analysis: A high-dimensional approach*, volume 41. Springer Science & Business Media.

[147] Shechtman, Y., Beck, A., and Eldar, Y. C. (2014). Gespar: Efficient phase retrieval of sparse signals. *IEEE transactions on signal processing*, 62(4):928–938.

[148] Shechtman, Y., Eldar, Y. C., Szameit, A., and Segev, M. (2011). Sparsity based sub-wavelength imaging with partially incoherent light via quadratic compressed sensing. *Optics express*, 19(16):14807–14822.

[149] Sion, M. et al. (1958). On general minimax theorems. *Pacific Journal of mathematics*, 8(1):171–176.

[150] Sivakumar, V., Banerjee, A., and Ravikumar, P. K. (2015). Beyond sub-gaussian measurements: High-dimensional structured estimation with sub-exponential designs. In *Advances in Neural Information Processing Systems*, pages 2197–2205.

[151] Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. (2018). The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878.

[152] Stojnic, M. (2009a). Block-length dependent thresholds in block-sparse compressed sensing. *arXiv preprint arXiv:0907.3679*.

[153] Stojnic, M. (2009b). Various thresholds for ℓ_1-optimization in compressed sensing. *arXiv preprint arXiv:0907.3666*.

[154] Stojnic, M. (2013a). A framework to characterize performance of lasso algorithms. *arXiv preprint arXiv:1303.7291*.

[155] Stojnic, M. (2013b). A framework to characterize performance of lasso algorithms. *arXiv preprint arXiv:1303.7291*.

[156] Stojnic, M. (2013c). Upper-bounding l1-optimization weak thresholds. *arXiv preprint arXiv:1303.7289*.

[157] Sur, P. and Candès, E. J. (2018). A modern maximum-likelihood theory for high-dimensional logistic regression. *arXiv preprint arXiv:1803.06964*.

[158] Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525.

[159] Sur, P., Chen, Y., and Candès, E. J. (2017). The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability Theory and Related Fields*, pages 1–72.

[160] Taeb, A., Maleki, A., Studer, C., and Baraniuk, R. (2013). Maximin analysis of message passing algorithms for recovering block sparse signals. *arXiv preprint arXiv:1303.2389*.

[161] Taheri, H., Pedarsani, R., and Thrampoulidis, C. (2019). Sharp guarantees for solving random equations with one-bit information. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 765–772. IEEE.

[162] Tan, P. H., Rasmussen, L. K., and Lim, T. J. (2001). Constrained maximum-likelihood detection in cdma. *IEEE Transactions on Communications*, 49(1):142–153.

[163] Tanaka, T. (2002). A statistical-mechanics approach to large-system analysis of cdma multiuser detectors. *IEEE Transactions on Information theory*, 48(11):2888–2910.

[164] Thrampoulidis, C., Abbasi, E., and Hassibi, B. (2015a). Lasso with non-linear measurements is equivalent to one with linear measurements. In *Advances in Neural Information Processing Systems*, pages 3420–3428.

[165] Thrampoulidis, C., Abbasi, E., and Hassibi, B. (2016a). Precise error analysis of regularized M-estimators in high-dimensions. *arXiv preprint arXiV: 1601.06233*.

[166] Thrampoulidis, C., Abbasi, E., Xu, W., and Hassibi, B. (2016b). Ber analysis of the box relaxation for bpsk signal recovery. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 3776–3780. IEEE.

[167] Thrampoulidis, C. and Hassibi, B. (2014). Estimating structured signals in sparse noise: A precise noise sensitivity analysis. In *Communication, Control, and Computing (Allerton), 2014 52nd Annual Allerton Conference on*, pages 866–873. IEEE.

[168] Thrampoulidis, C. and Hassibi, B. (2015). Isotropically random orthogonal matrices: Performance of lasso and minimum conic singular values. In *Information Theory (ISIT), 2015 IEEE International Symposium on*. IEEE.

[169] Thrampoulidis, C., Oymak, S., and Hassibi, B. (2014). Simple error bounds for regularized noisy linear inverse problems. *Information Theory, 2014. f 2014. Proceedings. International Symposium on*, pages 3007–3011.

[170] Thrampoulidis, C., Oymak, S., and Hassibi, B. (2015b). Recovering structured signals in noise: Least-squares meets compressed sensing. In *Compressed Sensing and Its Applications*, pages 97–141. Springer.

[171] Thrampoulidis, C., Oymak, S., and Hassibi, B. (2015c). Regularized linear regression: A precise analysis of the estimation error. In *Proceedings of The 28th Conference on Learning Theory*, pages 1683–1709.

[172] Thrampoulidis, C., Panahi, A., Guo, D., and Hassibi, B. (2015d). Precise error analysis of the lasso. In *40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2015, arxivPreprint arXiv:1502.04977*.

[173] Thrampoulidis, C., Panahi, A., and Hassibi, B. (2015e). Asymptotically exact error analysis for the generalized ℓ_2^2-lasso. In *Information Theory (ISIT), 2015 IEEE International Symposium on*. IEEE.

[174] Thrampoulidis, C., Xu, W., and Hassibi, B. (2018). Symbol error rate performance of box-relaxation decoders in massive mimo. *IEEE Transactions on Signal Processing*, 66(13):3377–3392.

[175] Thrampoulidis, C., Zadik, I., and Polyanskiy, Y. (2019). A simple bound on the ber of the map decoder for massive mimo systems. *arXiv preprint arXiv:1903.03949*.

[176] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

[177] Tropp, J. A. (2006). Just relax: Convex programming methods for identifying sparse signals in noise. *Information Theory, IEEE Transactions on*, 52(3):1030–1051.

[178] Tropp, J. A. (2014). Convex recovery of a structured signal from independent random linear measurements. *arXiv preprint arXiv:1405.1102*.

[179] Tropp, J. A. and Gilbert, A. C. (2007). Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on information theory*, 53(12):4655–4666.

[180] Tropp, J. A., Laska, J. N., Duarte, M. F., Romberg, J. K., and Baraniuk, R. G. (2009). Beyond nyquist: Efficient sampling of sparse bandlimited signals. *arXiv preprint arXiv:0902.0026*.

[181] Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of clinical epidemiology*, 49(11):1225–1231.

[182] Van de Geer, S. A. et al. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645.

[183] Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

[184] Vapnik, V. (1982). Estimation of dependences based on empirical data berlin.

[185] Verdu, S. (1998). *Multiuser detection*. Cambridge university press.

[186] Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

[187] Vershynin, R. (2012). How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 25(3):655–686.

[188] Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). *Information Theory, IEEE Transactions on*, 55(5):2183–2202.

[189] Wainwright, M. J. (2014). Structured regularizers for high-dimensional problems: Statistical and computational issues. *Annual Review of Statistics and Its Application*, 1:233–253.

[190] Waldspurger, I., d'Aspremont, A., and Mallat, S. (2015). Phase recovery, maxcut and complex semidefinite programming. *Mathematical Programming*, 149(1-2):47–81.

[191] Walther, A. (1963). The question of phase retrieval in optics. *Journal of Modern Optics*, 10(1):41–49.

[192] Wang, L. (2013). The l1 penalized lad estimator for high dimensional linear regression. *Journal of Multivariate Analysis*, 120:135–151.

[193] Wen, C.-K., Chen, J.-C., Wong, K.-K., and Ting, P. (2014). Message passing algorithm for distributed downlink regularized zero-forcing beamforming with cooperative base stations. *Wireless Communications, IEEE Transactions on*, 13(5):2920–2930.

[194] White, C. D., Sanghavi, S., and Ward, R. (2015). The local convexity of solving systems of quadratic equations. *arXiv preprint arXiv:1506.07868*.

[195] Wu, W. B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90(4):831–844.

[196] Wu, Y. and Verdú, S. (2012). Optimal phase transitions in compressed sensing. *IEEE Transactions on Information Theory*, 58(10):6241–6263.

[197] Xu, J. and Hsu, D. (2019). How many variables should be entered in a principal component regression equation? *arXiv preprint arXiv:1906.01139*.

[198] Yener, A., Yates, R. D., and Ulukus, S. (2002). Cdma multiuser detection: A nonlinear programming approach. *IEEE Transactions on Communications*, 50(6):1016–1024.

[199] Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35.

.1 Proof of Theorem 2

Here, we prove Theorem 2. The proof consists of several steps and intermediate results, that are stated as lemmas. The proofs of those are all deferred to Appendix .2.

Preliminaries

$$\hat{\beta} := \arg\min_{\beta} \mathcal{L}(\mathbf{y} - \mathbf{X}\beta) + \lambda f(\beta).$$

Recall that $\mathbf{y} = \mathbf{X}\beta_0 + \mathbf{z}$. Our goal is to characterize the nontrivial limiting behavior of $\|\hat{\beta} - \beta_0\|_2/\sqrt{p}$. We start with a simple change of variables $\mathbf{w} := (\beta - \beta_0)/\sqrt{p}$, to directly get a handle on the *error vector* $\mathbf{w}$. Also, we normalize the objective by dividing with p so that the optimal cost is of constant order. Then,

$$\hat{\mathbf{w}} := \arg\min_{\mathbf{w}} \frac{1}{p} \left\{ \mathcal{L}\left(\mathbf{z} - \sqrt{p}\mathbf{X}\mathbf{w}\right) + \lambda f\left(\beta_0 + \sqrt{p}\mathbf{w}\right) \right\}. \tag{1}$$

Instead of the optimization problem above, we will analyze a simpler Auxiliary Optimization (AO) that is tightly related to the Primary Optimization (PO) in (1) via the CGMT.

The CGMT for M-estimators

In this section, we show how the CGMT Theorem 8 can be applied to predict the limiting behavior of the solution $\|\hat{\mathbf{w}}\|_2$ to the minimization in (1). The main challenge here is to express (1) as a (convex-concave) minimax optimization in which the involved random matrix (here $\mathbf{X}$) appears in a bilinear form, exactly as in (2.126a). Also, some side technical details need to be taken care of. For example, in (2.126a) the optimization constraints are required by Theorem 8 to be bounded, which is not the case with (1). We start with addressing this immediately next.

Boundedness of the Error

The constraint set over which $\mathbf{w}$ is optimized in (2.126a) is unbounded. We will introduce “artificial" boundedness constraints that allow applying Theorem 8, while they do not affect the optimization itself. For this purpose, recall our goal of proving that $\|\hat{\mathbf{w}}\|_2$ converges to some (finite) α_* defined in Theorem 2. Define the set $\mathcal{S}_{\mathbf{w}} = \{\mathbf{w} \mid \|\mathbf{w}\|_2 \leq K_\alpha\}$, where

$$K_\alpha := \alpha_* + \zeta \tag{2}$$

for a constant $\zeta > 0$, and, consider the “bounded" version of (1):

$$\hat{\mathbf{w}}^B := \arg\min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}} \frac{1}{p} \left\{ \mathcal{L}\left(\mathbf{z} - \sqrt{p}\mathbf{X}\mathbf{w}\right) + \lambda f\left(\beta_0 + \sqrt{p}\mathbf{w}\right) \right\}. \tag{3}$$

We expect that the additional constraint $\mathbf{w} \in \mathcal{S}_{\mathbf{w}}$ in (3) will not affect the optimization with high probability when p is large enough. The idea here is that the minimizer of the original unconstrained problem in (1) satisfies $\|\hat{\mathbf{w}}\|_2 \approx \alpha_* < K_\alpha$ w.h.p.. Of course, this latter statement is yet to be proven; but onnce this is done, we can return and confirm that our initial expectation is met. Lemma 26 below shows that if $\|\hat{\mathbf{w}}^B\| \xrightarrow{P} \alpha_* < K_\alpha$, then, the same is true for the optimal of (1).

Lemma 26 *For the two optimizations in* (1) *and* (3)*, let* $\hat{\mathbf{w}}$ *and* $\hat{\mathbf{w}}^B$ *be optimal solutions. Also, recall the definition of* K_α *in* (2)*. If* $\|\hat{\mathbf{w}}^B\|_2 \xrightarrow{P} \alpha_*$*, then* $\|\hat{\mathbf{w}}\|_2 \xrightarrow{P} \alpha_*$*.*

Owing to the result of the lemma, henceforth, we work with the bounded optimization in (3). Using some abuse of notation, we will refer to optimal solution of (3) as $\hat{\mathbf{w}}$, rather than $\hat{\mathbf{w}}^B$.

Identifying the (PO)

Here, we bring the minimization in (3) it in the form of the (PO) in (2.126a). For this purpose, we will use Lagrange duality. Note that the former can be equivalently expressed as

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}},\mathbf{v}} \frac{1}{p}\left\{\mathcal{L}(\sqrt{p}\mathbf{v}) + \lambda f(\beta_0 + \sqrt{p}\mathbf{w})\right\} \quad \text{subject to} \quad \mathbf{v} = \mathbf{z} - \sqrt{p}\mathbf{X}\mathbf{w}.$$

Associating a dual variable $\mathbf{u}$ to the equality constraint above, we write it as

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}},\mathbf{v}} \max_{\mathbf{u}} \frac{1}{\sqrt{p}}\left\{-\mathbf{u}^T(\sqrt{p}\mathbf{X})\mathbf{w} + \mathbf{u}^T\mathbf{z} - \mathbf{u}^T\mathbf{v}\right\} + \frac{1}{p}\left\{\mathcal{L}(\mathbf{v}) + \lambda f(\beta_0 + \sqrt{p}\mathbf{w})\right\}. \tag{4}$$

It takes no much effort to check that the objective function above is in the desired format of (2.126a): the random matrix $\mathbf{X}$ appears in a bilinear term $\mathbf{u}^T\mathbf{X}\mathbf{w}$, and, the rest of the terms form a convex-concave function in $\mathbf{u}, \mathbf{w}$. Furthermore, we can use Assumption 1 to show that the optimal $\mathbf{u}_*$ is bounded, which is a requirement of Theorem 8. In the same lines as in Section .1, we henceforth work with the "bounded" version of (4), namely,

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}},\mathbf{v}} \max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}} \frac{1}{\sqrt{p}}\left\{-\mathbf{u}^T(\sqrt{p}\mathbf{X})\mathbf{w} + \mathbf{u}^T\mathbf{z} - \mathbf{u}^T\mathbf{v}\right\} + \frac{1}{p}\left\{\mathcal{L}(\mathbf{v}) + \lambda f(\beta_0 + \sqrt{p}\mathbf{w})\right\}. \tag{5}$$

for $\mathcal{S}_{\mathbf{u}} := \{\mathbf{u} \mid \|\mathbf{u}\|_2 \leq K_\beta\}$ and $K_\beta > 0$ a sufficiently large constant.

Lemma 27 *If Assumption 1(b) holds, then there exists sufficiently large constant K_β, such that the optimization problem in* (5) *is equivalent to that in* (3)*, with probability approaching 1 in the limit of $p \to \infty$.*

As a last step, before writing down the corresponding (AO) problem, it will be useful for the analysis of the latter, to express f in a variational form through its Fenchel conjugate, which gives,

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}\in\mathcal{S}_\mathbf{w},\mathbf{v}} \max_{\mathbf{u}\in\mathcal{S}_\mathbf{u},\mathbf{s}} \frac{1}{\sqrt{p}} \left\{-\mathbf{u}^T(\sqrt{p}\mathbf{X})\mathbf{w} + \mathbf{u}^T\mathbf{z} - \mathbf{u}^T\mathbf{v}\right\} + \frac{1}{p}\left\{\mathcal{L}(\mathbf{v}) + \lambda\mathbf{s}^T\beta_0 + \lambda\sqrt{p}\mathbf{s}^T\mathbf{w} - \lambda f^*(\mathbf{s})\right\}. \tag{6}$$

The (AO)

Having identified (6) as the (PO) in our application, it is straightforward to write the corresponding (AO) problem following (2.126b):

$$\min_{\mathbf{w}\in\mathcal{S}_\mathbf{w},\mathbf{v}} \max_{\mathbf{u}\in\mathcal{S}_\mathbf{u},\mathbf{s}} \frac{1}{\sqrt{p}} \left\{\|\mathbf{w}\|_2\mathbf{g}^T\mathbf{u} - \|\mathbf{u}\|_2\mathbf{h}^T\mathbf{w} + \mathbf{u}^T\mathbf{z} - \mathbf{u}^T\mathbf{v}\right\} + \frac{1}{p}\left\{\mathcal{L}(\mathbf{v}) + \lambda\mathbf{s}^T\beta_0 + \lambda\sqrt{p}\mathbf{s}^T\mathbf{w} - \lambda f^*(\mathbf{s})\right\}. \tag{7}$$

Once we have identified the (AO) problem, Corollary 3 suggests analyzing that one instead of the (PO). Our goal is showing that $\|\hat{\mathbf{w}}\|_2 \xrightarrow{P} \alpha_*$. For this, we wish to apply the corollary to the following set

$$\mathcal{S} = \{\mathbf{w} \mid |\|\mathbf{w}\|_2 - \alpha_*| > \epsilon\},$$

for arbitrary $\epsilon > 0$.

Asymptotic min-max property of the (AO)

It turns out that verifying the conditions of the corollary for the (AO) as it appears in (7) is not directly easy. In short, what makes the analysis cumbersome is the fact that the optimization in (7) is not convex (e.g. if $\mathbf{g}^T\mathbf{u}$ is negative, then $\|\mathbf{w}\|_2\mathbf{g}^T\mathbf{u}$ is not convex). Thus, flipping the order of min-max operations that would simplify the analysis is not directly justified.

At this point, recall that the (PO) in (6) is itself convex. In fact, for it, all conditions of Sion's min-max Theorem [149] are met, thus, the order of min-max operations can be flipped. According to the CGMT, the (PO) and the (AO) are tightly related in

an asymptotic setting. We use this, to translate the convexity properties of the (PO) to the (AO). In essence, we show that when dimensions grow, the order of min-max operations in the (AO) can be flipped. Thus, we will instead consider the following problem as the (AO):

$$\phi(\mathbf{g},\mathbf{h}) := \max_{\substack{0\le\beta\le K_\beta \\ \mathbf{s}}} \min_{\substack{\|\mathbf{w}\|_2\le K_\alpha \\ \mathbf{v}}} \max_{\|\mathbf{u}\|_2=\beta} \frac{1}{\sqrt{p}}(\|\mathbf{w}\|_2\mathbf{g}+\mathbf{z}-\mathbf{v})^T\mathbf{u} - \frac{1}{\sqrt{p}}\|\mathbf{u}\|_2\mathbf{h}^T\mathbf{w} + \frac{1}{p}\mathcal{L}(\mathbf{v}) + \frac{\lambda}{p}\mathbf{s}^T\beta_0 + \frac{\lambda}{\sqrt{p}}\mathbf{s}^T\mathbf{w} - \frac{\lambda}{p}f^*(\mathbf{s}). \tag{8}$$

Observe that the objective function remains the same; it is only the order of min-max operations that is slightly modified compared to (7). Since the objective function is not necessarily convex-concave in its arguments, there is no immediate guarantee that the two problems in (7) and (8) are equivalent for all realizations of $\mathbf{g}$ and $\mathbf{h}$. However, the lemma below essentially shows that such a strong duality holds with high probability over $\mathbf{g}$ and $\mathbf{h}$ in high dimensions. Hence, the problem in (8) can be as well used, instead of the one in (7), in order to analyze the (PO). For this reason, henceforth, we refer to (8) as the (AO) problem.

Lemma 28 *Let $\hat{\mathbf{w}}(\mathbf{X})$ denote an optimal solution of* (1)*. Consider the (AO) problem in* (8)*. Let α_* be as defined in Theorem 2. For any $\epsilon > 0$ define the set $\mathcal{S} := \{\mathbf{w} \mid |\|\mathbf{w}\|_2 - \alpha_*| < \epsilon\}$, and, $\phi_{\mathcal{S}^c}(\mathbf{g},\mathbf{h})$ be the optimal cost of the same optimization as in* (8)*, only this time the minimization over $\mathbf{w}$ is further constrained such that $\mathbf{w} \notin \mathcal{S}$. Assume that for any $K_\alpha > \alpha_*$ and for any sufficiently large K_β, there exist constants $\overline{\phi} < \overline{\phi}_{\mathcal{S}^c}$ such that for all $\eta > 0$, with probability approaching one in the limit of $p \to \infty$ the following hold:*

1. $\phi(\mathbf{g},\mathbf{h}) < \overline{\phi} + \eta$,
2. $\phi_{\mathcal{S}^c}(\mathbf{g},\mathbf{h}) > \overline{\phi}_{\mathcal{S}^c} - \eta$.

Then,

$$\lim_{p\to\infty} \mathbb{P}\left(|\|\hat{\mathbf{w}}(\mathbf{X})\|_2 - \alpha_*| < \epsilon \right) = 1.$$

In view of Lemma 28, what remains in order to prove Theorem 2 is satisfying the conditions of the lemma. This involves a thorough analysis of the (AO) problem in (8), which is the subject of the next few sections.

Scalarization

Observe that the optimization in (8) is over vectors. The purpose of this section is to simplify the (AO) into an optimization involving only scalar variables. Of course, one of this has to play the role of the norm of $\mathbf{w}$, which is the quantity of interest. The main idea behind the "scalarization" step of the (AO) is to perform the optimization over only the direction of the vector variables while keeping their magnitude constant. This is already hinted by the rearrangement of the order of min-max operations going from (7) to (8). Also, this process is facilitated by the following two:

1. The bilinear term $\mathbf{u}^T\mathbf{X}\mathbf{w}$ that appears in the (PO) conveniently "splits" into the two terms $\|\mathbf{w}\|_2\mathbf{g}^T\mathbf{u}$ and $\|\mathbf{u}\|_2\mathbf{h}^T\mathbf{w}$ in the (AO),
2. The term involving the regularizer, i.e. $f(\beta_0 + \mathbf{w})$ has been expressed in a variational form as $\sup_{\mathbf{s}} \mathbf{s}^T\beta_0 + \mathbf{s}^T\mathbf{w} - f^*(\mathbf{s})$.

The details of the reduction step are all summarized in Lemma 29 below which shows that the (AO) reduces to the following *convex* minimax problem on four *scalar* optimization variables:

$$\inf_{\substack{0\le\alpha\le K_\alpha \\ \tau_g>0}} \sup_{\substack{0\le\beta\le K_\beta \\ \tau_h>0}} \frac{\beta\tau_g}{2} + \frac{1}{p}\mathrm{e}_{\mathcal{L}}\left(\alpha\mathbf{g}+\mathbf{z};\frac{\tau_g}{\beta}\right) - \begin{cases} \frac{\alpha\tau_h}{2} + \frac{\beta^2\alpha}{2\tau_h}\frac{\|\mathbf{h}\|^2}{p} - \lambda\cdot\frac{1}{p}\mathrm{e}_f\left(\frac{\beta\alpha}{\tau_h}\mathbf{h}+\beta_0;\frac{\alpha\lambda}{\tau_h}\right) & ,\alpha>0 \\ \frac{\lambda}{p}f(\beta_0) & ,\alpha=0 \end{cases}, \tag{9}$$

where recall that

$$\mathrm{e}_\omega(\mathbf{u};\tau) := \min_{\mathbf{v}}\{\frac{1}{2\tau}\|\mathbf{u}-\mathbf{v}\|_2^2 + \omega(\mathbf{v})\}$$

denotes the (vector) τ-Moreau envelope of a function $\omega : \mathbb{R}^d \to \mathbb{R}$ evaluated at $\mathbf{u} \in \mathbb{R}^d$.

Lemma 29 (Scalarization of the (AO)) *The following statements are true regarding the two minimax optimization problems in* (8) *and* (9)*:*

1. *They have the same optimal cost.*
2. *The objective function in* (9) *is continuous on its domain, (jointly) convex in* (α, τ_g) *and (jointly) concave in* (β, τ_h).
3. *The order of inf-sup in* (9) *can be flipped without changing the optimization.*

Convergence Analysis

The goal of this section is to show that the (AO) satisfies the conditions of Lemma 28. This requires a convergence analysis of its optimal cost. We work with the scalarized version of the (AO) that was derived in the previous section:

$$\phi(\mathbf{g},\mathbf{h},\mathbf{z},\beta_0) = \inf_{\substack{0\le\alpha\le K_\alpha\\ \tau_g>0}} \sup_{\substack{0\le\beta\le K_\beta\\ \tau_h>0}} \mathcal{R}_p(\alpha,\tau_g,\beta,\tau_h;\mathbf{g},\mathbf{h},\mathbf{z},\beta_0), \tag{10}$$

$$\mathcal{R}_p = \frac{\beta\tau_g}{2} + \frac{1}{p}\left\{\mathrm{e}_{\mathcal{L}}\left(\alpha\mathbf{g}+\mathbf{z};\frac{\tau_g}{\beta}\right) - \mathcal{L}(\mathbf{z})\right\} - \begin{cases} \frac{\alpha\tau_h}{2} + \frac{\beta^2\alpha}{2\tau_h}\frac{\|\mathbf{h}\|^2}{p} - \frac{\lambda}{p}\left\{\mathrm{e}_f\left(\frac{\beta\alpha}{\tau_h}\mathbf{h}+\beta_0;\frac{\alpha\lambda}{\tau_h}\right) - f(\beta_0)\right\} & ,\alpha>0 \\ 0 & ,\alpha=0 \end{cases}.$$

Here, when compared to (9), we have subtracted from the objective the terms $\mathcal{L}(\mathbf{z})$ and $f(\beta_0)$, which of course does not affect the optimization. The optimization is of course random over the realizations of $\mathbf{g},\mathbf{h},\mathbf{z}$ and β_0, and, by the WLLN, it is easy to identify the converging value of the objective function $\mathcal{R}_p$ for fixed parameter values $\alpha,\tau_g,\beta,\tau_h$. For our goals, we need to show that minimax of the converging sequence of objectives converges to the minimax of the objective of the (SOP). Convexity of $\mathcal{R}_p$ plays a crucial role here since is being use to conclude local uniform convergence from the pointwise convergence. Uniform convergence is a requirement to conclude the desired.[1]

Lemma 30 (Convergence properties of the (AO)) *Let*

$$\mathcal{R}_p(\alpha,\tau_g,\beta,\tau_h) := \mathcal{R}_p(\alpha,\tau_g,\beta,\tau_h;\mathbf{g},\mathbf{h},\mathbf{z},\beta_0)$$

be defined as in (10)*, and,*

$$\phi_{\mathcal{A}} := \phi_{\mathcal{A}}(\mathbf{g},\mathbf{h},\mathbf{z},\beta_0) := \inf_{\substack{\alpha\in\mathcal{A}\\ \tau_g>0}} \sup_{\substack{0\le\beta\le K_\beta\\ \tau_h>0}} \mathcal{R}_p(\alpha,\tau_g,\beta,\tau_h), \tag{11}$$

for $\mathcal{A}\subseteq[0,\infty)$*. Further consider the following deterministic convex program*

$$\overline{\phi}_{\mathcal{A}} := \inf_{\substack{\alpha\in\mathcal{A}\\ \tau_g>0}} \sup_{\substack{\beta\ge0\\ \tau_h>0}} \mathcal{D}(\alpha,\tau_g,\beta,\tau_h) := \begin{cases} \frac{\beta\tau_g}{2} + \delta\cdot L\left(\alpha,\frac{\tau_g}{\beta}\right) & ,\beta>0 \\ -\delta\cdot L_0 & ,\beta=0 \end{cases} - \begin{cases} \frac{\alpha\tau_h}{2} + \frac{\alpha\beta^2}{2\tau_h} - \lambda\cdot F\left(\frac{\alpha\beta}{\tau_h},\frac{\alpha\lambda}{\tau_h}\right) & ,\alpha>0 \\ 0 & ,\alpha=0 \end{cases}. \tag{12}$$

[1] Interestingly, some of the tools used for this part of the proof are similar to those classically used for the study of consistency of M-estimators in the classical regime where p is fixed and n goes to infinity, see for example the Arg-min theorem in [107, Thm. 7.70], [125, Thm. 2.7].

where L and F as in Theorem 2. If Assumption 1 hold, then,

1. $\mathcal{R}_n(\alpha, \tau_g, \beta, \tau_h) \xrightarrow{P} \mathcal{D}(\alpha, \tau_g, \beta, \tau_h)$*, for all* $(\alpha, \tau_g, \beta, \tau_h)$*, and,* $\mathcal{D}(\alpha, \tau_g, \beta, \tau_h)$ *is convex in* (α, τ_g) *and concave in* (β, τ_h)*.*

2. *Assume* α_* *is the unique minimizer in* (12) *with* $\mathcal{A} := [0, \infty)$*. For any* $\epsilon > 0$*, define* $\mathcal{S}_\epsilon := \{\alpha \mid |\alpha - \alpha_*| < \epsilon\}$*. Then, for any sufficiently large constants* $K_\alpha > \alpha_*$ *and* $K_\beta > 0$*, and for all* $\eta > 0$*, it holds with probability approaching 1 as* $p \to \infty$*:*

   a) $\phi_{[0,K_\alpha]} < \overline{\phi}_{[0,\infty)} + \eta$*,*

   b) $\phi_{[0,K_\alpha]\setminus\mathcal{S}_\epsilon} \geq \overline{\phi}_{[0,\infty)\setminus\mathcal{S}_\epsilon} - \eta$*,*

   c) $\overline{\phi}_{[0,\infty)\setminus\mathcal{S}_\epsilon} > \overline{\phi}_{[0,\infty)}$*.*

Putting all the Pieces Together

We are now ready to conclude the proof of Theorem 2.

Proof 16 (Proof of Theorem 2) *Fix any* $\epsilon > 0$*. Consider the set* $\mathcal{S}_\epsilon = \{\mathbf{w} \mid |\|\mathbf{w}\|_2 - \alpha_*\|_2 < \epsilon$ *as in Lemma 28. We use the same notation as in the lemma. Let* $K_\alpha > \alpha_*$ *and arbitrarily large (but finite)* $K_\beta > 0$*. From Lemma 29(i)* $\phi(\mathbf{g}, \mathbf{h})$ *is equal to the optimal cost of the optimization in* (9)*. But, from Lemma 30(b)(i), the latter converges in probability to some constant* $\overline{\phi}$ *(see Lemma 30 for the exact value constant). The same line of arguments applies to* $\phi_{\mathcal{S}_\epsilon^c}(\mathbf{g}, \mathbf{h})$*, showing that it converges to another constant* $\overline{\phi}_{\mathcal{S}_\epsilon^c}$*. Again from Lemma 30(iii):* $\overline{\phi}_{\mathcal{S}_\epsilon^c} > \overline{\phi}$*. Thus, the conditions of Lemma 28 are satisfied, and, it implies that the magnitude of any optimal minimizer (say)* $\hat{\mathbf{w}}^{(PO)}$ *of the (PO) problem in* (6) *satisfies* $\hat{\mathbf{w}}^{(PO)} \in \mathcal{S}$ *in probability, in the limit of* $p \to \infty$*.*
Once, we have a min-max optimization that consits of only scalars, we only have to investigate the optimality conditions to get to the system of non-linear equations in Theorem 2.

.2 Proofs for Section .1

Proof of Theorem 8(iii)

Consider the following event

$$\mathcal{E} = \{\Phi_{\mathcal{S}^c}(\mathbf{G}) \geq \overline{\phi}_{\mathcal{S}^c} - \eta \ , \ \Phi(\mathbf{G}) \leq \overline{\phi} + \eta\}.$$

In this event, it is not hard to check using assumption (a) that $\Phi_{\mathcal{S}^c} > \Phi$, or equivalently $\mathbf{w}_\Phi \in \mathcal{S}$. Thus, it suffices to show that $\mathcal{E}$ occurs with probability at least $1 - 4p$.

Indeed, from statement (i) of the theorem and assumption (c),

$$\mathbb{P}(\Phi_{\mathcal{S}^c}(\mathbf{G}) < \overline{\phi}_{\mathcal{S}^c} - \eta) \leq 2\mathbb{P}(\phi_{\mathcal{S}^c}(\mathbf{g}, \mathbf{h}) \leq \overline{\phi}_{\mathcal{S}^c} - \eta) \leq 2p.$$

Also, from statement (ii) of the theorem and assumption (b),

$$\mathbb{P}(\Phi(\mathbf{G}) > \overline{\phi} + \eta) \leq 2\mathbb{P}(\phi(\mathbf{g}, \mathbf{h}) \geq \overline{\phi} + \eta) \leq 2p.$$

Combining the above displays the claim follows from a union bound.

Proof of Corollary 3

Call $\eta := (\overline{\phi}_{\mathcal{S}^c} - \overline{\phi})/3 > 0$. By assumption, for any $p > 0$ there exists $N := N(\eta, p)$ such that the events $\{\phi < \overline{\phi} + \eta\}$ and $\{\phi_{\mathcal{S}^c} > \overline{\phi}_{\mathcal{S}^c} - \eta\}$ occur with probability at least $1 - p$ each, for all $p > N$. Then, for all $p > N$, we can apply Theorem 8(iii) to conclude that $\mathbf{w}_\Phi(\mathbf{G}) \in \mathcal{S}$ with probably at least $1 - 4p$. Since this holds for all $p > 0$, the proof is complete.

Proof of Lemma 26

For convenience, denote with $M(\mathbf{w})$ the objective function in (1). For some $\epsilon > 0$ such that $\alpha + \epsilon < K_\alpha$ (e.g. $\epsilon = \zeta/2$ in (2)), denote $\mathcal{D} := \{\mathbf{w} \mid \alpha - \epsilon \leq \|\mathbf{w}\|_2 \leq \alpha + \epsilon\}$. By assumption, with probability approaching 1 (w.p.a. 1).

$$\hat{\mathbf{w}}^B \in \mathcal{D}. \tag{13}$$

For the shake of a contradiction, assume that there exists optimal solution $\hat{\mathbf{w}}$ of (1) such that $\hat{\mathbf{w}} \notin \mathcal{D}$ w.p.a. 1. Clearly,

$$M(\hat{\mathbf{w}}) \leq M(\hat{\mathbf{w}}^B). \tag{14}$$

Suppose $\hat{\mathbf{w}} \in \mathcal{S}_{\mathbf{w}}$, then $\hat{\mathbf{w}}$ is optimal for (3) and satisfies (13), which contradicts our assumption. Thus, $\hat{\mathbf{w}} \notin \mathcal{S}_{\mathbf{w}}$. Next, let $\mathbf{w}_\theta := \theta\hat{\mathbf{w}} + (1 - \theta)\hat{\mathbf{w}}^B$ for $\theta \in (0, 1)$ such that $\mathbf{w}_\theta \notin \mathcal{D}$ and $\mathbf{w}_\theta \in \mathcal{S}_w$ (always possible, by definition of $\mathcal{D}$). By the convexity of F and (14), it follows that $M(\hat{\mathbf{w}}_\theta) \leq M(\hat{\mathbf{w}}^B)$. Hence, $\hat{\mathbf{w}}_\theta$ is optimal for (3) and satisfies (13), which, again, is a contradiction. This completes the proof.

Proof of Lemma 27

It suffices to prove the equivalence of the optimization (4) and (5). Let $\mathbf{w}_*, \mathbf{v}_*, \mathbf{u}_*$ be optimal in (4). To prove the claim, we show that $\mathbf{u}_* \in \mathcal{S}_{\mathbf{u}}$ $\left(\Leftrightarrow \|\mathbf{u}_*\|_2 \leq K_\beta\right)$ w.p.a. 1. From the first order optimality conditions in (4), we find that

$$\mathbf{u}_* \in \frac{1}{\sqrt{p}} \partial \mathcal{L}(\mathbf{v}_*) \tag{15}$$

$$\mathbf{v}_* = \mathbf{z} - \sqrt{p}\mathbf{X}\mathbf{w}_*. \tag{16}$$

Recall Assumption 1 and consider two cases. First, if $\sup_{\mathbf{v}\in\mathbb{R}^n} \sup_{\mathbf{s}\in\partial\mathcal{L}(\mathbf{v})} \|\mathbf{s}\|_2 < \infty$, the claim follows directly by (15). Next, assume that w.h.p., $\|\mathbf{z}\|_2 \leq C_1\sqrt{p}$ for constant $C_1 > 0$. Also, a standard high probability bound on the spectral norm of Gaussian matrices gives $\|\mathbf{X}\|_2 \leq C_2$, e.g. [186]. Using these, boundedness of $\mathbf{w}_*$ and (16), we find that $\|\mathbf{v}_*\|_2 \leq C_3\sqrt{p}$ w.h.p.. Then, the normalization condition $\frac{1}{\sqrt{p}} \sup_{\mathbf{s}\in\partial\mathcal{L}(\mathbf{v})} \|\mathbf{s}\|_2 \leq C$ for all $\|\mathbf{v}\|_2 \leq c\sqrt{p}$ and all $p \in \mathbb{N}$, yields the desired, i.e. $\|\mathbf{u}_*\|_2 \leq C$ holds with probability approaching 1 as $p \to \infty$.

Proof of Lemma 28

Let $\mathbf{w}_*$ denote an optimal solution of the "bounded" optimization in (6). It will suffice to prove that $\mathbf{w}_* \in \mathcal{S}$ in probability. To see this, recall from Lemma 27 that (6) is asymptotically equivalent to (3). Then, Lemma 26 and the assumption $\alpha_* < K_\alpha$ guarantee that $\hat{\mathbf{w}}(\mathbf{X}) \in \mathcal{S}$ in probability, as desired.

Denote $\Phi := \Phi(\mathbf{X})$ the optimal cost of the minimization in (6) and $\Phi_{\mathcal{S}^c} := \Phi_{\mathcal{S}^c}(\mathbf{X})$ the optimal cost of the same problem when the minimization is further restricted to be over the set $\mathbf{w} \in \mathcal{S}^c$. Note that $\mathbf{w}_* \in \mathcal{S}$ iff $\Phi_{\mathcal{S}^c}(\mathbf{X}) > \Phi(\mathbf{X})$; hence, it will suffice to prove that the latter event occurs in probability.

We do so by relating the (PO) in (6) to the Auxiliary Optimization (AO) in (8) using Theorem 8. For concreteness, denote the objective function in (8) with $A(\mathbf{w}, \mathbf{v}, \mathbf{u}, \mathbf{s})$, and, recall $\mathcal{S}_{\mathbf{w}} := \{\mathbf{w} \mid \|\mathbf{w}\|_2 \leq K_\alpha\}$, $\mathcal{S}_{\mathbf{u}} := \{\mathbf{u} \mid \|\mathbf{u}\|_2 \leq K_\beta\}$. With these, define

$$\begin{aligned} \phi^P &:= \phi^P(\mathbf{g}, \mathbf{h}) := \min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}, \mathbf{v}} \max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}, \mathbf{s}} A(\mathbf{w}, \mathbf{v}, \mathbf{u}, \mathbf{s}) \quad \text{and} \\ \phi^D &:= \phi^D(\mathbf{g}, \mathbf{h}) := \max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}, \mathbf{s}} \min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}, \mathbf{v}} A(\mathbf{w}, \mathbf{v}, \mathbf{u}, \mathbf{s}). \end{aligned} \tag{17}$$

Observe here that the order of min-max in ϕ^P is exactly as in the original formulation of the CGMT, cf. (2.126b); ϕ^D is the dual of it, while ϕ in (8) involves yet another change in the order of the optimizations. The reason we prefer to work with the

later problem, is that this particular order allows for a number of simplifications performed in Section .1.

As done before, denote with ${\phi^P}_{\mathcal{S}^c}, {\phi^D}_{\mathcal{S}^c}$ the optimal costs of the optimization problems in (17) under the additional constraint $\mathbf{w} \in \mathcal{S}^c$. The two problems in (17) are related to the one in (8) as follows:

$$\begin{aligned} {\phi^P}_{\mathcal{S}^c} &= \min_{\substack{\mathbf{w}\in\mathcal{S}_\mathbf{w},\mathbf{v} \\ \mathbf{w}\in\mathcal{S}^c}} \max_{\mathbf{u}\in\mathcal{S}_\mathbf{u},\mathbf{s}} A(\mathbf{w},\mathbf{v},\mathbf{u},\mathbf{s}) = \min_{\substack{\mathbf{w}\in\mathcal{S}_\mathbf{w},\mathbf{v} \\ \mathbf{w}\in\mathcal{S}^c}} \max_{\beta,\mathbf{s}} \max_{\|\mathbf{u}\|_2=\beta} A(\mathbf{w},\mathbf{v},\mathbf{u},\mathbf{s}) \\ &\geq \max_{\beta,\mathbf{s}} \min_{\substack{\mathbf{w}\in\mathcal{S}_\mathbf{w},\mathbf{v} \\ \mathbf{w}\in\mathcal{S}^c}} \max_{\|\mathbf{u}\|_2=\beta} A(\mathbf{w},\mathbf{v},\mathbf{u},\mathbf{s}) = \phi_{\mathcal{S}^c}, \end{aligned} \tag{18}$$

where the inequality follows from the min-max inequality [136, Lem. 36.1]. Similarly,

$$\begin{aligned} \phi^D &= \max_{\mathbf{u}\in\mathcal{S}_\mathbf{u},\mathbf{s}} \min_{\mathbf{w}\in\mathcal{S}_\mathbf{w},\mathbf{v}} A(\mathbf{w},\mathbf{v},\mathbf{u},\mathbf{s}) = \max_{\beta,\mathbf{s}} \max_{\|\mathbf{u}\|_2=\beta} \min_{\mathbf{w}\in\mathcal{S}_\mathbf{w},\mathbf{v}} A(\mathbf{w},\mathbf{v},\mathbf{u},\mathbf{s}) \\ &\leq \max_{\beta,\mathbf{s}} \min_{\mathbf{w}\in\mathcal{S}_\mathbf{w},\mathbf{v}} \max_{\|\mathbf{u}\|_2=\beta} A(\mathbf{w},\mathbf{v},\mathbf{u},\mathbf{s}) = \phi. \end{aligned} \tag{19}$$

Furthermore, they are related to the (PO) via the CGMT. From Theorem 8(i), for all $c \in \mathbb{R}$:

$$\mathbb{P}(\Phi_{\mathcal{S}^c} < c) \leq 2\mathbb{P}({\phi^P}_{\mathcal{S}^c} \leq c). \tag{20}$$

Also, from Theorem 8(ii)[2]:

$$\mathbb{P}(\Phi > c) \leq 2\mathbb{P}(\phi^D \geq c). \tag{21}$$

The remaining of the proof is in the same lines as the proof of 8(iii), but is included for clarity. Let $\eta := (\phi_{\mathcal{S}^c} - \phi)/3 > 0$. We may apply (20) for $c = \overline{\phi}_{\mathcal{S}^c} - \eta$ and combine with (18) to find that

$$\mathbb{P}(\Phi_{\mathcal{S}^c} < \overline{\phi}_{\mathcal{S}^c} - \eta) \leq 2\mathbb{P}({\phi^P}_{\mathcal{S}^c} \leq \overline{\phi}_{\mathcal{S}^c} - \eta) \leq 2\mathbb{P}(\phi_{\mathcal{S}^c} \leq \overline{\phi}_{\mathcal{S}^c} - \eta). \tag{22}$$

From assumption (b) the last term above tends to zero as $p \to \infty$. In a similar way, combining (21), (19) and assumption (a), we find that

$$\mathbb{P}(\Phi > \overline{\phi} + \eta) \leq 2\mathbb{P}(\phi^D \geq \overline{\phi} + \eta) \leq 2\mathbb{P}(\phi \geq \overline{\phi} + \eta), \tag{23}$$

goes to zero with $p \to \infty$. Denote the event $\mathcal{E} = \{\Phi_{\mathcal{S}^c} \geq \overline{\phi}_{\mathcal{S}^c} - \eta \text{ and } \Phi \leq \overline{\phi} + \eta\}$. From (22) and (23) the event occurs with probability approaching 1. Furthermore, in this event, after using assumption (a), we have ${\Phi^b}_{\mathcal{S}^c} \geq \overline{\phi}_{\mathcal{S}^c} - \eta > \overline{\phi} + \eta \geq \Phi^b$; equivalently, the optimal minimizer satisfies $\mathbf{w}_* \in \mathcal{S}$, which completes the proof.

[2] more precisely, please refer to equation (32) in [171].

Proof of Lemma 29

(i) We start by showing how the vector optimization in (8) can be reduced to the scalar one that appears in (9). This requires the following steps.

Optimizing over the direction of $\mathbf{u}$*:* Performing the inner maximization is easy. In particular, using the fact that $\max_{\|\mathbf{u}\|_2=\beta} \mathbf{u}^T\mathbf{t} = \beta\|\mathbf{t}\|_2$ for all $\beta \geq 0$ the problem simplifies to a max-min one:

$$\max_{0\leq\beta\leq K_\beta,\mathbf{s}} \min_{\|\mathbf{w}\|\leq K_\alpha,\mathbf{v}} \frac{\beta}{\sqrt{p}}\| \|\mathbf{w}\|_2\mathbf{g}+\mathbf{z}-\mathbf{v} \|_2 - \frac{\beta}{\sqrt{p}}\mathbf{h}^T\mathbf{w} + \frac{1}{p}\mathcal{L}(\mathbf{v}) + \frac{\lambda}{p}\mathbf{s}^T\mathbf{x}_0 + \frac{\lambda}{\sqrt{p}}\mathbf{s}^T\mathbf{w} - \frac{\lambda}{p}f^*(\mathbf{s}), \tag{24}$$

Optimizing over the direction of $\mathbf{w}$*:* Next, we fix $\|\mathbf{w}\|_2 = \alpha$, and, similar to what was done above, minimize over its direction:

$$\max_{0\leq\beta\leq K_\beta,\mathbf{s}} \min_{0\leq\alpha\leq K_\alpha,\mathbf{v}} \frac{\beta}{\sqrt{p}}\| \alpha\mathbf{g}+\mathbf{z}-\mathbf{v} \|_2 + \frac{1}{p}\mathcal{L}(\mathbf{v}) - \frac{\alpha}{\sqrt{p}}\|\beta\mathbf{h}-\lambda\mathbf{s}\|_2 + \frac{\lambda}{p}\mathbf{s}^T\mathbf{x}_0 - \frac{\lambda}{p}f^*(\mathbf{s}). \tag{25}$$

Changing the orders of min-max: Denote with $M(\alpha,\beta,\mathbf{v},\mathbf{s})$ the objective function above. It can be checked that M is jointly convex in $(\alpha,\mathbf{v})$ and jointly concave in $(\beta,\mathbf{s})$ (cf. Lemma 34). Thus, $\min_{\mathbf{v}} M$ is convex in α and jointly concave in $(\beta,\mathbf{s})$. Furthermore, the constraint sets are all convex and the one over which minimization over α occurs is bounded. Hence, as in [149, Cor. 3.3] we can flip the order of $\max_{\beta,\mathbf{s}}\min_{\alpha}$, to conclude with

$$\min_{0\leq\alpha\leq K_\alpha} \max_{0\leq\beta\leq K_\beta} \max_{\mathbf{s}} \min_{\mathbf{v}} M(\alpha,\beta,\mathbf{v},\mathbf{s}).$$

Also, observe that the order of optimization among $\mathbf{v}$ and $\mathbf{s}$ does not affect the outcome.

The square-root trick: We apply the fact that $\sqrt{\chi} = \inf_{\tau>0}\{\frac{\tau}{2}+\frac{\chi}{2\tau}\}$ to both the terms $\frac{1}{\sqrt{p}}\|\alpha\mathbf{g}+\mathbf{z}-\mathbf{v}\|_2$ and $\frac{1}{\sqrt{p}}\|\beta\mathbf{h}-\lambda\mathbf{s}\|_2$:

$$\min_{0\leq\alpha\leq K_\alpha} \max_{0\leq\beta\leq K_\beta} \inf_{\tau_g>0} \sup_{\tau_h>0} \frac{\beta\tau_g}{2} + \frac{1}{p}\min_{\mathbf{v}}\left\{\frac{\beta}{2\tau_g}\| \alpha\mathbf{g}+\mathbf{z}-\mathbf{v} \|_2^2 + \mathcal{L}(\mathbf{v})\right\} - \frac{\alpha\tau_h}{2} - \frac{1}{p}\min_{\mathbf{s}}\left\{\frac{\alpha}{2\tau_h}\|\beta\mathbf{h}-\lambda\mathbf{s}\|_2^2 - \lambda\mathbf{s}^T\mathbf{x}_0 + \lambda f^*(\mathbf{s})\right\}. \tag{26}$$

Identifying the Moreau envelope: Arguing as before, we can change the order of optimization between β and τ_g. Also, it takes only a few algebra steps and using

basic properties of Moreau envelope functions (in particular, Lemma 35(ii)) in order to rewrite the last summand in (26) as below. If $\alpha > 0$, then,

$$\min_{\mathbf{s}} \left\{ \frac{\alpha}{2\tau_h} \|\beta\mathbf{h} - \lambda\mathbf{s}\|_2^2 - \lambda \mathbf{s}^T \mathbf{x}_0 + \lambda f^*(\mathbf{s}) \right\}$$

$$= -\frac{\tau_h}{2\alpha} \|\mathbf{x}_0\|_2^2 - \beta \mathbf{h}^T \mathbf{x}_0 + \lambda \cdot \mathrm{e}_{f^*} \left(\frac{\beta}{\lambda}\mathbf{h} + \frac{\tau_h}{\alpha\lambda}\mathbf{x}_0; \frac{\tau_h}{\alpha\lambda} \right) \tag{27}$$

$$= \frac{\beta^2\alpha}{2\tau_h} \|\mathbf{h}\|^2 - \lambda \cdot \mathrm{e}_f \left(\frac{\beta\alpha}{\tau_h}\mathbf{h} + \mathbf{x}_0; \frac{\alpha\lambda}{\tau_h} \right). \tag{28}$$

Otherwise, if $\alpha = 0$, then the same term equals $-\lambda f(\mathbf{x}_0)$ since $\max_{\mathbf{s}} \mathbf{s}^T\mathbf{x}_0 - f^*(\mathbf{s}) = f(\mathbf{x}_0)$.

(ii) The continuity of the objective function in (9) follows directly from the continuity of the Moreau envelope functions, cf. [137, Lem. 1.25, 2.26]. In particular, regarding the two branches of the objective: it can be checked, using the continuity of the Moreau envelope, that the limit of the RHS in (28) as $\alpha \to 0$ evaluates to $-\lambda f(\mathbf{x}_0)$. (In fact, this is the unique extension of the upper branch to a continuous finite convex function on the whole $\alpha \geq 0, \tau > 0$, as per [136, Thm. 10.3]).

Convexity of (9) can be checked from (26). By applying Lemma 34, after minimization over $\mathbf{v}$ the Moreau Envelope remains jointly convex with respect to α and τ_g and concave in β. The same argument (and similar lemma) holds for the last term of (26) in which after minimization over $\mathbf{s}$ it remains jointly convex in β and τ_h and concave in α. Then the negative sign before this term makes it jointly concave in β and τ_h and convex over α.

Proof of Lemma 30

(a) By Assumption 1 the normalized Moreau envelope functions in (10) converge in probability to L and F, respectively. Also, $\|\mathbf{h}\|_2^2/p \xrightarrow{P} 1$ by the WLLN. This proves the convergence part.

Lemma 29(i) showed $\mathcal{R}_p$ to be convex-concave. Then, the same holds for $\mathcal{D}$ by point-wise convergence and the fact that convexity is preserved by point wise limits.

(b) Call

$$M_n(\alpha) = \sup_{\substack{0 \leq \beta \leq K_\beta \\ \tau_h > 0}} \inf_{\tau_g > 0} \mathcal{R}_n(\alpha, \tau_g, \beta, \tau_h) \quad \text{and} \quad M(\alpha) = \sup_{\substack{0 \leq \beta \\ \tau_h > 0}} \inf_{\tau_g > 0} \mathcal{D}_n(\alpha, \tau_g, \beta, \tau_h). \tag{29}$$

The bulk of the proof consists of showing that the following two statements hold

$$\forall \text{ compact susets } \mathcal{A} \subset (0,\infty) \text{ and sufficiently large } K_\beta := K_\beta(\mathcal{A}) > 0$$
$$: \inf_{\alpha\in\mathcal{A}} M_n(\alpha) \xrightarrow{P} \inf_{\alpha\in\mathcal{A}} M(\alpha) \tag{30}$$

and,

$$\forall \epsilon > 0, \text{w.p.a.1} : \; M_n(0) < M(0) + \epsilon. \tag{31}$$

Before proceeding with the proof of those, let us show how the conclusion of the lemma is reached once (30) and (31) are established.

Using (30) and (31) to prove the lemma : Fix $K_\alpha > \alpha_*$, any $\delta > 0$ such that $\mathcal{A} := [\alpha_* - 2\delta, \alpha_* + 2\delta] \subset (0, K_\alpha]$ and $K_\beta > 0$ large enough such that (30) and (31) both hold. Then, for all $\epsilon > 0$, w.p.a.1:

$$\min_{0\le\alpha\le K_\alpha} M_n(\alpha) \le \min_{\alpha\in\mathcal{A}} M_n(\alpha) \le M_n(\alpha_*) < M(\alpha_*) + \epsilon. \tag{32}$$

For the last inequality above: if $\alpha_* = 0$, it follows from (31), or otherwise from (30).

Next, consider the compact set $\mathcal{A}_l = \{\alpha > 0 \mid \alpha \in [\alpha_* - 2\delta, \alpha - \delta] \}$ and $\mathcal{A}_r = \{\alpha > 0 \mid \alpha \in [\alpha_* + \delta, \alpha + 2\delta] \}$. (Note that if $\alpha_* = 0$, then $\mathcal{A}_l$ is empty.) From (30), we know that for all $\epsilon > 0$, w.p.a.1

$$\min_{\alpha\in\mathcal{A}_l} M_n(\alpha) > \min_{\alpha\in\mathcal{A}_l} M(\alpha) - \epsilon \quad \text{and} \quad \min_{\alpha\in\mathcal{A}_u} M_n(\alpha) > \min_{\alpha\in\mathcal{A}_u} M(\alpha) - \epsilon.$$

Let $\mathcal{A}_{lu} = \mathcal{A}_l \cup \mathcal{A}_u$ and combine the above to find

$$\min_{\alpha\mathcal{A}_{lu}} M_n(\alpha) > \min_{\alpha\in\mathcal{A}_{lu}} M(\alpha) - \epsilon. \tag{33}$$

By assumption on uniqueness of α_* and on convexity of M, we have

$$M(\alpha_*) < \min_{\alpha\in\mathcal{A}_{lu}} M(\alpha) \tag{34}$$

and $M(\alpha_*) = \min_{\alpha\in\mathcal{A}} M(\alpha)$. Thus, Applying (32) and (33) for $\epsilon = (\min_{\alpha\in\mathcal{A}_{lu}} M(\alpha) - M(\alpha_*))/3$ yields w.p.a.1 :

$$\min_{\alpha\in\mathcal{A}_{lu}} M_n(\alpha) > \min_{\alpha\in\mathcal{A}_{lu}} M(\alpha) - \epsilon > M(\alpha_*) + \epsilon > \min_{\alpha\in[\alpha_*-2\delta,\alpha_*+2\delta]} M_n(\alpha). \tag{35}$$

Thus, w.p.a.1,

$$\hat{\alpha}_n := \arg\min_{\alpha\in\mathcal{A}} M_n(\alpha) \in (\alpha_* - \delta, \alpha_* + \delta).$$

In this event, for any $\alpha \notin \mathcal{A}$, there is a convex combination $\alpha_\theta := \theta\hat{\alpha}_n + (1-\theta)\alpha$, $(\theta < 1)$ that equals either $\alpha_* - 2\delta$ or $\alpha_* + 2\delta$. By convexity,

$$M_n(\alpha_\theta) \leq \theta M_n(\hat{\alpha}_n) + (1-\theta)M_n(\alpha)$$

Also, from (35), $M_n(\hat{\alpha}_n) < M_n(\alpha_\theta)$. Combining those, we find $M_n(\hat{\alpha}_n) < M_n(\alpha)$, implying that $\hat{\alpha}_n$ is the minimizer of M_n over the entire $[0, K_\alpha]$ w.p.a.1. In other words, for all ϵ w.p.a. 1,

$$\min_{\alpha\in[0,K_\alpha]\setminus(\alpha_*-\delta,\alpha_*+\delta)} M_n(\alpha) \geq \min_{\alpha\in\mathcal{A}_{lu}} M_n(\alpha) > \min_{\alpha\in\mathcal{A}_{lu}} M(\alpha) - \epsilon. \tag{36}$$

To establish a connection with the three statements (i)-(iii) of the lemma, observe that $\overline{\phi}_{[0,\infty)} = M(\alpha_*)$. Also, $\overline{\phi}_{[0,\infty)\setminus\mathcal{S}_\delta} = \min_{\alpha\in\mathcal{A}_{lu}} M(\alpha)$ (by convexity). With these, (i) corresponds directly to (32), (ii) to (36), and, (iii) to (34).

Proof of (30) and (31) : From the first statement of the lemma, the objective function $\mathcal{R}_n$ of the (AO) converges point-wise to $\mathcal{D}$. We will use this to show that the minimax value of $\mathcal{R}_p$ converges to the corresponding minimax of $\mathcal{D}$. The proof is based on a repeated use of Lemma 31 below, about convergence of the infimum of a sequence of convex converging stochastic processes. This fact is essentially a consequence of what is known in the literature as the *convexity lemma*, according to which point wise convergence of convex functions implies uniform convergence in compact subsets. Please refer to Section .2 for the proof.

Lemma 31 (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, $\inf_{x>0} M_n(x) \xrightarrow{P} \inf_{x>0} F(x)$.

1) Fix $\alpha \geq 0, \beta > 0$, and, $\tau_h > 0$. Consider

$$M_n^{\alpha,\beta,\tau_h}(\tau_g) := R_n(\alpha, \tau_g, \beta, \tau_h), \tag{37}$$

$$M^{\alpha,\beta,\tau_h}(\tau_g) := \mathcal{D}(\alpha, \tau_g, \beta, \tau_h). \tag{38}$$

The functions $\{M_n\}$ are convex. Furthermore, $M_n^{\alpha,\beta,\tau_h}(\tau_g) \xrightarrow{P} M^{\alpha,\beta,\tau_h}(\tau_g)$ point wise in τ_g. Next, we show that M^{α,β,τ_h} is level-bounded, i.e. it satisfies condition (b) of Lemma 31. In view of Lemma 32, it suffices to show that $\lim_{\tau_g\to\infty} M^{\alpha,\beta,\tau_h}(\tau_g) = +\infty$, or $\lim_{\tau_g\to\infty}\left(\frac{\beta}{2} + \delta\cdot\frac{L(\alpha,\tau_g/\beta)}{\tau_g}\right) > 0$. By assumption 1, $\lim_{\tau_g\to\infty} L(\alpha,\tau_g/\beta) = -L_0$. There is two cases to be considered. Either $L_0 < \infty$, or else, Assumption 1 holds. Either way, $\lim_{\tau_g\to\infty} L(\alpha,\tau_g/\beta)/\tau_g = 0$ and we are done. Now, we can apply Lemma 31 to conclude that

$$\inf_{\tau_g>0} M_n^{\alpha,\beta,\tau_h}(\tau_g) \xrightarrow{P} \inf_{\tau_g>0} M^{\alpha,\beta,\tau_h}(\tau_g). \tag{39}$$

2)

Next, again for fixed $\alpha \geq 0, \tau_h > 0$, consider (we use some abuse of notation here, with the purpose of not overloading notation)

$$M_n^{\alpha,\tau_h}(\beta) := \inf_{\tau_g>0} M_n^{\alpha,\beta,\tau_h}(\tau_g)$$

$$M^{\alpha,\tau_h}(\beta) := \inf_{\tau_g>0} M^{\alpha,\beta,\tau_h}(\tau_g)$$

The functions $\{M_n^{\alpha,\tau_h}\}$ are concave in β, as the point wise minima of concave functions. Furthermore, $M_n^{\alpha,\tau_h}(\beta) \xrightarrow{P} M^{\alpha,\tau_h}(\beta)$ point wise in $\beta > 0$, by (39).

$\underline{\alpha > 0:}$ For now and until further notice, restrict attention to the case $\alpha > 0$. Also, consider first $\beta > 0$. We show that M^{α,τ_h} is level-bounded, i.e. it satisfies condition (b) of Lemma 31. In view of Lemma 32, it suffices to show that $\lim_{\beta\to+\infty} M^{\alpha,\tau_h}(\beta) = -\infty$, or $\lim_{\beta\to+\infty} \inf_{\tau_g>0} M^{\alpha,\beta,\tau_h}(\tau_g) = -\infty$. This condition is equivalent to the following

$$(\forall M > 0)(\exists B > 0)\ \left[\beta > B \Rightarrow (\exists\{\tau_g\}_k)\ [D(\alpha,\tau_g,\beta,\tau_h) < -M]\ \right]. \tag{40}$$

First, we show that

$$\lim_{\beta\to+\infty} \frac{\alpha\beta^2}{2\tau_h} - \lambda\cdot F\left(\frac{\alpha\beta}{\tau_h}, \frac{\alpha\lambda}{\tau_h}\right) = +\infty \tag{41}$$

This follows by Assumption 1 when applied for $c = \alpha\beta/\tau_h$ and $\tau = \alpha\lambda/\tau_h$ (recall here that $\alpha > 0$).

Next, choose $\{\tau_g\}_k \to 0$. For that choice, $\frac{\beta\tau_g}{2} + L(\alpha,\tau_g/\beta) \to \lim_{\tau\to 0} L(\alpha,\tau) < \infty$, where boundedness follows by Assumption 1. Thus, (40) is correct and we may apply Lemma 31 to conclude that

$$\sup_{\beta>0} M_n^{\alpha,\tau_h}(\beta) \xrightarrow{P} \sup_{\beta>0} M^{\alpha,\tau_h}(\beta). \tag{42}$$

Now, we investigate the case $\beta = 0$. We have, $M_n^{\alpha,\tau_h}(0) = -\frac{1}{p}\mathcal{L}(\mathbf{z}) - \frac{\alpha\tau_h}{2} + \frac{\lambda}{p}\left(\mathrm{e}_f\left(\mathbf{x}_0; \frac{\alpha\lambda}{\tau_h}\right) - f(\mathbf{x}_0)\right)$ and $M_n^{\alpha,\tau_h}(0) = -\delta L_0 - \frac{\alpha\tau_h}{2} + F(0, \frac{\alpha\lambda}{\tau_h})$.

If $L_0 < \infty$, then by assumption, $M_n^{\alpha,\tau_h}(0) \xrightarrow{P} M^{\alpha,\tau_h}(0)$. Combined with (42), we find

$$\sup_{\beta\geq 0} M_n^{\alpha,\tau_h}(\beta) \xrightarrow{P} \sup_{\beta\geq 0} M^{\alpha,\tau_h}(\beta). \tag{43}$$

Now, consider the case $L_0 = +\infty$. Clearly, the optimal β for M^{α,τ_h} is not at zero; thus, $\sup_{\beta\geq 0} M^{\alpha,\tau_h}(\beta) = \sup_{\beta>0} M^{\alpha,\tau_h}(\beta)$. Also, by assumption, for all M, $\lim_{p\to\infty} \mathbb{P}\left(\frac{1}{p}\mathcal{L}(\mathbf{z}) > M\right) = 1$. Letting, $\epsilon > 0$ and $M := -\sup_{\beta>0} M^{\alpha,\tau_h}(\beta) + \epsilon + \frac{\alpha\tau_h}{2} - F(0, \frac{\alpha\lambda}{\tau_h})$, then w.p.a.1, $M_n^{\alpha,\tau_h}(0) < \sup_{\beta>0} M^{\alpha,\tau_h}(\beta) - \epsilon \leq \sup_{\beta>0} M_n^{\alpha,\tau_h}(\beta)$, where the last inequality follows because of (42). Again, this leads to (43). To sum up, (43) holds for all $\alpha > 0$.

$\underline{\alpha = 0:}$ We show that for all $\epsilon > 0$, the following holds w.p.a.1:

$$\sup_{\beta\geq 0} M_n^{\alpha=0,\tau_h}(\beta) < \sup_{\beta\geq 0} M^{\alpha=0,\tau_h}(\beta) + \epsilon. \tag{44}$$

To begin with, note that for all p,

$$\sup_{\beta\geq 0} M_n^{\alpha=0,\tau_h}(\beta) \leq \sup_{\beta>0} \lim_{\tau_g\to 0} \frac{\beta\tau_g}{2} + \frac{1}{p}\min_{\mathbf{v}}\left\{\frac{\beta}{2\tau_g}\|\mathbf{z}-\mathbf{v}\|_2^2 + \mathcal{L}(\mathbf{v}) - \mathcal{L}(\mathbf{z})\right\} = 0, \tag{45}$$

where we have used Lemma 40(ix). Next, we show that

$$M^{\alpha=0,\tau_h}(\beta) = 0. \tag{46}$$

Using Assumption 1 on the non-negativity of L_0 and Assumption 1 that $\lim_{\tau\to 0} L(c,\tau) = 0$, it follows that $M^{\alpha=0,\tau_h}(\beta) \leq \sup_{\beta>0} \lim_{\tau_g\to 0} \frac{\beta\tau_g}{2} + L(0, \tau_g/\beta) = 0$. Thus, it will suffice for the claim if we prove

$$\lim_{\beta\to\infty} \inf_{\tau_g>0} \frac{\beta\tau_g}{2} + L(0, \tau_g/\beta) = 0, \tag{47}$$

or equivalently,

$$\lim_{\beta\to\infty} \inf_{\kappa>0} \kappa\left(\frac{\beta^2}{2} + \frac{L(0,\kappa)}{\kappa}\right) = 0.$$

Fix some $\beta > 0$. Note that $\lim_{\kappa\to 0} \frac{\kappa\beta^2}{2} + L(0,\kappa) = 0$, where we have used Assumption 1 that $\lim_{\tau\to 0} L(0,\tau) = 0$. Also, $\lim_{\kappa\to\infty} \kappa\left(\frac{\beta^2}{2} + \frac{L(0,\kappa)}{\kappa}\right) = +\infty$, using Assumption

1 this time. Now, consider only $\beta > \sqrt{-L_{2,+}(0,0)}$ (see Assumption 1). Then, the function $\frac{\kappa\beta^2}{2} + L(0,\kappa)$ has a positive derivative at $\kappa \to 0^+$. From this and convexity, it follows that for all $\kappa > 0$,

$$\frac{\kappa\beta^2}{2} + L(0,\kappa) \geq \lim_{\kappa\to 0} \frac{\kappa\beta^2}{2} + L(0,\kappa) = 0.$$

This proves (47) as desired.

To complete the argument, (44) follows by (45) and (46), and with this we have completed the proof of (31).

3) Keep $\alpha > 0$ fixed and consider

$$M_n^{\alpha}(\tau_h) := \sup_{\beta\geq 0} M_n^{\alpha,\tau_h}(\beta),$$

$$M^{\alpha}(\tau_h) := \sup_{\beta\geq 0} M^{\alpha,\tau_h}(\beta),$$

The functions $\{M_n^{\alpha}\}$ and F are all concave in τ_h, as the point wise maxima of jointly concave functions. Furthermore, $M_n^{\alpha}(\tau_h) \xrightarrow{P} M^{\alpha}(\tau_h)$ point wise in τ_h, by (43). Next, we show that M^{τ_h} is level-bounded, i.e. it satisfies condition (b) of Lemma 31. In view of Lemma 32, it suffices to show that $\lim_{\tau_h\to\infty} M^{\alpha}(\tau_h) = +\infty$, or $\lim_{\tau_h\to\infty} \sup_{\beta>0} \inf_{\tau_g>0} \mathcal{D}(\alpha,\tau_g,\beta,\tau_h) = -\infty$. This is equivalent to the following

$$(\forall M > 0)(\exists T > 0)\ \left[\tau_h > T \Rightarrow (\forall\{\beta\}_k)(\exists\{\tau_g\}_k)\ [D(\alpha,\tau_g,\beta,\tau_h) < -M]\ \right]. \tag{48}$$

Consider the function

$$\mathcal{H}(\beta,\tau_h) := \frac{\alpha\tau_h}{2} + \frac{\alpha\beta^2}{2\tau_h} - \lambda\cdot F(\frac{\alpha\beta}{\tau_h}, \frac{\alpha\lambda}{\tau_h}).$$

We show that

$$\mathcal{H}(\beta,\tau_h) \geq \frac{\alpha\tau_h}{2}.$$

To see this note that $\mathrm{e}_f\,(c\mathbf{h}+\mathbf{x}_0;\tau) \leq \frac{c^2\|\mathbf{h}\|^2}{2\tau} + f(\mathbf{x}_0)$. Thus, $\frac{1}{p}\left\{\mathrm{e}_f\,(c\mathbf{h}+\mathbf{x}_0;\tau) - f(\mathbf{x}_0)\right\} \leq \frac{c^2\|\mathbf{h}\|^2}{2\tau p}$. The LHS converges to $F(c,\tau)$ by Assumption 1 and the RHS converges to $\frac{c^2}{2\tau}$. Therefore, $F(c,\tau) \leq \frac{c^2}{2\tau}$. Applying this for $c = \frac{\alpha\beta}{\tau_h}$ and $\tau = \frac{\alpha\lambda}{\tau_h}$, we have that $\frac{\alpha\beta^2}{2\tau_h} - \lambda\cdot F(\frac{\alpha\beta}{\tau_h}, \frac{\alpha\lambda}{\tau_h}) \geq 0$, as desired.

Then,

$$\mathcal{D}(\alpha,\tau_g,\beta,\tau_h) \leq \frac{\beta\tau_g}{2} + \delta\cdot L\left(\alpha, \frac{\tau_g}{\beta}\right) - \frac{\alpha\tau_h}{2}.$$

Also, note that for all $\beta > 0$, we can choose (sequence) of τ_g, such that $\beta\tau_g, \frac{\tau_g}{\beta} \to 0$. Then, $\frac{\beta\tau_g}{2} + \delta \cdot L\left(\alpha, \frac{\tau_g}{\beta}\right) \to \lim_{\tau\to 0} L(\alpha, \tau) =: A < \infty$. It can then be seen that (48) holds for (say) $T := T(M) = 4(A + M)/\alpha$.

We can apply Lemma 31 to conclude that

$$\sup_{\tau_h>0} M_n^{\alpha}(\tau_h) \xrightarrow{P} \sup_{\tau_h>0} M^{\alpha}(\tau_h). \tag{49}$$

4) Finally, consider

$$M_n(\alpha) := \sup_{\tau_h>0} M_n^{\alpha}(\tau_h),$$

$$M(\alpha) := \sup_{\tau_h>0} M^{\alpha}(\tau_h). \tag{50}$$

The functions $\{M_n\}$ and F are all convex in τ_h, as the point wise maxima of convex functions. Furthermore, $M_n(\alpha) \xrightarrow{P} M(\alpha)$ point wise in α, by (49). By assumption of the lemma, F has a unique minimizer α_*, which of course implies level boundedness. Thus, we can apply Lemma 10 to conclude that

$$\inf_{\alpha>0} M_n(\alpha) \xrightarrow{P} \inf_{\alpha>0} M(\alpha). \tag{51}$$

Besides, pointwise convergence $M_p(\alpha) \xrightarrow{P} M(\alpha)$ translates to uniform convergence over any compact subset $\mathcal{A} \subset (0, \infty)$ by the Convexity lemma [7, Cor.. II.1] ,[107, Lem. 7.75]. Hence,

$$\inf_{\alpha\in\mathcal{A}} M_n(\alpha) \xrightarrow{P} \inf_{\alpha\in\mathcal{A}} M(\alpha).$$

This is of course same as the desired in (30). Recall, (31) was established in (44). The only thing remaining is showing that there exists an optimal β_* in $\sup_{\beta\geq 0} M^{\alpha,\tau_h}(\beta)$ that is bounded by some sufficiently large $K_\beta(\mathcal{A})$. This follows from the level-boundedness arguments above as detailed immediately next.

Boundedness of solutions : For a compact subset $\mathcal{A} \subset (0, \infty)$, we argue that there exists *bounded* β_* and sequences $\{\tau_{g_*}\}_k, \{\tau_{h*}\}_k$ such that $(\alpha_*, \{\tau_{g_*}\}_k, \beta_*, \{\tau_{h*}\}_k)$ approaches
$\min_{\alpha\in\mathcal{A}} \sup_{\substack{\tau_h>0\\ \beta\geq 0}} \inf_{\tau_g>0} \mathcal{D}(\alpha, \tau_g, \beta, \tau_h)$. This follows from the work above. In particular, at each step in the proof of (30) above, we showed level-boundedness of the corresponding functions. For example, (48) shows that there exists (sufficiently large) $T_h(\alpha) > 0$ such that $\sup_{\tau_h>0} M^{\alpha}(\tau_h)$ is equal to $\sup_{T_h(\alpha)\geq\tau_h>0} M^{\alpha}(\tau_h)$. This holds for all α; so, in particular, is true for $T_h := \max_{\alpha\in\mathcal{A}} T_h(\alpha)$. Next, from (41) there

exists $K_\beta(\alpha, T_h)$, such that $\sup_{\beta\geq 0} M^{\alpha,\tau_h}(\beta)$ is equal to $\sup_{K_\beta(\alpha,T_h)\geq\beta\geq 0} M^{\alpha,\tau_h}(\beta)$. Again, this holds for all $\alpha \in \mathcal{A}$, thus there exists sufficiently large $K_\beta > 0$ such that (see also Lemma 33)

$$\min_{\substack{\alpha\in\mathcal{A}\\ \beta\geq 0}} \sup_{\tau_h>0} \inf_{\tau_g>0} \mathcal{D}(\alpha,\tau_g,\beta,\tau_h) = \inf_{\alpha\in\mathcal{A}} \sup_{\substack{\tau_h>0\\ K_\beta\geq\beta\geq 0}} \inf_{\tau_g>0} \mathcal{D}(\alpha,\tau_g,\beta,\tau_h)$$

The objective function $\mathcal{D}$ above is convex-concave. Also, the constraint sets over α and β are compact. Furthermore, the optimization of $\mathcal{D}$ over τ_g and τ_h is separable. With these and an application of Sion's minimax theorem, the order of inf–sup between the four optimization variables can be flipped arbitrarily without affecting the outcome. Thus, for example,

$$\inf_{\substack{\alpha\in\mathcal{A}\\ \beta\geq 0}} \sup_{\tau_h>0} \inf_{\tau_g>0} \mathcal{D}(\alpha,\tau_g,\beta,\tau_h) = \inf_{\substack{\alpha\in\mathcal{A}\\ \tau_g>0}} \sup_{\substack{\tau_h>0\\ \geq\beta\geq 0}} \mathcal{D}(\alpha,\tau_g,\beta,\tau_h)$$

The same is of course true for the corresponding random optimizations (also, Lemma 29(iii)).

Auxiliary Lemmas

Proof 17 (Proof of Lemma 31) *First, convexity is preserved by point wise limits, so that $F(x)$ is also convex. Using this and level-boundedness condition (b) of the lemma, it is easy to show that $\inf_{x>0} F(x) > -\infty$. Since F is proper and (lower) level-bounded, the only way $\inf_{x>0} F(x) = -\infty$ is if $\lim_{x\to 0} F(x) = -\infty$. But, this is not possible as follows: Fix $0 < x_1 < x_2 < x_3$. Then, for any $0 < x < x_1$ and $\theta := \frac{x_3-x_2}{x_3-x}$, convexity gives*

$$F(x) \geq \frac{1}{\theta}F(x_2) - \left(1-\frac{1}{\theta}\right)F(x_3) \geq -\frac{x_3-x_1}{x_3-x_2}|F(x_2)| - \frac{x_2-x_1}{x_3-x_2}|F(x_3)|.$$

Next, we show that for sufficiently small $\epsilon > 0$, there exist $x_0 > x_\epsilon > 0$:

$$\inf_{x>0} F(x) \leq F(x_\epsilon) \leq \inf_{x>0} F(x) + \epsilon \quad \text{and} \quad F(x_\epsilon) < F(x_0). \tag{52}$$

We show the claim for all $0 < \epsilon < \epsilon_1 := (F(z) - \inf_{x>0} F(x))$. Since $\inf_{x>0} F(x)$ is finite, there exists $x_\epsilon > 0$ such that $F(x_\epsilon) - \epsilon \leq \inf_{x>0} F(x)$. Without loss of generality, $x_\epsilon < z$. Pick any $x_0 > z$. For the shake of contradiction, assume $F(x_0) \leq F(x_\epsilon)$. Then, by convexity, for some $\theta \in (0,1)$

$$F(z) \leq \theta F(x_\epsilon) + (1-\theta)F(x_0) \leq F(x_\epsilon) \leq \inf_{x>0} F(x) + \epsilon < F(z).$$

Thus, $F(x_\epsilon) < F(x_0)$.

In order to establish the desired, it suffices that for all arbitrarily small $\delta > 0$*, w.p.a. 1,*

$$|\inf_{x>0} F_n(x) - \inf_{x>0} F(x)| < \delta. \tag{53}$$

Fix some $0 < \epsilon < \min\{\epsilon_1, \delta\}$ *such that* (52) *holds, and, also some*

$$0 < \epsilon' < \min\{(F(x_0) - F(x_\epsilon))/4, \delta/4, \delta - \epsilon\}. \tag{54}$$

Let $K = [a, b] \subset (0, \infty)$ *be compact subset such that* $a < x_\epsilon < x_0 \leq b$ *and* $a = \frac{\delta - 2\epsilon'}{2\delta - \epsilon'} x_\epsilon$ *. The functions* $\{F_p\}$ *are convex and they converge point wise to* F *in the open set* $(0, \infty)$*. This implies uniform convergence in compact sets by the Convexity lemma [7, Cor.. II.1] ,[107, Lem. 7.75]. That is, there exists sufficiently large* N_1 *such that the event*

$$\sup_{x \in K} |F_n(x) - F(x)| < \epsilon' \tag{55}$$

occurs w.p.a. 1, for all $p > N_1$*. In this event,*

$$\inf_{x>0} F_n(x) \leq F_n(x_\epsilon) < F(x_\epsilon) + \epsilon' \leq \inf_{x>0} F(x) + \epsilon + \epsilon' \leq \inf_{x>0} F(x) + \delta$$

It remains to prove the other side of (53)*. In what follows, take* $p \geq N_1$ *and condition on the high probability event in* (55)*.*

Let us first show level-boundedness of F_n*. Consider the event* $\inf_{x>x_0} F_n(x) < \inf_{x \leq x_0} F_n(x)$*. If this happens, then,* $\inf_{x>x_0} F_n(x) < F_n(x_\epsilon)$*, in which case there exists (by continuity of* F_n*),* $x_n > x_0$ *such that* $F_n(x_n) < F_n(x_\epsilon)$*. But then, convexity implies that for some* $0 < \theta_n < 1$*,*

$$F_n(x_0) \leq \theta_n F_n(x_n) + (1 - \theta_n) F_n(x_\epsilon) < F_n(x_\epsilon) \leq F(x_\epsilon) + \epsilon' < F(x_0) - \epsilon', \tag{56}$$

where we also used (55) *and* (54)*. Of course, this contradicts* (55)*. Thus,*

$$\inf_{x>0} F_n(x) = \inf_{x \leq x_0} F_n(x). \tag{57}$$

Using (57)*, convexity and properness of* $\{F_n\}$*, it can be shown that* $\inf_{x>0} F_n(x) > -\infty$*. The argument is the same as the one used in the beginning of the proof for* F*, thus is omitted for brevity.*

Overall, for all $p > N_1$, conditioned on (55)*, there is some $0 < x_n \leq x_0$ such that*

$$\inf_{x>0} F_n(x) \geq F_n(x_n) - \epsilon'. \tag{58}$$

If $a \leq x_n \leq b$, then a direct application of (55) *gives the desired*

$$F_n(x_n) \geq F(x_n) - \epsilon' \geq \inf_{x>0} F(x) - \epsilon' \Rightarrow \inf_{x>0} F_n(x) \geq \inf_{x>0} F(x) - 2\epsilon' \geq \inf_{x>0} F(x) - \delta.$$

Next, assume that $0 < x_n < a$. There exists $\theta_n \in (0,1)$ such that $\theta_n x_n + (1-\theta_n)x_\epsilon = a$. In fact,

$$\theta_n = \frac{x_\epsilon - a}{x_\epsilon - x_n} \geq (1 - a/x_\epsilon) = \frac{\delta - 2\epsilon'}{2\delta - \epsilon'}. \tag{59}$$

Then, by convexity and (55)*, $F_n(a) \leq \theta_n F_n(x_n) + (1-\theta_n)F_n(x_\epsilon)$. Rearranging and using* (55)

$$\begin{aligned} F_n(x_n) &\geq \frac{1}{\theta_n} F_n(a) - \frac{1-\theta_n}{\theta_n} F_n(x_\epsilon) \\ &\geq \frac{1}{\theta_n}(F(a) - \epsilon') - \frac{1-\theta_n}{\theta_n}(F(x_\epsilon) + \epsilon') \\ &\geq \frac{1}{\theta_n}\left(\inf_{x>0} F(x) - \epsilon\right) - \frac{1-\theta_n}{\theta_n}\left(\inf_{x>0} F(x) + \delta\right) \end{aligned}$$

Combining this with (58) *and* (59)*, yields the desired $\inf_{x>0} F_n(x_n) \geq \inf_{x>0} F(x) - \delta$.*

Lemma 32 *(Level-bounded convex fcns) Let $F : (0,\infty) \to \mathbb{R}$ be convex. Then, the following two statements are equivalent:*

1. *There exists $z > 0$ such that $F(x) > \inf_{x>0} F(x)$ for all $x \geq z$.*
2. *$\lim_{x\to\infty} F(x) = +\infty$.*

Proof 18 *(a)⇒(b): Clearly, there exists $0 < x_0 < z$, such that $F(z) > F(x_0)$. Then, by convexity, for all $x > z$ it holds*

$$F(x) \geq F(z) + \underbrace{\frac{F(z) - F(x_0)}{z - x_0}}_{>0}(x - z).$$

Taking limits of $x \to \infty$ on both sides above, proves the claim.

(a)⇐(b): A a proper functions, F has a nonempty domain in $(0,\infty)$. Hence, $\inf_{x>0} F(x) < \infty$ and can choose some $M > \inf_{x>0} F(x)$. From (b), there exists $z > 0$ such that $F(x) \geq M$ for all $x \geq z$, as desired.

Lemma 33 (Saddle-points) *For a convex-concave function $F : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, consider the minimax optimization $\inf_x \sup_y F(x, y)$. Let C, D be compact subsets such that there exists at least one saddle point $(x_*, y_*) \in C \times D$. Then,*

$$\inf_x \sup_y F(x, y) = \inf_{x \in C} \sup_{y \in D} F(x, y).$$

Proof 19 *First observe that,*

$$\inf_x \sup_y F(x, y) = \inf_{x \in C} \sup_y F(x, y)$$

Since F has a saddle-point, the LHS above is equal to $\sup_y \inf_x F(x, y)$ [136, Lem. 36.2]. Also, from Sion's minimax theorem, the RHS is equal to $\sup_y \inf_{x \in C} F(x, y)$. Thus, it suffices to prove that

$$\sup_y \inf_{x \in C} F(x, y) = \sup_{y \in D} \inf_{x \in C} F(x, y).$$

Clearly, this holds with a "$\geq$" sign. To prove equality, let (x_*, y_*) be a saddle point. Then,

$$\sup_y \inf_{x \in C} F(x, y) = \inf_{x \in C} \sup_y f(x, y) \leq \sup_y f(x_*, y) \leq f(x_*, y_*) = \sup_{y \in D} \inf_{x \in C} F(x, y).$$

Lemma 34 *The function $h(\alpha, \tau, \mathbf{v}) = \frac{1}{2\tau} \|\alpha \mathbf{x} + \mathbf{z} - \mathbf{v}\|_2^2$ is jointly convex in its arguments.*

Proof 20 *The function $\|\alpha \mathbf{x} - \mathbf{v}\|_2^2$ is trivially jointly convex in α and $\mathbf{v}$. So its perspective function which is $\frac{1}{\tau} \|\alpha \mathbf{x} - \mathbf{v}\|_2^2$ is also jointly convex in all its arguments, same as its shifted version which is $h(\alpha, \tau, \mathbf{v})$.*

Lemma 35 *Let $f : \mathbb{R}^p \to \mathbb{R}$ be convex. Then,*

1. $\operatorname{prox}_f(\mathbf{x}; \tau) + \tau \cdot \operatorname{prox}_{f^*}\left(\mathbf{x}/\tau; \tau^{-1}\right) = \mathbf{x}$,
2. $\mathrm{e}_f(\mathbf{x}; \tau) + \mathrm{e}_{f^*}(\mathbf{x}/\tau; 1/\tau) = \frac{\|\mathbf{x}\|^2}{2\tau}$.

.3 Proof of Auxiliary Lemmas

Proof of Lemma 10

First, convexity is preserved by point wise limits, so that $F(x)$ is also convex. Using this and level-boundedness condition (b) of the lemma, it is easy to show that $\inf_{x>0} F(x) > -\infty$. Since F is proper and (lower) level-bounded, the only way $\inf_{x>0} F(x) = -\infty$ is if $\lim_{x\to 0} F(x) = -\infty$. But, this is not possible as follows: Fix $0 < x_1 < x_2 < x_3$. Then, for any $0 < x < x_1$ and $\theta := \frac{x_3 - x_2}{x_3 - x}$, convexity gives

$$F(x) \geq \frac{1}{\theta} F(x_2) - \left(1 - \frac{1}{\theta}\right) F(x_3) \geq -\frac{x_3 - x_1}{x_3 - x_2}|F(x_2)| - \frac{x_2 - x_1}{x_3 - x_2}|F(x_3)|.$$

Next, we show that for sufficiently small $\epsilon > 0$, there exist $x_0 > x_\epsilon > 0$:

$$\inf_{x>0} F(x) \leq F(x_\epsilon) \leq \inf_{x>0} F(x) + \epsilon \quad \text{and} \quad F(x_\epsilon) < F(x_0). \tag{60}$$

We show the claim for all $0 < \epsilon < \epsilon_1 := (F(z) - \inf_{x>0} F(x))$. Since $\inf_{x>0} F(x)$ is finite, there exists $x_\epsilon > 0$ such that $F(x_\epsilon) - \epsilon \leq \inf_{x>0} F(x)$. Without loss of generality, $x_\epsilon < z$. Pick any $x_0 > z$. For the shake of contradiction, assume $F(x_0) \leq F(x_\epsilon)$. Then, by convexity, for some $\theta \in (0, 1)$

$$F(z) \leq \theta F(x_\epsilon) + (1 - \theta) F(x_0) \leq F(x_\epsilon) \leq \inf_{x>0} F(x) + \epsilon < F(z).$$

Thus, $F(x_\epsilon) < F(x_0)$.

In order to establish the desired, it suffices that for all arbitrarily small $\delta > 0$, w.p.a. 1,

$$|\inf_{x>0} F_n(x) - \inf_{x>0} F(x)| < \delta. \tag{61}$$

Fix some $0 < \epsilon < \min\{\epsilon_1, \delta\}$ such that (60) holds, and, also some

$$0 < \epsilon' < \min\{(F(x_0) - F(x_\epsilon))/4, \delta/4, \delta - \epsilon\}. \tag{62}$$

Let $K = [a, b] \subset (0, \infty)$ be compact subset such that $a < x_\epsilon < x_0 \leq b$ and $a = \frac{\delta - 2\epsilon'}{2\delta - \epsilon'} x_\epsilon$. The functions $\{F_n\}$ are convex and they converge point wise to F in the open set $(0, \infty)$. This implies uniform convergence in compact sets by the Convexity lemma ,[107, Lem. 7.75]. That is, there exists sufficiently large N_1 such that the event

$$\sup_{x \in K} |F_n(x) - F(x)| < \epsilon' \tag{63}$$

occurs w.p.a. 1, for all $n > N_1$. In this event,

$$\inf_{x>0} F_n(x) \leq F_n(x_\epsilon) < F(x_\epsilon) + \epsilon' \leq \inf_{x>0} F(x) + \epsilon + \epsilon' \leq \inf_{x>0} F(x) + \delta \tag{64}$$

It remains to prove the other side of (61). In what follows, take $n \geq N_1$ and condition on the high probability event in (63).

Let us first show level-boundedness of F_n. Consider the event $\inf_{x>x_0} F_n(x) < \inf_{x\leq x_0} F_n(x)$. If this happens, then, $\inf_{x>x_0} F_n(x) < F_n(x_\epsilon)$, in which case there exists (by continuity of F_n), $x_n > x_0$ such that $F_n(x_n) < F_n(x_\epsilon)$. But then, convexity implies that for some $0 < \theta_n < 1$,

$$F_n(x_0) \leq \theta_n F_n(x_n) + (1-\theta_n) F_n(x_\epsilon) < F_n(x_\epsilon) \leq F(x_\epsilon) + \epsilon' < F(x_0) - \epsilon', \tag{65}$$

where we also used (63) and (62). Of course, this contradicts (63). Thus,

$$\inf_{x>0} F_n(x) = \inf_{x\leq x_0} F_n(x). \tag{66}$$

Using (66), convexity and properness of $\{F_n\}$, it can be shown that $\inf_{x>0} F_n(x) > -\infty$. The argument is the same as the one used in the beginning of the proof for F, thus is omitted for brevity.

Overall, for all $n > N_1$, conditioned on (63), there is some $0 < x_n \leq x_0$ such that

$$\inf_{x>0} F_n(x) \geq F_n(x_n) - \epsilon'. \tag{67}$$

If $a \leq x_n \leq b$, then a direct application of (63) gives the desired

$$F_n(x_n) \geq F(x_n) - \epsilon' \geq \inf_{x>0} F(x) - \epsilon' \Rightarrow \inf_{x>0} F_n(x) \geq \inf_{x>0} F(x) - 2\epsilon' \geq \inf_{x>0} F(x) - \delta.$$

Next, assume that $0 < x_n < a$. There exists $\theta_n \in (0,1)$ such that $\theta_n x_n + (1-\theta_n) x_\epsilon = a$. In fact,

$$\theta_n = \frac{x_\epsilon - a}{x_\epsilon - x_n} \geq (1 - a/x_\epsilon) = \frac{\delta - 2\epsilon'}{2\delta - \epsilon'}. \tag{68}$$

Then, by convexity and (63), $F_n(a) \leq \theta_n F_n(x_n) + (1-\theta_n) F_n(x_\epsilon)$. Rearranging and using (63)

$$\begin{aligned} F_n(x_n) &\geq \frac{1}{\theta_n} F_n(a) - \frac{1-\theta_n}{\theta_n} F_n(x_\epsilon) \\ &\geq \frac{1}{\theta_n} (F(a) - \epsilon') - \frac{1-\theta_n}{\theta_n} (F(x_\epsilon) + \epsilon') \\ &\geq \frac{1}{\theta_n} \left(\inf_{x>0} F(x) - \epsilon \right) - \frac{1-\theta_n}{\theta_n} \left(\inf_{x>0} F(x) + \delta \right) \end{aligned}$$

Combining this with (67) and (68), yields the desired $\inf_{x>0} F_n(x_n) \geq \inf_{x>0} F(x) - \delta$.

Lemma 36 *(Level-bounded convex fcns) Let $F : (0, \infty) \to \mathbb{R}$ be convex. Then, the following two statements are equivalent:*

1. *There exists $z > 0$ such that $F(x) > \inf_{x>0} F(x)$ for all $x \geq z$.*
2. *$\lim_{x\to\infty} F(x) = +\infty$.*

Proof 21 *(a)⇒(b): Clearly, there exists $0 < x_0 < z$, such that $F(z) > F(x_0)$. Then, by convexity, for all $x > z$ it holds*

$$F(x) \geq F(z) + \underbrace{\frac{F(z) - F(x_0)}{z - x_0}}_{>0}(x - z).$$

Taking limits of $x \to \infty$ on both sides above, proves the claim.

(a)⇐(b): A a proper functions, F has a nonempty domain in $(0, \infty)$. Hence, $\inf_{x>0} F(x) < \infty$ and can choose some $M > \inf_{x>0} F(x)$. From (b), there exists $z > 0$ such that $F(x) \geq M$ for all $x \geq z$, as desired.

Lemma 37 (Saddle-points) *For a convex-concave function $F : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, consider the minimax optimization $\inf_x \sup_y F(x, y)$. Let C, D be compact subsets such that there exists at least one saddle point $(x_*, y_*) \in C \times D$. Then,*

$$\inf_x \sup_y F(x, y) = \inf_{x\in C} \sup_{y\in D} F(x, y).$$

Proof 22 *First observe that,*

$$\inf_x \sup_y F(x, y) = \inf_{x\in C} \sup_y F(x, y)$$

Since F has a saddle-point, the LHS above is equal to $\sup_y \inf_x F(x, y)$ [136, Lem. 36.2]. Also, from Sion's minimax theorem, the RHS is equal to $\sup_y \inf_{x\in C} F(x, y)$. Thus, it suffices to prove that

$$\sup_y \inf_{x\in C} F(x, y) = \sup_{y\in D} \inf_{x\in C} F(x, y).$$

Clearly, this holds with a "$\geq$" sign. To prove equality, let (x_*, y_*) be a saddle point. Then,

$$\sup_y \inf_{x\in C} F(x, y) = \inf_{x\in C} \sup_y f(x, y) \leq \sup_y f(x_*, y) \leq f(x_*, y_*) = \sup_{y\in D} \inf_{x\in C} F(x, y).$$

Lemma 38 *The function* $h(\alpha, \tau, \mathbf{v}) = \frac{1}{2\tau}\|\alpha\mathbf{x} + \mathbf{z} - \mathbf{v}\|_2^2$ *is jointly convex in its arguments.*

Proof 23 *The function* $\|\alpha\mathbf{x} - \mathbf{v}\|_2^2$ *is trivially jointly convex in* α *and* $\mathbf{v}$*. So its perspective function which is* $\frac{1}{\tau}\|\alpha\mathbf{x} - \mathbf{v}\|_2^2$ *is also jointly convex in all its arguments, same as its shifted version which is* $h(\alpha, \tau, \mathbf{v})$*.*

Lemma 39 *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be convex. Then,*

1. $\text{prox}_f(\mathbf{x}; \tau) + \tau \cdot \text{prox}_{f^*}\left(\mathbf{x}/\tau; \tau^{-1}\right) = \mathbf{x}$,
2. $\text{e}_f(\mathbf{x}; \tau) + \text{e}_{f^*}(\mathbf{x}/\tau; 1/\tau) = \frac{\|\mathbf{x}\|^2}{2\tau}$.

Properties of the Moreau envelope:

In this section we have gathered some very useful properties of Moreau envelopes of convex functions. We have made heavy use of those results for the proofs of Theorem (1) and (2). Some of the results are standard, while others are more tailored towards our interests.

Lemma 40 (Properties of the Moreau envelope) *Let* $\ell : \mathbb{R} \to \mathbb{R}$ *be a proper, closed, convex function. For* $\tau > 0$*, consider its Moreau envelope function and its proximal operator:*

$$\text{e}_\ell(\chi; \tau) := \min_v \frac{1}{2\tau}(\chi - v)^2 + \ell(v), \tag{69a}$$

$$\text{prox}_\ell(\chi; \tau) := \arg\min \frac{1}{2\tau}(\chi - v)^2 + \ell(v) \tag{69b}$$

The following statements are true:

1. $\text{prox}_\ell(\chi; \tau)$ *is single valued and continuous. Furthermore,*

$$\ell'_{\chi,\tau} := \frac{1}{\tau}(\chi - \text{prox}_\ell(\chi; \tau)) \in \partial\ell(\text{prox}_\ell(\chi; \tau)). \tag{70}$$

2. $\text{e}_\ell(\chi; \tau)$ *is jointly convex in* (χ, τ)*.*

3. $\text{e}_\ell(\chi; \tau)$ *is continuously differentiable with respect to both* x *and* τ*. The gradients are given by:*

$$E_1(\chi, \tau) := \frac{\partial e_\ell}{\partial \chi} = \frac{1}{\tau}(\chi - \text{prox}_\ell(\chi; \tau)) = \ell'_{\chi,\tau}, \tag{71}$$

$$E_2(\chi, \tau) := \frac{\partial e_\ell}{\partial \tau} = -\frac{1}{2\tau^2}(\chi - \text{prox}_\ell(\chi; \tau))^2 = -\frac{1}{2}\left(\ell'_{\chi,\tau}\right)^2. \tag{72}$$

4. *Fix χ and $\tau > 0$. Consider the function $\Delta : \mathbb{R} \times (-\tau, \infty) \to \mathbb{R}$:*

$$\Delta(x, y) := (E_1(\chi + x, \tau + y) - E_1(\chi, \tau))x + (E_2(\chi + x, \tau + y) - E_2(\chi, \tau))y$$

Then,

$$\Delta(x, y) \geq \left(\tau + \frac{y}{2}\right)(\ell'_{\chi+x,\tau+y} - \ell'_{\chi,\tau})^2. \tag{73}$$

5. $\mathrm{e}_\ell(x; \tau)$ *is non-increasing in* τ.

6. $\lim_{\tau\to\infty} \mathrm{e}_\ell(x; \tau) = \min_v \ell(v)$.

7. $\lim_{\tau\to\infty} \frac{1}{\tau}|x - \mathrm{prox}_\ell(x; \tau)| = 0$.

8. *If* $0 \in \arg\min_v \ell(v)$, *then* $\mathrm{prox}_\ell(x; \tau)\, x \geq 0$, $|\mathrm{prox}_\ell(x; \tau)| \leq |x|$ *and* $|\ell'_{\mathrm{prox}_\ell(x;\tau),\tau}| \leq |\ell'_{x,\tau}|$.

9. $\mathrm{e}_\ell(x_n; \tau_n) \to \ell(x)$ *whenever* $x_n \to x$ *while* $\tau_n \to 0^+$ *in such a way that the sequence* $\{|x_n - x|/\tau_n\}_{n\in\mathbb{N}}$ *is bounded.*

Proof 24 *(i) From [137, Thm. 2.26(a)],* $\mathrm{prox}_\ell(\chi; \tau)$ *is known to be continuous single valued mapping. Besides, from standard optimality conditions:*

$$\frac{1}{\tau}(\chi - \mathrm{prox}_\ell(\chi; \tau)) \in \partial\ell(\mathrm{prox}_\ell(\chi; \tau)) \tag{74}$$

For convenience, we have define $\ell'_{\chi,\tau} := \frac{1}{\tau}(\chi - \mathrm{prox}_\ell(\chi; \tau) \in \partial\ell(\mathrm{prox}_\ell(\chi; \tau))$. *Note that if* ℓ *is differentiable at* $\mathrm{prox}_\ell(\chi; \tau)$, *then* $\ell'_{\chi,\tau}$ *is the derivative of* ℓ *at that point.*

(ii) Trivially, $h(\chi, v) := (\chi - v)^2$ *is a jointly convex function of* v *and* x. *Thus, its perspective function* $\tau h(\frac{\chi}{\tau}, \frac{v}{\tau}) = \frac{1}{\tau}(\chi - v)^2$ *is also jointly convex over* τ, x *and* v *and so after minimization over* v, *the function remains jointly convex over* x *and* τ *(cf. [137, Prop. 2.22]).*

(iii) See [137, Thm. 2.26(b)] for differentiability with respect to x. *Next, we mimic the argument to conclude about differentiability with respect to* τ. *It suffices to show that* $h(y) := \mathrm{e}_\ell(\chi; \tau + y) - \mathrm{e}_\ell(\chi; \tau) + \frac{y}{2\tau^2}(\chi - \mathrm{prox}_\ell(\chi; \tau))^2$ *is differentiable at* $y = 0$ *with* $\frac{\partial h}{\partial y} = 0$. *We know* $\mathrm{e}_\ell(\chi; \tau) = \frac{1}{2\tau}(\chi - \mathrm{prox}_\ell(\chi; \tau))^2 + \ell(\mathrm{prox}_\ell(\chi; \tau))$, *whereas* $\mathrm{e}_\ell(\chi; \tau + y) \leq \frac{1}{2(\tau+y)}(\chi - \mathrm{prox}_\ell(\chi; \tau))^2 + \ell(\mathrm{prox}_\ell(\chi; \tau))$. *Thus,*

$$\begin{aligned} h(y) &\leq \frac{1}{2(\tau + y)}(\chi - \mathrm{prox}_\ell(\chi; \tau))^2 - \frac{1}{2\tau}(\chi - \mathrm{prox}_\ell(\chi; \tau))^2 + \frac{y}{2\tau^2}(\chi - \mathrm{prox}_\ell(\chi; \tau))^2 \\ &= \frac{y^2}{2\tau^2(\tau + y)}(\chi - \mathrm{prox}_\ell(\chi; \tau))^2. \end{aligned} \tag{75}$$

Besides because of convexity of $h(y)$, $0 = h(0) \leq \frac{1}{2}h(y) + \frac{1}{2}h(-y)$ *or equivalently* $h(y) \geq -h(-y)$. *Thus,* (75) *gives:*

$$h(y) \geq \frac{y^2}{2\tau^2(\tau - y)}(\chi - \text{prox}_\ell(\chi;\tau))^2. \tag{76}$$

Combining (75) *and* (76) *leads to the following*

$$\frac{y^2}{2\tau^2(\tau - y)}(\chi - \text{prox}_\ell(\chi;\tau))^2 \leq h(y) \leq \frac{y^2}{2\tau^2(\tau + y)}(\chi - \text{prox}_\ell(\chi;\tau))^2 \tag{77}$$

Here, $h(y)$ *is sandwiched between two continuously differentiable functions at 0 with zero derivatives. This completes the proof.*

(iv) From (71) *and* (72)*, we have*

$$\Delta(x,y) = (\ell'_{\chi+x,\tau+y} - \ell'_{\chi,\tau})x - \left(\ell'^2(\chi + x, \tau + y) - \ell'^2(\chi,\tau)\right)\frac{y}{2} \tag{78}$$

$$= (\ell'_{\chi+x,\tau+y} - \ell'_{\chi,\tau})\left(x - \frac{y}{2}\left(\ell'_{\chi+x,\tau+y} + \ell'_{\chi,\tau}\right)\right). \tag{79}$$

On the other hand, due to optimality conditions in (70)*,*

$$\begin{aligned}\text{prox}_\ell(\chi + x; \tau + y) - \text{prox}_\ell(\chi;\tau) &= x - (\tau + y)\ell'_{\chi+x,\tau+y} + \tau\ell'_{\chi,\tau} \\ &= \left(x - \frac{y}{2}(\ell'_{\chi+x,\tau+y} + \ell'_{\chi,\tau})\right) - (\tau + \frac{y}{2})(\ell'_{\chi+x,\tau+y} - \ell'_{\chi,\tau}).\end{aligned} \tag{80}$$

Finally, from convexity of ℓ*, it follows from the monotonicity property of the subdifferential that*

$$(\ell'_{\chi+x,\tau+y} - \ell'_{\chi,\tau})(\text{prox}_\ell(\chi + x; \tau + y) - \text{prox}_\ell(\chi;\tau)) \geq 0.$$

Combining the three displays above gives the desired inequality.

(v) This follows directly by non-positivity of the derivative as in (72)*.*

(vi) Using the decreasing nature of $\text{e}_\ell(x;\tau)$ *w.r.t.* τ*, we have*

$$\lim_{\tau\to\infty} \text{e}_\ell(x;\tau) = inf_{\tau>0} \min_v \frac{1}{2\tau}(x-v)^2 + \ell(v) = \min_v inf_{\tau>0} \frac{1}{2\tau}(x-v)^2 + \ell(v) = \min_v \ell(v). \tag{81}$$

(vii) Fix an $\epsilon > 0$*. Since* $\lim_{\tau\to\infty} \text{e}_\ell(x;\tau) = \min_v \ell(v)$*, there exist* T'_ϵ *such that for all* $\tau \geq T_\epsilon := \max\{2, T'_\epsilon\}$*,*

$$|\text{e}_\ell(x;\tau) - \min_v \ell(v)| = \frac{1}{2\tau}(x - \text{prox}_\ell(x;\tau))^2 + (\ell(\text{prox}_\ell(x;\tau)) - \min_v \ell(v)) < \epsilon^2 \tag{82}$$

Then, $\frac{1}{2\tau}(x - \text{prox}_\ell\,(x;\tau))^2 < \epsilon^2$*, which gives*

$$\frac{1}{\tau}|x - \text{prox}_\ell\,(x;\tau)\,| < \epsilon\sqrt{\frac{2}{\tau}} \le \epsilon\sqrt{\frac{2}{T_\epsilon}} \le \epsilon \tag{83}$$

Therefore, $\lim_{\tau\to\infty} \frac{1}{\tau}|x - \text{prox}_\ell\,(x;\tau)\,| = 0$.

(viii) By (70) *and the assumption* $0 \in \arg\min_v \ell(v)$*, we find* $\text{prox}_\ell\,(0;\tau) = 0$*. Monotonicity of the prox operator [137, Prop. 12.19], gives* $(\text{prox}_\ell\,(x;\tau) - \text{prox}_\ell\,(0;\tau))x \ge 0$*, which then shows* $\text{prox}_\ell\,(x;\tau)\,x \ge 0$*. Also, monotonicity of the subdifferential of* ℓ *gives* $\ell'_{x,\tau}x \ge 0$*. Those two, when combined with optimality conditions in* (70) *give*

$$x - \text{prox}_\ell\,(x;\tau) = \tau\ell'_{x,\tau} \implies x^2 \ge \text{prox}_\ell\,(x;\tau)\,x \implies |x| \ge |\text{prox}_\ell\,(x;\tau)\,x|.$$

It remains to show that $\max_{s\in\partial\ell(\text{prox}_\ell(x;\tau))}|s| \le \max_{s\in\partial\ell(x)}|s|$*. Since* $0 \in \arg\min_v \ell(v)$*, it follows by convexity that*

$$(0 \le x_1 \le x_2 \;\; or \;\; x_2 \le x_1 \le 0) \implies \max_{s\in\partial\ell(x_1)}|s| \le \max_{s\in\partial\ell(x_2)}|s|$$

Observe that the LHS of the implication above is equivalent to $(|x_2| \ge |x_1| \;\; and \;\; x_1x_2 \ge 0)$*. Then apply it for* $x_1 = \text{prox}_\ell\,(x;\tau)$ *and* $x_2 = x$*, to conclude.*

(ix) Please see [137, Thm. 1.25].

On Remark 14

Substituting the envelope function of $|\cdot|$ in (2.40) gives:

$$\frac{\beta}{2} + \bar{s}\mathbb{E}\begin{cases} -\frac{\beta(\alpha G+Z)^2}{2\tau^2} & , |\alpha G + Z| \le \frac{\tau}{\beta} \\ -\frac{1}{2\beta} & , \text{otherwise} \end{cases} + (\delta - \bar{s})\mathbb{E}\begin{cases} -\frac{\beta\alpha^2G^2}{2\tau^2} & , |\alpha G| \le \frac{\tau}{\beta} \\ -\frac{1}{2\beta} & , \text{otherwise} \end{cases} \ge 0, \tag{84a}$$

$$\bar{s}\mathbb{E}\begin{cases} \frac{\beta G(\alpha G+Z)}{2} & , |\alpha G + Z| \le \frac{\tau}{\beta} \\ G\;\text{sign}(\alpha G + Z) & , \text{otherwise} \end{cases} + (\delta - \bar{s})\mathbb{E}\begin{cases} \frac{\alpha G^2\beta}{\tau} & , |\alpha G\| \le \frac{\tau}{\beta} \\ G\;\text{sign}(G) & , \text{otherwise} \end{cases} - \beta\sqrt{\overline{D}_{\mathcal{K}}} \ge 0, \tag{84b}$$

$$\frac{\tau}{2} + \bar{s}\mathbb{E}\begin{cases} \frac{(\alpha G+Z)^2}{2\tau} & , |\alpha G + Z| \le \frac{\tau}{\beta} \\ -\frac{\tau}{2} & , \text{otherwise} \end{cases} + (\delta - \bar{s})\mathbb{E}\begin{cases} \frac{\alpha^2G^2}{2\tau} & , |\alpha G| \le \frac{\tau}{\beta} \\ -\frac{\tau}{2} & , \text{otherwise} \end{cases} - \alpha\sqrt{\overline{D}_{\mathcal{K}}} \le 0. \tag{84c}$$

Define $\kappa := \frac{\tau}{\beta\alpha}$ and $\rho := \frac{\tau}{\beta}$. In order to find a sufficient condition for α to be zero, we assume $\alpha \to 0$, $\tau \to 0$, $\rho \to 0$ and $\kappa \ge 0$ and look for conditions under which

the equations in (84) are consistent. Under these assumptions, one can check that (84c) is satisfied (the argument converges to zero), while, (84b) and (84a) become

$$2\frac{(\delta-\bar{s})}{\kappa}\int_0^{\kappa} G^2\phi(G)\mathrm{d}G + (\delta-\bar{s})\int_{\kappa}^{\infty} G\phi(G)\mathrm{d}G \geq \beta\sqrt{\overline{D}_{\mathcal{K}}}, \tag{85a}$$

$$\beta^2 \geq \bar{s} + 2\frac{\delta-\bar{s}}{\kappa^2}\int_0^{\kappa} G^2\phi(G)\mathrm{d}G + 2(\delta-\bar{s})\int_{\kappa}^{\infty}\phi(G)\mathrm{d}G, \tag{85b}$$

where $\phi(G) = e^{-G^2/2}/\sqrt{2\pi}$ and we multiplied (84a) by β^2 to get (85b). Observe that (85a) upper bounds β while (85b) derives a lower bound on it. Thus, consistency of the set of equations (85) is achieved if the following holds:

$$\frac{1}{\overline{D}_{\mathcal{K}}}(2\frac{(\delta-\bar{s})}{\kappa}\int_0^{\kappa} G^2\phi(G)\mathrm{d}G + (\delta-\bar{s})\int_{\kappa}^{\infty} G\phi(G)\mathrm{d}G)^2$$
$$\geq \bar{s} + 2\frac{\delta-\bar{s}}{\kappa 2}\int_0^{\kappa} G^2\phi(G)\mathrm{d}G + 2(\delta-\bar{s})\int_{\kappa}^{\infty}\phi(G)\mathrm{d}G.$$

Or, equivalently,

$$\overline{D}_{\mathcal{K}} \leq \frac{(2\frac{(\delta-\bar{s})}{\kappa}\int_0^{\kappa} G^2\phi(G)\mathrm{d}G + (\delta-\bar{s})\int_{\kappa}^{\infty} G\phi(G)\mathrm{d}G)^2}{\bar{s} + 2\frac{\delta-\bar{s}}{\kappa^2}\int_0^{\kappa} G^2\phi(G)\mathrm{d}G + 2(\delta-\bar{s})\int_{\kappa}^{\infty}\phi(G)\mathrm{d}G}. \tag{86}$$

Thus if maximum of the right side of (86) with respect to κ is greater than $\overline{D}_{\mathcal{K}}$, all our variables satisfy (2.40) and the optimal value in (2.39) occurs when $\alpha \to 0$, $\tau \to 0$ and $\frac{\tau}{\alpha\beta} \to \kappa$ which means $\alpha_* = 0$. We will show that

$$\max_{\kappa>0} \frac{(2\frac{(\delta-\bar{s})}{\kappa}\int_0^{\kappa} G^2\phi(G)\mathrm{d}G + (\delta-\bar{s})\int_{\kappa}^{\infty} G\phi(G)\mathrm{d}G)^2}{\bar{s} + 2\frac{\delta-\bar{s}}{\kappa^2}\int_0^{\kappa} G^2\phi(G)\mathrm{d}G + 2(\delta-\bar{s})\int_{\kappa}^{\infty}\phi(G)\mathrm{d}G}$$
$$\geq \delta - \min_{\kappa>0}\bar{s}(1+\kappa^2) + 2(\delta-\bar{s})\int_{\kappa}^{\infty}(G-\kappa)^2\phi(G)\mathrm{d}G \tag{87}$$

If both this and (2.45) are true then, there will be a κ for which (86) holds and as we discussed, this implies $\alpha_* = 0$.

For convenience, we define $A_\kappa = \int_{\kappa}^{\infty} G^2\phi(G)\mathrm{d}G$, $B_\kappa = \int_{\kappa}^{\infty} G\phi(G)\mathrm{d}G$ and $C_\kappa = \int_{\kappa}^{\infty}\phi(G)\mathrm{d}G$. The optimal κ for the right side of (87) satisfies the following due to the first optimality condition

$$2(\delta-\bar{s})\hat{\kappa}B_{\hat{\kappa}} - 2(\delta-\bar{s})\hat{\kappa}^2 C_{\hat{\kappa}} = \hat{\kappa}^2\bar{s} \tag{88}$$

For this value of κ, the left side of (87) becomes

$$\frac{(2\frac{(\delta-\bar{s})}{\hat{\kappa}}\int_0^{\hat{\kappa}} G^2\phi(G)\,dG + (\delta-\bar{s})\int_{\hat{\kappa}}^{\infty} G\phi(G)\,dG)^2}{\bar{s} + 2\frac{\delta-\bar{s}}{\hat{\kappa}^2}\int_0^{\hat{\kappa}} G^2\phi(G)\,dG + 2(\delta-\bar{s})\int_{\hat{\kappa}}^{\infty}\phi(G)\,dG} = (\delta-\bar{s})(1-2A_{\hat{\kappa}}+2\hat{\kappa}B_{\hat{\kappa}})$$
$$= \delta - \bar{s} - 2(\delta-\bar{s})\hat{\kappa}B_{\hat{\kappa}} + 2(\delta-\bar{s})\hat{\kappa}^2C_{\hat{\kappa}} - 2(\delta-\bar{s})A_{\hat{\kappa}} + 4(\delta-\bar{s})\hat{\kappa}B_{\hat{\kappa}} - 2(\delta-\bar{s})\hat{\kappa}^2C_{\hat{\kappa}}$$
$$= \delta - \bar{s}(1+\hat{\kappa}^2) - 2(\delta-\bar{s})(A_{\hat{\kappa}} - 2B_{\hat{\kappa}} + \hat{\kappa}^2C_{\hat{\kappa}}) = \delta - \bar{s}(1+\hat{\kappa}^2)$$
$$+ 2(\delta-\bar{s})\int_{\hat{\kappa}}^{\infty}(G-\hat{\kappa})^2\phi(G)\,dG,$$

where the first and third equalities follow after substituting $\bar{s}$ using (88). This proves (87) as desired to conclude the claim of the remark.

Satisfying Assumption 1(a) on Section 2.1

It only takes a few calculations to show that

$$\frac{1}{n}\mathrm{e}_{\sqrt{p}\|\cdot\|_2}(\alpha\mathbf{g}+\mathbf{z};\tau) = \begin{cases} \frac{1}{\sqrt{n\delta}}\|\alpha\mathbf{g}+\mathbf{z}\|_2 - \frac{\tau}{2\delta} & \text{, if } \frac{\sqrt{\delta}\|\alpha\mathbf{g}+\mathbf{z}\|_2}{\sqrt{n}} \geq \tau, \\ \frac{1}{2\tau}\frac{\|\alpha\mathbf{g}+\mathbf{z}\|_2^2}{n} & \text{, otherwise.} \end{cases} \tag{89}$$

Assume that $0 < \mathbb{E}\frac{\|\mathbf{z}\|_2^2}{n} =: \sigma^2 < \infty$. From (89), it can be seen that $\frac{1}{n}\mathrm{e}_{\sqrt{p}\|\cdot\|_2}(\alpha\mathbf{g}+\mathbf{z};\tau)$ is a Lipschitz convex function of $\frac{\|\alpha\mathbf{g}+\mathbf{z}\|}{\sqrt{n}}$. Also, $\frac{\|\alpha\mathbf{g}+\mathbf{z}\|}{\sqrt{n}}$ converges in probability to $\sqrt{\alpha^2+\sigma^2}$, thus

$$\frac{1}{n}(\mathrm{e}_{\sqrt{p}\|\cdot\|_2}(\alpha\mathbf{g}+\mathbf{z};\tau) - \|\mathbf{z}\|_2\sqrt{p}) \to L(\alpha,\tau) = \begin{cases} \frac{\sqrt{\alpha^2+\sigma^2}-\sigma}{\sqrt{\delta}} - \frac{\tau}{2\delta} & \text{, if } \delta(\alpha^2+\sigma^2) \geq \tau^2, \\ \frac{1}{2\tau}(\alpha^2+\sigma^2) - \frac{\sigma}{\sqrt{\delta}} & \text{, otherwise.} \end{cases}$$