Computational and Statistical Aspects of High-Dimensional Structured Estimation

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Sheng Chen

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Prof. Arindam Banerjee

May, 2018

Acknowledgements

There are many people that have earned my gratitude for their contribution to my six-year PhD life.

First I would like to express my sincerest gratitude and appreciation to my advisor Prof. Arindam Banerjee, for his guidance, support and encouragement. He is so knowledgeable that every discussion with him was thought-provoking. He is also passionate about delving into technical details, which has inspired several threads of my research. Besides I have also benefited from his extraordinary skills on writing, presentation and communication. Overall I was extremely fortunate to work with Prof. Banerjee during the past few years.

Second, I am deeply indebted to Prof. Rui Kuang, who has guided me through the initial years in graduate study. His tremendous patience and thoughtful advice helped me start the PhD career in a very positive way. Without his support, there would be more obstacles and difficulties in pursuing my PhD degree. I am also grateful to Prof. Daniel Boley, Prof. George Karypis, and Prof. Jarvis Haupt for serving as my dissertation committee and for their helpful suggestions and feedbacks.

Third, I would like to extend my great thanks to my supervisors and colleagues while interning at Yahoo! Research in 2016, including Troy Chevaliar, Nikolay Laptev, Tina Liu, Yashar Mehdad, Aasish Pappu, Rao Shen, Akshay Soni, and Kapil Thadani. They have opened a door for me so that I could learn the cutting-edge technologies used in industry.

Last but not the least, I would also like to show my gratitude to my lab mates and fellow graduate students: Soumyadeep Chatterjee, Konstantina Christakopoulou, Chintan Dalal, Miao Fan, Farideh Fazayeli, Robert Giaquinto, Hardik Goel, Andre Goncalves, Qilong Gu, Sijie He, Nicholas Johnson, Xinyan Li, Xiaoli Liu, Igor Melnyk, Dave Roe, Vidyashankar Sivakumar, Amir Taheri, Shaozhe Tao, Huahua Wang, Huanan Zhang, Wei Zhang and Yingxue Zhou. It has always been enjoyable and fruitful to discuss and work with them.

The research in this thesis was supported in part by NSF grants IIS-1563950, IIS-1447566, IIS-1447574, IIS-1422557, CCF-1451986, CNS- 1314560, IIS-0953274, IIS-1029711, NASA grant NNX12AQ39A, and gifts from Adobe, IBM, and Yahoo.

Dedication

To God and my family

Abstract

Modern statistical learning often faces high-dimensional data, for which the number of features that should be considered is very large. In consideration of various constraints encountered in data collection, such as cost and time, however, the available samples for applications in certain domains are of small size compared with the feature sets. In this scenario, statistical estimation becomes much more challenging than in the large-sample regime. Since the information revealed by small samples is inadequate for finding the optimal model parameters, the estimator may end up with incorrect models that appear to fit the observed data but fail to generalize to unseen ones. Owning to the prior knowledge about the underlying parameters, additional structures can be imposed to effectively reduce the parameter space, in which it is easier to identify the true one with limited data. This simple idea has inspired the study of high-dimensional statistics since its inception.

Over the last two decades, sparsity has been one of the most popular structures to exploit when we estimate a high-dimensional parameter, which assumes that the number of nonzero elements in parameter vector/matrix is much smaller than its ambient dimension. For simple scenarios such as linear models, L_1-norm based convex estimators like Lasso and Dantzig selector, have been widely used to find the true parameter with reasonable amount of computation and provably small error. Recent years have also seen a variety of structures proposed beyond sparsity, e.g., group sparsity and low-rankness of matrix, which are demonstrated to be useful in many applications. On the other hand, the aforementioned estimators can be extended to leverage new types of structures by finding appropriate convex surrogates like the L_1 norm for sparsity. Despite their success on individual structures, current developments towards a unified

understanding of various structures are still incomplete in both computational and statistical aspects. Moreover, due to the nature of the model or the parameter structure, the associated estimator can be inherently non-convex, which may need additional care when we consider such unification of different structures.

In this thesis, we aim to make progress towards a unified framework for the estimation with general structures, by studying the high-dimensional structured linear model and other semi-parametric and non-convex extensions. In particular, we introduce the generalized Dantzig selector (GDS), which extends the original Dantzig selector for sparse linear models. For the computational aspect, we develop an efficient optimization algorithm to compute the GDS. On statistical side, we establish the recovery guarantees of GDS using certain *geometric measures*. Then we demonstrate that those geometric measures can be bounded by utilizing simple information of the structures. These results on GDS have been extended to the matrix setting as well. Apart from the linear model, we also investigate one of its semi-parametric extension – the single-index model (SIM). To estimate the true parameter, we incorporate its structure into two types of simple estimators, whose estimation error can be established using similar geometric measures. Besides we also design a new semi-parametric model called sparse linear isotonic model (SLIM), for which we provide an efficient estimation algorithm along with its statistical guarantees. Lastly, we consider the non-convex estimation for structured multi-response linear models. We propose an alternating estimation procedure to estimate the parameters. In spite of dealing with non-convexity, we show that the statistical guarantees for general structures can be also summarized by the geometric measures.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In recent years, data-driven approaches have gained unprecedented popularity in a wide range of disciplines, such as social science, linguistics, healthcare and finance, to name a few. Numerous applications of data analysis have greatly impacted our daily life. For example, useful patterns and information are extracted from data to help people make decisions (e.g., disease diagnosis [147], portfolio selection [103] and product recommendation [142]). Emerging intelligent systems trained using massive data, like voice assistant and autonomous vehicles, can emancipate people from time-consuming or tedious tasks. Moreover, the recent victory of the AlphaGo [149] against the top human go players has created a tremendous sensation, registering a peak of "Big Data".

The success of data science critically relies on the methodology developed in machine learning and statistics. To harness the power of data, many statistical models have been proposed to describe intrinsic structures hidden in the data, and searching for the model that best explains the collected data often requires the estimation of the model parameters. Classical statistical machine learning typically deals with data arising in the low dimension, meaning that the number of features/predictors is relatively small, for which the model estimation can be performed with moderate amount of data [101]. In recent years, however, high-dimensional data are frequently encountered in practice [27], where one has to consider a large set of features. Due to the expensive cost of data collection process or other constraints, it is yet difficult to gather large samples in certain scientific domain of applications, e.g., bioinformatics, climate informatics, ecology and etc. The limited sample size in comparison to the data dimension has posed significant challenges for the analysis.

In principle, the challenges brought by high-dimensional data are two-fold. In terms of methodology, data scarcity usually leads to multiple, even infinitely many models that seemingly well fit the observed data but fail to capture the true underlying patterns. To address the issue, we need methods that can distinguish the true model from the spurious ones. On the other hand, theoretical study for high-dimensional data also needs new treatments. In the high-dimensional regime, large-sample based asymptotic analysis [170] is not suitable for characterizing the behavior of estimators under small sample. Therefore it is necessary to derive non-asymptotic results, which provide finite-sample guarantees that hold with high probability. Aiming at the two main challenges, the research on high-dimensional statistics has made substantial progress over the last two decades. Simply put, the key philosophy behind the study of high-dimensional data is the exploitation of prior knowledge on the model structure. Generally speaking, the source of such knowledge can be domain-specific expertise, experimental evidence or certain subjective beliefs. By enforcing the consistency between the model and the prior knowledge, we can effectively eliminate the incorrect models without using lots of data, which explains, at high level, why we can survive the high dimension. Though many previous works have demonstrated, both empirically and theoretically, that certain structural priors can significantly benefit the estimation of models, attention has rarely been devoted to understanding different apriori structures in a unified framework.

To some extent, a general framework can facilitate both algorithmic design and theoretical analysis of the estimator, as well as reveal the essence that plays a role in the estimation. Conversely, a unified understanding may inspire better ways to encode the prior knowledge. This thesis is motivated by this thread of thought.

1.1 High-Dimensional Statistics

1.1.1 Statistical Estimation and Curse of High Dimensions

Suppose that a parametric model $\mathcal{P} = \{f_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p\}$ is proposed for a sample space $\mathcal{Z}$, from which an independent and identically distributed (i.i.d.) data sample $\mathcal{Z}_n = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\}$ is generated with a specific parameter $\boldsymbol{\theta}^*$. The size of a data point usually reflects the *ambient dimension* p of the parameter space Θ. Given the data $\mathcal{Z}_n$, one of the central goals of statistical learning is to find an accurate approximation of $\boldsymbol{\theta}^*$. An *estimator* $\hat{\boldsymbol{\theta}}(\mathcal{Z}_n)$ is defined as a function that maps the (random) sample $\mathcal{Z}_n$ to an estimate in the parameter space, which is abbreviated as $\hat{\boldsymbol{\theta}}_n$ or $\hat{\boldsymbol{\theta}}$ when the context is clear. One common way to design estimators is through the *empirical risk minimization* (ERM) [171] framework. In order to characterize the fitness between a single observation $\mathbf{z}_i$ and a parameter $\boldsymbol{\theta}$, a *loss function* $\ell : \mathcal{Z} \times \Theta \mapsto \mathbb{R}$ is associated with the model $\mathcal{P}$, and the ERM estimator tries to minimize the average of ℓ over $\mathcal{Z}_n$, i.e.,

$$\hat{\boldsymbol{\theta}}_{\text{ERM}} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell\left(\mathbf{z}_i, \boldsymbol{\theta}\right) \ . \tag{1.1}$$

Particularly the *maximum likelihood* principle is often used to specify the ERM estimator, where the loss function ℓ is the negative log-likelihood of the model, i.e., $\ell(\mathbf{z}, \boldsymbol{\theta}) = -\log f_{\boldsymbol{\theta}}(\mathbf{z})$. In general, the estimators designed in classical statistical learning are focused on the *low-dimensional* setting in which $n \gg p$, and the parameter space Θ is usually unrestricted and equal to $\mathbb{R}^p$. The setup of the corresponding theoretical studies typically assumes that $n \to +\infty$ while p is fixed. To be specific, let us consider the following simple *linear model*,

$$y = \langle \mathbf{x}, \boldsymbol{\theta}^* \rangle + \epsilon \ , \tag{1.2}$$

where $\mathbf{x} \in \mathbb{R}^p$ and $y \in \mathbb{R}$ are *predictor vector* and *response* respectively, and the stochastic noise $\epsilon \sim \mathcal{N}(0, 1)$ is standard Gaussian. Given observed data $\mathcal{Z}_n = \{\mathbf{z}_i = (\mathbf{x}_i, y_i)\}_{i=1}^n$ with $n > p$, the maximum likelihood principle gives rise to the *ordinary least squares* (OLS) estimator, which estimates $\boldsymbol{\theta}^*$ by solving

$$\hat{\boldsymbol{\theta}}_{\text{OLS}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\operatorname{argmin}} \ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle \mathbf{x}_i, \boldsymbol{\theta} \rangle)^2 = \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 \ , \tag{1.3}$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^T$ is called *design matrix*, and $\mathbf{y} = [y_1, y_2, \ldots, y_n]^T$ is called *response vector*. The unique solution to (1.3) can be compactly written as

$$\hat{\boldsymbol{\theta}}_{\text{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \ , \tag{1.4}$$

as long as $\mathbf{X}^T\mathbf{X}$ is invertible, and numerical methods can efficiently compute this solution in polynomial time [97]. Regarding the theoretical analysis, based on central limit theorem (CLT) and delta method [39], one has *asymptotic normality* for $\hat{\boldsymbol{\theta}}_{\text{OLS}}$ as $n \to +\infty$,

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{\text{OLS}} - \boldsymbol{\theta}^*\right) \ \xrightarrow{d} \ \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}^{-1}\right) \ , \tag{1.5}$$

in which $\boldsymbol{\Sigma} = \mathbb{E}\left[\mathbf{x}\mathbf{x}^T\right]$ is the *covariance matrix* for $\mathbf{x}$. That is to say, for sufficiently large sample, the distribution of $\hat{\boldsymbol{\theta}}_{\text{OLS}}$ is close to $\mathcal{N}\left(\boldsymbol{\theta}^*, \frac{\boldsymbol{\Sigma}^{-1}}{n}\right)$, which can be further applied to inferential tasks, such as constructing *hypothesis test* and *confidence set*. Therefore, the study of linear models is rather complete in the low dimension for both computational and statistical aspects. The same estimation problem, however, exhibits rather different characteristics in the high-dimensional setting. First, the OLS solution is not unique when $n < p$, as the columns of $\mathbf{X}$ are linearly dependent. In fact, there can be infinitely many $\boldsymbol{\theta}$ that fit the data perfectly (i.e., satisfy $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}$), from which by no means can $\boldsymbol{\theta}^*$ be distinguished. Second, the asymptotic normality may break down even if a $\hat{\boldsymbol{\theta}}$ can be specified, and the limiting case poorly captures the finite-sample behavior of $\hat{\boldsymbol{\theta}}$. In short, switching linear models to the high-dimensional regime renders the results for the low dimension meaningless. What is worse, such situation is prevalent in statistical learning.

1.1.2 Surviving High Dimension: Sparsity and Convexity

The striking differences between the high-dimensional estimation and that in low dimension inspire the development of high-dimensional statistics, which concerns the estimation of statistical models under small sample. Since its inception [165], the core idea behind high-dimensional estimation has been centered around imposing prior structure on the true parameter $\boldsymbol{\theta}^*$, which can be fulfilled by restricting the parameter space Θ to be a strict subset of $\mathbb{R}^p$. The restricted parameter space often represents a *parsimonious* structure, which reflects the natural appeal to simplicity as suggested by the old principle, *Occam's razor* [163]. Parsimony is not only a subjective preference in consideration of interpretability, but also supported by empirical evidence in real-world applications. One of the most well-known parsimonious structures in high-dimensional statistics is *sparsity* [165], which posits that $\boldsymbol{\theta}^*$ has only *few non-zero* elements. For instance, natural images admit sparse representations in the wavelet basis, and a text document is usually related to only a few topics out of thousands of categories. At first glance, confining the parameter space using prior knowledge seems trivial, but the subsequent estimation is in fact more challenging than it appears. Returning to the linear model, if sparsity is assumed and $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^p \mid \|\boldsymbol{\theta}\|_0 = |\operatorname{supp}(\boldsymbol{\theta})| \leq s \ll p\}$ is an s-sparse parameter space, a straightforward estimator can be obtained by extending (1.3),

$$\hat{\boldsymbol{\theta}}_0 = \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^p} \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\theta}\|_0 \leq s \ . \tag{1.6}$$

However, the combinatorial nature of (1.6) makes the optimization *NP-hard* in general, which prevents us from pursuing this direction. To bypass the computational intractability of (1.6), numbers of alternatives have been proposed to incorporate the sparsity. A big family of approaches are based on *convexification*, which basically replaces $\|\cdot\|_0$ by its *convex surrogate*, L_1 norm $\|\cdot\|_1$, leading to a convex program,

$$\hat{\boldsymbol{\theta}}_{\text{cs}} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^p} \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\theta}\|_1 \leq \lambda \ , \tag{1.7}$$

where λ is a tuning parameter. In fact, the more widely adopted formulation is the regularized estimator, known as *Lasso* [165],

$$\hat{\boldsymbol{\theta}}_{\text{rg}} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^p} \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_1 \ , \tag{1.8}$$

which is also a convex optimization problem. In the literature, earlier analyses have shown that under mild assumptions on the distribution of $\mathbf{x}$ and suitable choice of λ, the L_2-error of $\hat{\boldsymbol{\theta}}_{\text{rg}}$ satisfies

$$\left\| \hat{\boldsymbol{\theta}}_{\text{rg}} - \boldsymbol{\theta}^* \right\|_2 \leq O\left(\sqrt{\frac{s \log p}{n}} \right) \ , \tag{1.9}$$

with high probability if the true $\boldsymbol{\theta}^*$ is s-sparse. A similar result holds for the constrained estimator $\hat{\boldsymbol{\theta}}_{\text{cs}}$ as well. Unlike the asymptotic result, the finite-sample bound gives an exact dependency of error on n, p and s. More importantly, the sample size only needs to satisfy $n = \omega(s \log p)$ in order to guarantee the estimation consistency, while the low dimension requires $n = \omega(p)$. The sharp contrast between the requirements on sample size conveys a key message that additional structures of $\boldsymbol{\theta}^*$ can greatly benefit the estimation.

The topic of sparsity has also been extensively investigated in the field of *compressed sensing* (CS). The goal of compressed sensing is to estimate a sparse vector (i.e., signal) from a small number of linear measurements, which is similar to the estimation of sparse linear models. The most significant difference between the two settings is that the design matrix $\mathbf{X}$ in CS is often well controlled by the experimenter. The ability to manipulate the design can guarantee many nice properties, based on which several algorithms are proposed for CS, including orthogonal matching pursuit (OMP) [167] and compressive sampling matching pursuit (CoSaMP) [126], just to name a few. Though being fast in practice, these algorithms are less extensible to other settings beyond sparsity and linear measurements. Moreover, the data gleaned in statistical learning are less controllable, and the methods above can be vulnerable to the violation of the desired properties.

1.2 Beyond Unstructured Sparsity

The sparsity structure introduced in Section 1.1.2 is sometimes termed as *unstructured sparsity*, since no additional pattern of sparsity is known. Recent years have witnessed a surge of development in other types of sparsity, which are considered as *structured sparsity* [11,12,76] (see Figure 1.1). A popular example is the *group sparsity* [183], where

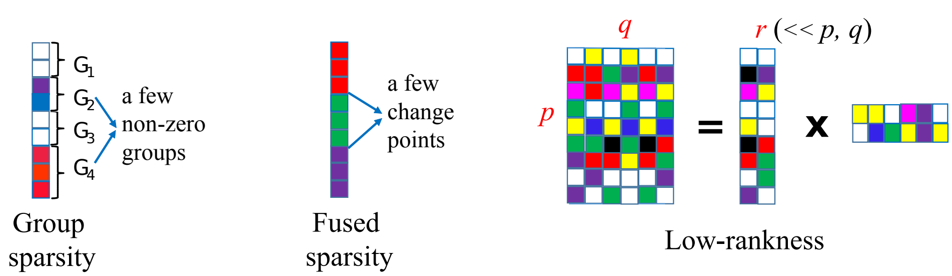


Figure 1.1: Examples of structures beyond unstructured sparsity

sparsity is imposed on predefined groups of entries of $\boldsymbol{\theta}^*$ rather than the individuals. The groups themselves can be structured as well, e.g., non-overlapping, overlapping with hierarchy, and etc. Group sparsity has found numbers of specific applications in real-world problems, such as expression quantitative trait loci (eQTL) mapping in genetics [95], and sparse coding in signal processing [91]. Another widely-used structured sparsity is the *fused sparsity* [166], where only a small fraction of neighboring pairs in $\boldsymbol{\theta}^*$ have different values from each other. That is to say, $\boldsymbol{\theta}^*$ is piecewise constant with only few change points. Apart from the adjacency induced by the inherent one-dimensional chain structure, the elements of $\boldsymbol{\theta}^*$ can be organized as nodes of a graph, and the fused sparsity can be defined over the edges of the graph. The applications of fused sparsity include time-varying network recovery [2], DNA copy number variation (CNV) detection [164] and so on. The notion of sparsity can also be suitably generalized to matrix setting, resulting in the *low-rank* structure, which has been extensively exploited in the context of recommender systems [96], natural language processing [50], image analysis [33]. The low-rank structure simply assumes that the true matrix to be estimated has relatively small rank, i.e., has only few non-zero singular values. Furthermore, more complex structures can be created from simpler ones. For instance, one may assume that the true parameter simultaneously has multiple different structures [143], or it is a superposition of two or more structured components [66, 87].

Given massive interesting structures, the key to extending the aforementioned idea of convexification is to find the corresponding convex surrogate functions (usually norms). For the group sparsity and fused sparsity, their convex surrogates are simply given by the $L_{2,1}$ group norm [183] and the total variation (TV) function [166] respectively, while the low-rank structure is usually captured by the nuclear norm [141]. In the literature, there are also systematic ways to define convex surrogates, for example, via *submodular function* [13] and *infimal convolution* [40]. Broadly speaking, the surrogate function encodes the constrained parameter space in which $\boldsymbol{\theta}$ has limited degree of freedom, such that the preferred structure has a small function value. Computationally, using either constrained or regularized estimator with a convex loss ℓ, we end up with a convex program that can be solved globally in polynomial time [10, 24, 26]. Statistically, however, the state-of-the-art understanding falls short for general structures. Earlier works [21, 174, 188] were simply focused on the unstructured sparsity, which were later extended to group sparsity [75], fused sparsity [113], and etc. Those case-by-case analyses lack a general view into the key factors that determine the performance of the convex surrogates. On the contrary, a unified framework for general structures can avoid complicacies and help the analysis when we cope with new structures.

In this thesis, our first goal is committed to have a deeper understanding towards such unification. First, we concentrate on the *Dantzig*-type estimator for linear models, which is less studied in the literature. In particular, we extend the original Dantzig selector [32] to the generalized Dantzig selector (GDS), in order to accommodate general structures. Unlike the loss-minimization formulation in (1.7) and (1.8), the objective of Dantzig-type estimator is the convex surrogate instead of the loss, which is often non-smooth and needs extra care. Therefore, we come up with an efficient alternating direction method of multipliers (ADMM) to solve the associated optimization problem. On the statistical side, we introduce the critical *geometric measures* – *Gaussian width* [63] and *restricted norm compatibility* – which describe the recovery guarantees of GDS. Following that, we turn to bounding the geometric measures by utilizing simple information of the structures, which largely simplifies the calculation. Moreover, we have extended those results to the matrix setting. Second, we focus on a semi-parametric extension of linear models, the single-index model (SIM), where the response is assumed to be an *unknown* transformation of the original linear measurement. To estimate the underlying parameter, we propose two types of simple estimators, the constrained and the regularized one, based on U-statistics [98]. Under suitable conditions, the L_2-error bound of both estimators can be established using similar geometric measures. In addition to SIMs, we also propose a new semi-parametric model called sparse linear isotonic model (SLIM) for the high-dimensional setting, which allows nonlinear monotone transformations of the features. For SLIM, we design the computational algorithm to estimate the unknown parameter, which also leverages U-statistics. At the same time, some statistical guarantees are derived to complement the computational development of SLIM.

1.3 Beyond Convexity

As discussed in previous sections, the convexification plays a crucial role in high-dimensional estimation, which addresses the computational challenge brought by the combinatorial structure of $\boldsymbol{\theta}^*$. If the loss ℓ is convex, the optimization problems associated with both the constrained and the regularized estimator can be solved globally, which avoids the local optima that could be statistically erroneous. However, pursuing convexity is not always a free lunch. For certain estimation problems, such as dictionary learning [1] and phase retrieval [34], the natural formulation of the loss is inherently non-convex, and exploring hidden convexity (if there is any) may require skillful reformulations [8,37]. Furthermore the structure of the estimator $\hat{\boldsymbol{\theta}}$ obtained by using convex surrogate may slightly differ from the desired one. In some tasks, e.g., variable selection, extra effort is needed to convert $\hat{\boldsymbol{\theta}}$ into the sought structure. Though convexity guarantees global optimality, solving convex estimators sometimes can be computationally expensive compared with local search heuristics applied to non-convex formulations, e.g., in low-rank matrix estimation [83, 84].

Given the above shortcomings of convex formulations, it is sometimes tempting to

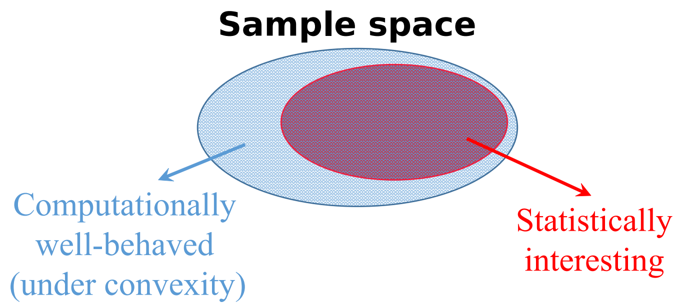


Figure 1.2: Though convexity guarantees computational optima for *all* data (blue) from the sample space, only a *subset* of them (red) are of statistical interest. The rest of data are fundamentally uninformative in the information-theoretic sense.

try non-convex estimators in high dimension, which could involve either non-convex losses or unconvexified functions that capture the structure of $\boldsymbol{\theta}^*$. As far as computation is concerned, non-convexity is notorious for the risk of getting trapped in local optima as well as the computational hardness, especially when discrete structures present (see (1.6) in Section 1.1.2). Despite those disadvantages, the statistical performance of non-convex estimators is often superb in practice. Such gap between the computational and the statistical aspects is rooted in the assumption on data. Without the access to unrestricted computational resources, convexity is essential for ensuring the computational global optima for *arbitrary* input data. On the contrary, statistical recovery is typically focused on *generic* data, since the worst-case scenario can be too pessimistic to encounter in practice. Moreover the computational results for untypical data could fail to make any statistical sense even though they are globally optima guaranteed by convexity. To see a concrete example, we revisit the linear model (1.2). Suppose that the noise ϵ is zero and the received data are of the form $(\mathbf{x}_i, y_i) = (\mathbf{0}, 0)$. In this scenario, both L_1-regularized and L_1-constrained estimator always yield the estimate $\hat{\boldsymbol{\theta}} = \mathbf{0}$, regardless of the true s-sparse $\boldsymbol{\theta}^*$. Although $\hat{\boldsymbol{\theta}} = \mathbf{0}$ is the computational optimum, its statistical error can be arbitrarily large due to the pathological data. Thus convexity, to some extent, is an unnecessarily strong notion in the statistical context, which is illustrated by Figure 1.2. With that being said, it is of little interest to study the computation alone without investigating the recovery guarantee, when it comes to statistical estimation. On the other side, the focus of statistical recovery may give us an opportunity to relax the convexity requirement and design non-convex methods tailored specifically for generic data. Guided by this thinking, the study of non-convex optimization/estimation has received considerable attention over the last few years. Several influential papers [20, 34, 60, 86, 159] have managed to show that some non-convex estimators can be empowered when generic data are considered. More precisely, under suitable stochastic assumptions on data, these estimators are able to recover the underlying true parameter with provably small error, which include the formulation (1.6) for sparse liner regression that is nevertheless computationally infeasible in the worst case. However, like the convex setting, so far most of the related works on non-convex estimation have not yet explored the general structure of parameter, with only few exceptions [130, 154].

Motivated by both the success of non-convex optimization and the inadequate attention on general structures, the second goal of this thesis is to investigate the unification of structured estimation under non-convexity, which parallels the goal for convex setting. In particular, we consider the problem of estimating multi-response linear models with general structures. Apart from the parameter vector $\boldsymbol{\theta}^*$ in vanilla linear models, here we also need to deal with the unknown noise covariance across the responses, which makes the estimation problem non-convex. We first propose an alternating estimation (AltEst) framework, a generalization of the popular alternating minimization (AltMin) procedure for non-convex optimization [82], and plug GDS in this framework to estimate both parameter vector and noise covariance. In the meanwhile, we derive the statistical guarantee for an idealized version of AltEst applied to multi-response linear models, which utilizes the same geometric measures as mentioned earlier. Second we aim at relaxing the requirement of a norm surrogate when using GDS, along with an improved statistical analysis without assuming any idealized conditions. Specifically the GDS in the proposed AltEst framework is replaced by a constrained estimator which, from a computational perspective, is more amenable to non-norm (or non-convex) characterization of the structure of $\boldsymbol{\theta}*$. For the statistical analysis, by using a modified proof strategy, we are able to concentrate on the practical version of AltEst instead of the idealized one, whose theoretical guarantee is confirmed by the empirical observations.

1.4 Contributions and Organization

The main theme of this thesis is to develop both computational and statistical framework for some high-dimensional estimation problems, with an emphasis on general structures. For the computational aspect, we embrace both the idea of convexification and the non-convexity as it is, and provide algorithmic recipes for different types of estimators. On the statistical side, we focus on the L_2-error analysis and establish the error bound in terms of certain geometric measures. Moreover, we demonstrate the usefulness of these geometric measures, by deriving their further bounds for a broad class of structures. Hence our theoretical results do not leave in the bound any quantities that is hard to calculate.

The organization of this thesis is as follows.

- In Chapter 2, we provide a review for some background knowledge in probability theory, convex analysis and optimization. Also, we introduce an important notion called *Gaussian width* [63] along with *generic chaining* [161], an advanced tool in probability theory, which plays a key role in establishing the statistical guarantees.

- In Chapter 3, we extend the celebrated *Dantzig selector* for sparse linear models to accommodate general structures. As to optimization, the proposed *generalized Dantzig selector* (GDS) [41] can be efficiently solved by a variant of basic alternating direction method of multipliers (ADMM). In terms of statistical analysis, we present a unified framework for various structures, which can succinctly characterize the error bound with certain geometric measures, such as Gaussian width.

- Chapter 4 is devoted to the study of the geometric measures introduced in Chapter 3. Those geometric measures essentially quantify the complexity of the associated structures, which need to be computed or bounded in order to determine the final error bound. For a broad class of structures that can be captured by *atomic norms*, we have managed to bound the geometric measures using simple information of the structure [43].

- In Chapter 5, we extend the results obtained in Chapter 3 and 4 to the matrix scenario [44], in which we have general bounds for the structures induced by the family of *unitarily invariant norm.*

- In Chapter 6, we study an important semi-parametric extension of linear models, the *single-index models* (SIMs), which allow the response to be an *unknown* transfer of the linear measurement. We develop two types of estimators for the recovery of model parameters [46]. With minimal assumption on noise, the statistical guarantees are established for the proposed estimators under suitable conditions, which also allow general structures of the underlying parameter. Moreover, the proposed estimator is novelly instantiated for SIMs with monotone transfer function, and the obtained estimator can better leverage the monotonicity.

- In Chapter 7, we make an attempt to introduce some nonlinearity in the features of linear models, as opposed to the nonlinear response considered by single-index models. In particular, we propose a novel model named sparse linear isotonic model (SLIM) [47], which hybridizes the ideas in both parametric sparse linear models and additive isotonic models (AIMs) that assume the response to be a summation of unknown monotone feature transformations. In the computational aspect, a two-step algorithm is designed for estimating the sparse parameter as well as the monotone functions. Under mild statistical assumptions, we show that the algorithm can accurately estimate the parameter.

- In Chapter 8, we focus on the non-convex estimation of structured multi-response linear models. By exploiting the noise correlations among different responses, we employ an alternating estimation (AltEst) procedure [45] to estimate the parameters based on GDS. Under suitable sample size requirement and the resampling assumption, we show that the error of the estimates generated by an variant of AltEst, with high probability, converges linearly to certain minimum achievable level, which can be tersely expressed by the geometric measures.

- In Chapter 9, we continue to investigate the structured multi-response linear models, with several extensions from Chapter 8. We allow the function encoding the structure of the parameter to be non-convex, through replacing the GDS in the AltEst framework by a constrained estimator, which results in an alternating-minimization-type algorithm. In the statistical analysis, we relax the assumption on the noise distribution. More importantly, we come up with a new analysis for the practical version of the estimator, which does not resort to any resampling assumptions. The result also reveals that random initializations of the estimation algorithm can even yield good recovery of the unknown parameter.

- Chapter 10 is dedicated to the conclusion, in which we summarize the contributions of this thesis.

Chapter 2

Preliminaries

2.1 Convex Analysis

In this section, we briefly review some basics of convex analysis. Since the scope of this topic is too wide, we will just cover those used in our works for the sake of simplicity as well as keeping the self-containedness. For more complete materials, we refer interested readers to [144].

2.1.1 Convex Set

We start with the definition of *convex set* in $\mathbb{R}^p$.

Definition 1 (convex set) A set $\mathcal{C} \subseteq \mathbb{R}^p$ is convex if the following holds for any $\mathbf{u}, \mathbf{v} \in \mathcal{C}$,

$$\lambda \mathbf{u} + (1-\lambda)\mathbf{v} \in \mathcal{C}, \quad \forall \;\; 0 \leq \lambda \leq 1 \; . \tag{2.1}$$

Examples of convex set include *affine set* $\{\mathbf{u} \mid \mathbf{A}\mathbf{u} = \mathbf{b}\}$ ($\mathbf{A} \in \mathbb{R}^{q \times p}$, $\mathbf{b} \in \mathbb{R}^q$ and q are fixed), *half-space* $\{\mathbf{u} \mid \langle \mathbf{w}, \mathbf{u} \rangle \geq \beta\}$ ($\mathbf{w} \in \mathbb{R}^p$ and $\beta \in \mathbb{R}$ are fixed), and so on. Another important instance of convex set is *convex cone.*

Definition 2 (cone/convex cone) A set $\mathcal{C} \subseteq \mathbb{R}^p$ is a *cone* if it satisfies that

$$\mathbf{u} \in \mathcal{C} \qquad \Longrightarrow \qquad \lambda \mathbf{u} \in \mathcal{C}, \ \forall \ \lambda > 0 \ . \tag{2.2}$$

If $\mathcal{C}$ is further convex, then it is a *convex cone*.

Given a set $\mathcal{A} \subseteq \mathbb{R}^p$, we can construct a cone by the operator $\text{cone}(\mathcal{A}) = \{c \cdot \mathbf{a} \mid c \geq 0, \ \mathbf{a} \in \mathcal{A}\}$. For an arbitrary set, we can also define a special convex set called *convex hull*, which is its smallest convex superset.

Definition 3 (convex hull) Given any set $\mathcal{S} \in \mathbb{R}^p$, its *convex hull*, denoted by $\text{conv}(\mathcal{S})$, is the smallest convex set containing $\mathcal{S}$. In particular, if $\mathcal{S} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ is finite, then $\text{conv}(\mathcal{S})$ consists of all *convex combinations* of $\mathbf{u}_1, \ldots, \mathbf{u}_n$, i.e.,

$$\text{cone}(\mathcal{S}) = \left\{ \lambda_1 \mathbf{u}_1 + \lambda_2 \mathbf{u}_2 + \ldots + \lambda_n \mathbf{u}_n \ \middle| \ \sum_{i=1}^{n} \lambda_i = 1, \ \lambda_1, \lambda_2, \ldots, \lambda_n \geq 0 \right\} \tag{2.3}$$

2.1.2 Convex Function

Based on the definition of convex set, the *convex function* can be defined as follows.

Definition 4 (convex function) A function $f : \mathbb{R}^p \mapsto \mathbb{R}$ is said to be convex if its domain $\text{dom}\, f$ is convex and f satisfies that for any $\mathbf{u}, \mathbf{v} \in \text{dom}\, f$

$$f\left(\lambda \mathbf{u} + (1-\lambda)\mathbf{v}\right) \leq \lambda f(\mathbf{u}) + (1-\lambda) f(\mathbf{v}), \quad \forall \ \ 0 \leq \lambda \leq 1 \ . \tag{2.4}$$

Specifically a convex function f is said to be *proper* if $f(\mathbf{u}) < +\infty$ for at least one $\mathbf{u}$ and $f(\mathbf{u}) > -\infty$ for all $\mathbf{u}$.

There are several useful notions related to convex functions, such as *convex conjugate* and *gauge function* (a.k.a. *Minkowski functional*).

Definition 5 (convex conjugate) For any function $f : \mathbb{R}^p \mapsto \mathbb{R}$, its *convex conjugate* $f^* : \mathbb{R}^p \mapsto \mathbb{R}$ is given by

$$f^*(\mathbf{u}) = \sup_{\mathbf{v} \in \operatorname{dom} f} \{\langle \mathbf{u}, \mathbf{v} \rangle - f(\mathbf{v})\} \tag{2.5}$$

Note that f^* is always convex even if f is not. f^* is also known as *Fenchel conjugate*. A special type of convex conjugate is *support function*, where f is the *indicator function* of a non-empty set $\mathcal{S}$, i.e.,

$$\mathbb{I}_{\mathcal{S}}(\mathbf{u}) = \begin{cases} 0, & \text{if } \mathbf{u} \in \mathcal{S} \\ +\infty, & \text{otherwise} \end{cases} . \tag{2.6}$$

Definition 6 (support function) The *support function* of a non-empty set $\mathcal{S}$ is given by

$$h_{\mathcal{S}}(\mathbf{u}) = \sup_{\mathbf{v} \in \mathbb{R}^p} \{\langle \mathbf{u}, \mathbf{v} \rangle - \mathbb{I}_{\mathcal{S}}(\mathbf{v})\} = \sup_{\mathbf{v} \in \mathcal{S}} \langle \mathbf{u}, \mathbf{v} \rangle \tag{2.7}$$

In some places, convex conjugate and support function are only considered for convex f and $\mathcal{S}$. In this thesis, it is also sufficient to just focus on convex case.

Definition 7 (gauge function) The *gauge function* (or simply *gauge*) of a non-empty convex set $\mathcal{C}$ is defined as

$$\gamma_{\mathcal{C}}(\mathbf{u}) = \inf \left\{ \lambda \geq 0 \;\middle|\; \mathbf{u} \in \lambda \mathcal{C} \right\} \tag{2.8}$$

The gauge function is convex as well, and a useful class of gauge is *norm*, for which the convex set $\mathcal{C}$ should be bounded, centrally symmetric about the origin (i.e., $\mathbf{u} \in \mathcal{C}$ iff. $-\mathbf{u} \in \mathcal{C}$), and include $\mathbf{0}$ in its interior.

Definition 8 (norm) A *norm* $\|\cdot\|$ is a function mapping from $\mathbb{R}^p$ to $\mathbb{R}$, which satisfies

- (positivity) $\|\mathbf{u}\| \geq 0 \;\; \forall \; \mathbf{u} \in \mathbb{R}^p$, and $\|\mathbf{u}\| = 0$ iff. $\mathbf{u} = \mathbf{0}$
- (absolute homogeneity) $\|\lambda\mathbf{u}\| = |\lambda| \cdot \|\mathbf{u}\| \;\; \forall \; \mathbf{u} \in \mathbb{R}^p, \; \lambda \in \mathbb{R}$
- (subadditivity) $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\| \;\; \forall \; \mathbf{u}, \mathbf{v} \in \mathbb{R}^p$

A *dual norm* $\|\cdot\|_*$ can be defined for the original norm $\|\cdot\|$ through support function,

$$\|\mathbf{u}\|_* = \sup_{\|\mathbf{v}\| \leq 1} \langle \mathbf{u}, \mathbf{v} \rangle \tag{2.9}$$

Simple examples of norm are L_2 norm $\|\mathbf{u}\|_2 = (\sum_{i=1}^p u_i^2)^{1/2}$, L_1 norm $\|\mathbf{u}\|_1 = \sum_{i=1}^p |u_i|$, L_∞ norm $\|\mathbf{u}\|_\infty = \max_{1 \leq i \leq p} |u_i|$, and etc. The dual norm of L_2 norm is itself, and L_1 and L_∞ norm are dual to each other. Norm plays a central role in high-dimensional statistics, which often acts as the convex surrogate for certain structure. One nice property of dual norm is the *Hölder's inequality*.

Proposition 1 (Hölder's inequality) *For any norm $\|\cdot\|$ and its dual norm $\|\cdot\|_*$, it holds that $|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\| \cdot \|\mathbf{v}\|_*$ for any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$.*

Encompassing the norm as a special case, gauge function provides a different perspective of view into Definition 8. The closure of the convex set $\mathcal{C}$ that induces the norm $\|\cdot\|$ is actually the (closed) *unit norm ball*

$$\Omega = \left\{ \mathbf{u} \in \mathbb{R}^p \;\middle|\; \|\mathbf{u}\| \leq 1 \right\} \; . \tag{2.10}$$

Thus one can define the norm by specifying its unit ball, instead of giving the arithmetic expression. Such correspondence is helpful when we introduce the *atomic norm* [40] below.

Definition 9 (atomic norm) Given a compact set $\mathcal{A}$ that is centrally symmetric about origin and satisfies $\text{span}(\mathcal{A}) = \mathbb{R}^p$, define the *atomic norm* $\|\cdot\|_{\mathcal{A}}$ of $\mathcal{A}$ by

$$\|\mathbf{u}\|_{\mathcal{A}} = \inf\left\{\sum_{\mathbf{a}\in\mathcal{A}} c_{\mathbf{a}} \;\middle|\; \mathbf{u} = \sum_{\mathbf{a}\in\mathcal{A}} c_{\mathbf{a}}\mathbf{a},\ c_{\mathbf{a}} \geq 0 \ \ \forall\, \mathbf{a} \in \mathcal{A}\right\} . \tag{2.11}$$

The set $\mathcal{A}$ is called *atomic set*, and its element $\mathbf{a} \in \mathcal{A}$ is called *atom.*

Though the expression of atomic norm seems complicated, the unit norm ball of $\|\cdot\|_{\mathcal{A}}$ turns out to be simple.

Proposition 2 (unit ball of atomic norm) *The unit ball of atomic norm $\|\cdot\|_{\mathcal{A}}$ is the convex hull of $\mathcal{A}$, i.e., $\Omega_{\mathcal{A}} = \text{conv}(\mathcal{A})$. It follow immediately from this fact that the dual norm of $\|\cdot\|_{\mathcal{A}}$ is*

$$\|\mathbf{u}\|_{\mathcal{A}}^{*} = \sup_{\mathbf{v}\in\text{conv}(\mathcal{A})} \langle \mathbf{u}, \mathbf{v}\rangle = \sup_{\mathbf{v}\in\mathcal{A}} \langle \mathbf{u}, \mathbf{v}\rangle \tag{2.12}$$

Now the definition of atomic norm may look tricky given that $\Omega_{\mathcal{A}} = \text{conv}(\mathcal{A})$, since any norm $\|\cdot\|$ can be made atomic norm by choosing the atomic set $\mathcal{A}$ to its unit ball Ω. In practice, typically we bring up this notion only when $\mathcal{A}$ is finite. L_1 and L_∞ norm are representative atomic norms, whose atomic sets are $\mathcal{A}_{L_1} = \{\pm\mathbf{e}_1, \pm\mathbf{e}_2, \ldots, \pm\mathbf{e}_p\}$ ($\{\mathbf{e}_i\}$ denotes the standard basis of $\mathbb{R}^p$) and $\mathcal{A}_{L_\infty} = \{\pm 1\}^p$, respectively.

2.2 Convex Optimization

Convex optimization is paramount in modern machine learning and statistics, as finding the optimal parameters for statistical models can be often formulated as convex optimization problem. We are not intended to give a comprehensive review for every popular algorithms in the literature. Instead we will cover the basic gradient descent, proximal gradient method, and alternating direction method of multipliers. Generally speaking, convex optimization problem (or convex program) can be cast as

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} g(\boldsymbol{\theta}) \quad \text{s.t.} \quad \boldsymbol{\theta} \in \mathcal{C} \ , \tag{2.13}$$

where both *feasible set* $\mathcal{C} \subseteq \mathbb{R}^p$ and *objective function* $g : \mathbb{R}^p \mapsto \mathbb{R}$ are convex. In particular, when $\mathcal{C} = \mathbb{R}^p$, we say that the optimization problem is *unconstrained*. Convex optimization algorithms usually employ an iterative procedure to generate a sequence of *iterates*, $\boldsymbol{\theta}_{(0)}, \boldsymbol{\theta}_{(1)}, \ldots, \boldsymbol{\theta}_{(T)} \in \mathcal{C}$, such that $\lim_{T\to\infty} f(\boldsymbol{\theta}_{(T)}) = f(\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta}\in\mathcal{C}} g(\boldsymbol{\theta})$.

2.2.1 Gradient Descent

In many scenarios, we deal with unconstrained convex problems with g being smooth, which is arguably the simplest case of convex optimization. Unconstrained smooth optimization can be solved by *gradient descent* (GD), which iteratively performs

$$\boldsymbol{\theta}_{(t+1)} = \boldsymbol{\theta}_{(t)} - \eta \nabla g(\boldsymbol{\theta}_{(t)}) \tag{2.14}$$

in which η is the *step size*. In practice, η can vary along the iterations, e.g., can be determined by line search. The full algorithm is given in Algorithm 1. Under suitable conditions on g and step size, GD converges at rate of $O(1/T)$, namely

$$g(\boldsymbol{\theta}_{(T)}) - g(\hat{\boldsymbol{\theta}}) \leq O\left(\frac{1}{T}\right) \tag{2.15}$$

Algorithm 1 Gradient Descent (GD)

Input: step size η, number of iterations T
Output: iterate $\boldsymbol{\theta}^T$

Initialize $\boldsymbol{\theta}_{(0)}$
for $t = 0$ to $T-1$ **do**
  $\boldsymbol{\theta}_{(t+1)} = \boldsymbol{\theta}_{(t)} - \eta \nabla g(\boldsymbol{\theta}_{(t)})$
end for
return $\boldsymbol{\theta}_{(T)}$

2.2.2 Proximal Gradient Method and Proximal Operator

In high-dimensional statistics, as shown in (1.7) and (1.8), we often face more complex problems, with either nontrivial constraint or non-smooth objective. The two types of estimators can be unified in a single optimization framework. Consider the following problem

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \; f(\boldsymbol{\theta}) + h(\boldsymbol{\theta}) \; , \tag{2.16}$$

where f is smooth while h is non-smooth. For constrained problem, the constraint $\boldsymbol{\theta} \in \mathcal{C}$ can be incorporated into h by setting $h(\cdot) = \mathbb{I}_{\mathcal{C}}(\cdot)$. Taking $f(\boldsymbol{\theta}) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$ and $h(\cdot) = \mathbb{I}_{\lambda\Omega_{L1}}(\cdot)$ (or $h(\cdot) = \lambda\|\cdot\|_1$), we recover (1.7) (or (1.8)). The problem (2.16) can be solved by *proximal gradient method* (PGM). The algorithmic description is provided in Algorithm 2. Essentially PGM executes a gradient-descent step followed by a *proximal operator* (or *proximal mapping*) (Line 4). The proximal operator is formally defined

Algorithm 2 Proximal Gradient Method (PGM)

Input: step size η, number of iterations T
Output: iterate $\boldsymbol{\theta}_{(T)}$

1: Initialize $\boldsymbol{\theta}_{(0)}$
2: **for** $t = 0$ to $T-1$ **do**
3:   $\boldsymbol{\theta}' = \boldsymbol{\theta}_{(t)} - \eta \nabla f(\boldsymbol{\theta}_{(t)})$
4:   $\boldsymbol{\theta}_{(t+1)} = \operatorname{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^p} \; \frac{1}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2 + \eta \cdot h(\boldsymbol{\theta})$
5: **end for**
6: **return** $\boldsymbol{\theta}_{(T)}$

below.

Definition 10 (proximal operator) The *proximal operator* $\text{prox}_h : \mathbb{R}^p \mapsto \mathbb{R}^p$ for a closed proper convex function h is given as

$$\text{prox}_h(\mathbf{u}) = \operatorname*{argmin}_{\mathbf{v}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{u}-\mathbf{v}\|_2^2 + h(\mathbf{v}) \ . \tag{2.17}$$

If $h = \mathbb{I}_{\mathcal{C}}$ is the indicator function for a set $\mathcal{C}$, the proximal operator is also called *projection operator*,

$$\text{proj}_{\mathcal{C}}(\mathbf{u}) = \text{prox}_{\mathbb{I}_{\mathcal{C}}}(\mathbf{u}) = \operatorname*{argmin}_{\mathbf{v}\in\mathcal{C}} \frac{1}{2}\|\mathbf{u}-\mathbf{v}\|_2^2 \tag{2.18}$$

It can be shown that the proximal operator exists for all $\mathbf{u} \in \mathbb{R}^p$ and is also unique. The success of PGM heavily relies on the computation of the proximal operator being inexpensive, which is often the case for many useful h. For suitable f, the convergence rate of PGM is also $O(1/T)$.

2.2.3 Alternating Direction Method of Multipliers

In machine learning and statistics, sometimes we may come across more complicated optimization problems that involve two blocks of variables, subject to equality constraints, i.e,

$$\min_{\substack{\boldsymbol{\theta}\in\mathbb{R}^p \\ \boldsymbol{\beta}\in\mathbb{R}^q}} f(\boldsymbol{\theta}) + g(\boldsymbol{\beta}) \quad \text{s.t} \quad \mathbf{A}\boldsymbol{\theta} + \mathbf{B}\boldsymbol{\beta} = \mathbf{c} \ , \tag{2.19}$$

where f and g are both convex, but not necessarily smooth. $\mathbf{A} \in \mathbb{R}^{r\times p}$, $\mathbf{B} \in \mathbb{R}^{r\times q}$ and $\mathbf{c} \in \mathbb{R}^r$ are generic matrices and vector. Since the smoothness of f and g is not guaranteed, we cannot solve (2.19) using PGM. In recent years, an popular approach to tackle such problem is the *alternating direction method of multipliers* (ADMM). The basic idea of ADMM is to form the *augmented Lagrangian*,

$$L_\rho(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\mu}) = f(\boldsymbol{\theta}) + g(\boldsymbol{\beta}) + \langle \boldsymbol{\mu}, \mathbf{A}\boldsymbol{\theta} + \mathbf{B}\boldsymbol{\beta} - \mathbf{c} \rangle + \frac{\rho}{2} \left\| \mathbf{A}\boldsymbol{\theta} + \mathbf{B}\boldsymbol{\beta} - \mathbf{c} \right\|_2^2 \ , \tag{2.20}$$

where $\boldsymbol{\mu} \in \mathbb{R}^r$ is the dual variable, and $\rho > 0$ is a tuning parameter. Then ADMM solves the original problem by iteratively minimizing L_ρ w.r.t. two blocks of primal variables, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, followed by an update of dual variable $\boldsymbol{\mu}$. Algorithm 3 gives the details of ADMM. Under mild conditions on the problem (2.19), ADMM enjoys $O(1/T)$ rate of convergence.

Algorithm 3 Alternating Direction Method of Multipliers (ADMM)

Input: tuning parameter ρ, number of iterations T
Output: iterates $\boldsymbol{\theta}_{(T)}$ and $\boldsymbol{\beta}_{(T)}$

```
Initialize β_(0) and μ_(0)
for t = 0 to T − 1 do
    θ_(t+1) = argmin_θ L_ρ(θ, β_(t), μ_(t))
    β_(t+1) = argmin_β L_ρ(θ_(t+1), β, μ_(t))
    μ_(t+1) = μ_(t) + ρ(Aθ_(t+1) + Bβ_(t+1) − c)
end for
return θ_(T) and β_(T)
```

2.3 Basics of Probability Theory

In this section, we will review the basics of probability theory, including the notions of sub-Gaussian and sub-exponential random variable and related concentration inequalities.

2.3.1 Gaussian Random Variable

Gaussian random variable (r.v. for short) is arguably the most well-known random variable in probability theory, whose density function is

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} , \tag{2.21}$$

where μ and σ^2 are mean and variance, respectively. The *standard* Gaussian r.v. has zero-mean and unit-variance. In asymptotic setting, the limiting distributions for many statistics follow Gaussian distributions, and lots of nice properties can be shown for Gaussian random variable. We present below a few useful facts about Gaussian r.v. that are frequently utilized in this thesis work.

Proposition 3 *Suppose that x and y are two Gaussian random variables. Then x and y are independent if and only if they are uncorrelated, i.e.,*

$$\mathrm{Cov}(x,y) = 0 \qquad \Longleftrightarrow \qquad x \perp y$$

In general, the equivalence above does not hold for other random variables, though independence always implies uncorrelatedness. The Gaussianity can be carried to random vector too. A Gaussian random vector (or multivariate Gaussian) $\mathbf{x} \in \mathbb{R}^p$ has the density of the form

$$f\left(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}\right) = \frac{1}{\sqrt{(2\pi)^p|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2}\right\} , \tag{2.22}$$

where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ is the covariance matrix. Standard Gaussian random vector is referred to the one with $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{I}$. The Gaussianity of random vector is preserved under linear transformation.

Proposition 4 *If* $\mathbf{x} \in \mathbb{R}^p$ *is a Gaussian random vector with mean* $\boldsymbol{\mu}$ *and covariance* $\boldsymbol{\Sigma}$*, i.e.,* $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*, then* $\mathbf{A}\mathbf{x} \in \mathbb{R}^q$ *is also Gaussian for any fixed* $\mathbf{A} \in \mathbb{R}^{q \times p}$*, with*

$$\mathbb{E}\left[\mathbf{A}\mathbf{x}\right] = \mathbf{A}\boldsymbol{\mu} \qquad \textit{and} \qquad \mathrm{Cov}\left[\mathbf{A}\mathbf{x}\right] = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T \ . \tag{2.23}$$

In particular, $\langle \mathbf{a}, \mathbf{x} \rangle \sim \mathcal{N}(\mathbf{a}^T\boldsymbol{\mu}, \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a})$ *for any* $\mathbf{a} \in \mathbb{R}^p$*.*

Lipschitz function of Gaussian random vector enjoys a dimensionality-independent type of concentration via isoperimetric inequalities.

Proposition 5 *Let* $\mathbf{x} \in \mathbb{R}^p$ *be a standard Gaussian random vector, and* $f : \mathbb{R}^p \mapsto \mathbb{R}$ *be an L-Lipschitz function. For any* $\epsilon \geq 0$*, we have*

$$\mathbb{P}\left(f(\mathbf{x}) - \mathbb{E}f(\mathbf{x}) > \epsilon\right) \leq \exp\left(-\frac{\epsilon^2}{2L^2}\right) \tag{2.24}$$

2.3.2 Sub-Gaussian and Sub-Exponential Random Variable

A random variable x is *sub-Gaussian* if the ψ_2*-norm* defined below is finite

$$|||x|||_{\psi_2} \triangleq \sup_{q \geq 1} \frac{(\mathbb{E}|x|^q)^{\frac{1}{q}}}{\sqrt{q}} < +\infty \tag{2.25}$$

A random vector $\mathbf{x} \in \mathbb{R}^p$ is sub-Gaussian if $\langle \mathbf{x}, \mathbf{u} \rangle$ is sub-Gaussian for any $\mathbf{u} \in \mathbb{R}^p$, and $|||\mathbf{x}|||_{\psi_2} = \sup_{\mathbf{u} \in \mathbb{R}^p} |||\langle \mathbf{x}, \mathbf{u} \rangle|||_{\psi_2}$. A complete introduction can be found in [172]. Next we present some useful properties of sub-Gaussian random variables/vectors.

Proposition 6 (sub-Gaussian tail) *A random variable* x *satisfies the following inequality iff* $|||x|||_{\psi_2} \leq \kappa$*,*

$$\mathbb{P}\left(|x| > \epsilon\right) \leq e \cdot \exp\left(-\frac{C\epsilon^2}{\kappa^2}\right) \ , \tag{2.26}$$

where $C > 0$ *is an absolute constant.*

Proposition 7 (rotation invariance) *If $x_1, x_2, \ldots, x_n$ are independent centered sub-Gaussian random variables, then $\sum_i x_i$ is also a centered sub-Gaussian random variable with*

$$\left\|\left\|\sum_{i=1}^{n} x_i\right\|\right\|_{\psi_2}^2 \leq C^2 \sum_{i=1}^{n} \|\|x_i\|\|_{\psi_2}^2 \ , \tag{2.27}$$

where C is an absolute constant.

The rotation invariance immediately implies the well-known *Hoeffding's inequality.*

Proposition 8 (Hoeffding-type inequality) *Let $x_1, x_2, \ldots, x_n$ be independent centered sub-Gaussian r.v.s, and let $\kappa = \max_i \|\|x_i\|\|_{\psi_2}$. Then for any $\mathbf{a} = [a_1, a_2, \ldots, a_n]^T \in \mathbb{R}^n$ and $t \geq 0$, we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} a_i x_i\right| \geq \epsilon\right) \leq e \cdot \exp\left(-\frac{C\epsilon^2}{\kappa^2 \|\mathbf{a}\|_2^2}\right) \ , \tag{2.28}$$

where $C > 0$ is an absolute constant.

Proposition 9 *If $x_1, x_2, \ldots, x_p$ are independent centered sub-Gaussian random variables (not necessarily identical), then $\mathbf{x} = [x_1, \ldots, x_p]^T$ is a centered sub-Gaussian random vector with*

$$\|\|\mathbf{x}\|\|_{\psi_2} \leq C \max_{1 \leq i \leq p} \|\|x_i\|\|_{\psi_2} \ , \tag{2.29}$$

where $C > 0$ is an absolute constant.

Essentially Proposition 9 can be shown using the definition of sub-Gaussian vector and Proposition 7, which we generalize to independent sub-Gaussian vectors as follows.

Proposition 10 *If $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^m$ are independent centered sub-Gaussian random vectors, then $\mathbf{x} = [\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T]^T \in \mathbb{R}^{mn}$ is also a centered sub-Gaussian random vector* *with*

$$|||\mathbf{x}|||_{\psi_2} \leq C \max_{1\leq i\leq n} |||\mathbf{x}_i|||_{\psi_2} \ , \tag{2.30}$$

where C is an absolute constant.

Proof: Define $\mathbf{a} = [\mathbf{a}_1^T, \mathbf{a}_2^T, \ldots, \mathbf{a}_n^T]^T \in \mathbb{S}^{mn-1}$, where each $\mathbf{a}_i$ is m-dimensional. We have

$$\begin{aligned} |||\langle \mathbf{x}, \mathbf{a}\rangle|||_{\psi_2} &= \left|\left|\left| \sum_{i=1}^n \langle \mathbf{x}_i, \mathbf{a}_i\rangle \right|\right|\right|_{\psi_2} \leq \sqrt{C^2 \sum_{i=1}^n |||\langle \mathbf{x}_i, \mathbf{a}_i\rangle|||_{\psi_2}^2} \leq \sqrt{C^2 \sum_{i=1}^n \|\mathbf{a}_i\|_2^2 |||\mathbf{x}_i|||_{\psi_2}^2} \\ &\leq \sqrt{C^2 \sum_{i=1}^n \|\mathbf{a}_i\|_2^2} \cdot \max_{1\leq i\leq n} |||\mathbf{x}_i|||_{\psi_2} = C \max_{1\leq i\leq n} |||\mathbf{x}_i|||_{\psi_2} \ , \end{aligned}$$

where we use Proposition 7 for the first inequality. Based on the definition of sub-Gaussian random vector, we complete the proof. ■

A random variable x is said to be *sub-exponential* if its ψ_1*-norm* is finite, i.e.,

$$|||x|||_{\psi_1} = \sup_{q\geq 1} \frac{(\mathbb{E}|x|^q)^{\frac{1}{q}}}{q} < +\infty \ . \tag{2.31}$$

Like sub-Gaussian random variable, some useful facts about sub-exponential variable are listed below.

Proposition 11 (sub-exponential tail) *A random variable x satisfies the following inequality iff $|||x|||_{\psi_1} \leq \kappa$,*

$$\mathbb{P}\left(|x| > \epsilon\right) \leq e \cdot \exp\left(-\frac{C\epsilon}{\kappa}\right) \ , \tag{2.32}$$

where $C > 0$ is an absolute constant.

In contrast to sub-Gaussian case, rotation invariance does not hold for sub-exponential random variable, which only yields a *Bernstein-type inequality*.

Proposition 12 (Bernstein-type inequality) *Let $x_1, x_2, \ldots, x_n$ be independent centered sub-exponential random variables, and let $\kappa = \max_i |||x_i|||_{\psi_1}$. Then for any $\mathbf{a} = [a_1, a_2, \ldots, a_n]^T$ and $\epsilon \geq 0$, we have*

$$\mathbb{P}\left(\left| \sum_{i=1}^{n} a_i x_i \right| \geq \epsilon \right) \leq 2 \exp\left(-C \cdot \min\left\{ \frac{\epsilon^2}{\kappa^2 \|\mathbf{a}\|_2^2}, \frac{\epsilon}{\kappa \|\mathbf{a}\|_\infty} \right\} \right) \ , \tag{2.33}$$

where $C > 0$ is an absolute constant.

Sub-exponential and sub-Gaussian random variables are connected by the following proposition.

Proposition 13 *A random variable x is sub-Gaussian if and only if x^2 is sub-exponential. Moreover, we have*

$$|||x|||_{\psi_2}^2 \leq \left|\left|\left| x^2 \right|\right|\right|_{\psi_1} \leq 2 |||x|||_{\psi_2}^2 \tag{2.34}$$

2.4 Gaussian Width and Generic Chaining

In this section, we briefly introduce the concept of Gaussian width and the important probability tool called *generic chaining*. These topics are Interested readers are recommended to

2.4.1 Gaussian Width

Gaussian width is defined for a set $\mathcal{A} \subseteq \mathbb{R}^p$, which roughly measures its size.

Definition 11 (Gaussian width) The *Gaussian width* $w(\mathcal{A})$ of a set $\mathcal{A} \subseteq \mathbb{R}^p$ is defined as

$$w(\mathcal{A}) \triangleq \mathbb{E}\left[\sup_{\mathbf{u}\in\mathcal{A}} \langle \mathbf{u}, \mathbf{g}\rangle\right] , \tag{2.35}$$

where $\mathbf{g} \in \mathbb{R}^p$ is a standard Gaussian random vector.

The Gaussian width $w(\mathcal{A})$ provides a geometric characterization of the complexity of the set $\mathcal{A}$. We present three perspectives of view to understand the Gaussian width. First, consider the Gaussian process $\{Z_{\mathbf{u}}\}_{\mathbf{u}\in\mathcal{A}}$ where the constituent Gaussian random variables $Z_u = \langle \mathbf{u}, \mathbf{g}\rangle$ are indexed by $\mathbf{u} \in \mathcal{A}$, and $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbb{I}_{p\times p})$. Then the Gaussian width $w(\mathcal{A})$ can be viewed as the expectation of the supremum of the Gaussian process $\{Z_{\mathbf{u}}\}$. Second, $\langle \mathbf{u}, \mathbf{g}\rangle$ can be viewed as a Gaussian random projection of each $\mathbf{u} \in \mathcal{A}$ to one dimension, and the Gaussian width simply measures the expectation of largest value of such projections. Third, if $\mathcal{A}$ is the unit ball of a norm $\|\cdot\|$, i.e., $\mathcal{A} = \Omega$, then $w(\mathcal{A}) = \mathbb{E}[\|\mathbf{g}\|_*]$ by definition of the dual norm. Thus, the Gaussian width is the expected value of the dual norm of a standard Gaussian random vector. For instance, if $\mathcal{A}$ is unit ball of L_1 norm, then $w(\mathcal{A}) = \mathbb{E}[\|\mathbf{g}\|_\infty]$. Below we provide some simple yet useful properties of the Gaussian width of set $\mathcal{A} \subseteq \mathbb{R}^p$:

- (monotonicity) $\quad w(\mathcal{A}) \leq w(\mathcal{B})$ for any $\mathcal{A} \subseteq \mathcal{B}$
- (positive homogeneity) $\quad w(\mathcal{A}) = c \cdot w(\mathcal{A})$ for any $c > 0$
- (convexification invariance) $\quad w(\mathcal{A}) = w(\mathrm{conv}(\mathcal{A}))$
- (rotation invariance) $\quad w(\mathbf{U}\mathcal{A}) = w(\mathcal{A})$ for any unitary matrix $\mathbf{U} \in \mathbb{R}^{p\times p}$
- (translation invariance) $\quad w(\mathcal{A} + \mathbf{b}) = w(\mathcal{A})$ for any fixed $\mathbf{b} \in \mathbb{R}^p$

The following result for Gaussian width is useful when we deal with union of sets, which is extracted from Lemma 2 in [118].

Lemma 1 (Gaussian width for union of sets) *Let* $M > 4$, $\mathcal{A}_1, \cdots, \mathcal{A}_M \subset \mathbb{R}^p$, *and* $\mathcal{A} = \cup_m \mathcal{A}_m$. *The Gaussian width of* $\mathcal{A}$ *satisfies*

$$w(\mathcal{A}) \leq \max_{1 \leq m \leq M} w(\mathcal{A}_m) + 2 \sup_{\mathbf{z} \in \mathcal{A}} \|\mathbf{z}\|_2 \sqrt{\log M} \tag{2.36}$$

The concept of Gaussian width can be directly extended to the matrix setting, and $w(\mathcal{A})$ for set $\mathcal{A} \subseteq \mathbb{R}^{d \times p}$ is given as

$$w(\mathcal{A}) \triangleq \mathbb{E}_{\mathbf{G}} \left[\sup_{\mathbf{Z} \in \mathcal{A}} \langle\langle \mathbf{G}, \mathbf{Z} \rangle\rangle \right] , \tag{2.37}$$

in which $\langle\langle \cdot, \cdot \rangle\rangle$ denotes the matrix inner product, i.e., $\langle\langle \mathbf{A}, \mathbf{B} \rangle\rangle = \mathrm{Tr}(\mathbf{A}^T \mathbf{B})$ for any $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d \times p}$. Here $\mathbf{G} \in \mathbb{R}^{d \times p}$ is a random matrix with i.i.d. standard Gaussian entries, i.e., $G_{ij} \sim \mathcal{N}(0, 1)$. The aforementioned properties also hold for the matrix case.

2.4.2 Generic Chaining

One important tool that we use in our probabilistic argument is *generic chaining* [161, 162], which is powerful for bounding the suprema of stochastic processes. Suppose $\{Z_{\mathbf{t}}\}_{\mathbf{t} \in \mathcal{T}}$ is a centered stochastic process, where each $Z_{\mathbf{t}}$ is a centered random variable. We assume the index set $\mathcal{T}$ is endowed with some metric (distance function) $s(\cdot, \cdot)$. A key notion in generic chaining is γ_2-functional $\gamma_2(\mathcal{T}, s)$, which is defined for the metric space $(\mathcal{T}, s)$. One can think of γ_2-functional as a measure of the size of set $\mathcal{T}$ w.r.t. metric s. For self-containedness, we give the expression of $\gamma_2(\mathcal{T}, s)$.

$$\gamma_2(\mathcal{T}, s) = \inf_{\{\mathcal{P}_n\}} \sup_{\mathbf{t} \in \mathcal{T}} \sum_{n \geq 0} 2^{n/2} \cdot \mathrm{diam}\left(\mathcal{P}_n(\mathbf{t}), s\right) , \tag{2.38}$$

where $\{\mathcal{P}_n\}_{n=0}^{\infty} = \{\mathcal{P}_0, \mathcal{P}_1, \ldots, \mathcal{P}_n, \ldots\}$ is a sequence of partitions for $\mathcal{T}$, which satisfy that $|\mathcal{P}_0| = 1$, $|\mathcal{P}_n| \leq 2^{2^n}$ for $n \geq 1$, and that $\mathcal{P}_{n+1}$ is a finer partition than $\mathcal{P}_n$, i.e., every

$\mathcal{Q} \in \mathcal{P}_{n+1}$ is a subset of some $\mathcal{Q}' \in \mathcal{P}_n$. $\mathcal{P}_n(\mathbf{t})$ denotes the subset of $\mathcal{T}$ that contains $\mathbf{t}$ in the n-th partition, and $\operatorname{diam}(\mathcal{P}_n(\mathbf{t}), s)$ measures the diameter of $\mathcal{P}_n(\mathbf{t})$ w.r.t. metric $s(\cdot,\cdot)$. Note that γ_2-functional is a purely geometric concept, which involves no probability. Given that γ_2-functional is fairly involved, we are not going to discuss any insights behind this definition, and refer interested readers to the introductory books [161, 162]. Based on its definition, we list a few straightforward properties of γ_2-functional here.

$$\gamma_2(\mathcal{T}, s_1) \le \gamma_2(\mathcal{T}, s_2) \quad \text{if } s_1(\mathbf{u},\mathbf{v}) \le s_2(\mathbf{u},\mathbf{v}), \forall\ \mathbf{u},\mathbf{v} \in \mathcal{T} \tag{2.39}$$

$$\gamma_2(\mathcal{T}, \beta s) = \beta \cdot \gamma_2(\mathcal{T}, s) \quad \text{for any } \beta > 0 \ . \tag{2.40}$$

$$\gamma_2(\mathcal{T}_1, s_1) = \gamma_2(\mathcal{T}_2, s_2) \quad \text{if } \exists \text{ a global isometry between } (\mathcal{T}_1, s_1) \text{ and } (\mathcal{T}_2, s_2) \tag{2.41}$$

The following lemma concerned with the suprema of $\{Z_\mathbf{t}\}$ combines Theorem 2.2.22 and 2.2.27 from [162].

Lemma 2 *Given metric space $(\mathcal{T}, s)$, if the associated centered stochastic process $\{Z_\mathbf{t}\}_{\mathbf{t}\in\mathcal{T}}$ satisfies the condition*

$$\mathbb{P}\left(|Z_\mathbf{u} - Z_\mathbf{v}| \ge \epsilon\right) \ \le \ C_0 \exp\left(-\frac{C_1\epsilon^2}{s^2(\mathbf{u},\mathbf{v})}\right), \quad \forall \epsilon > 0 \ \textit{and}\ \mathbf{u},\mathbf{v} \in \mathcal{T} \ , \tag{2.42}$$

then the following inequalities hold

$$\mathbb{E}\left[\sup_{\mathbf{t}\in\mathcal{T}} Z_\mathbf{t}\right] \le C_2 \gamma_2\left(\mathcal{T}, s\right) \ , \tag{2.43}$$

$$\mathbb{P}\left(\sup_{\mathbf{u},\mathbf{v}\in\mathcal{T}} |Z_\mathbf{u} - Z_\mathbf{v}| \ge C_3\left(\gamma_2(\mathcal{T}, s) + \epsilon \cdot \operatorname{diam}(\mathcal{T}, s)\right)\right) \le C_4 \exp\left(-\epsilon^2\right) \ , \tag{2.44}$$

where C_0, C_1, C_2, C_3 and C_4 are all absolute constants.

Another useful result based on generic chaining is the Theorem D in [125].

Lemma 3 (Theorem D in [125]) *There exist absolute constants C_1, C_2 for which the following holds. Let (Ω, μ) be a probability space on which x is defined, and $x_1, \ldots, x_n$ be independent copies of x. Let set $\mathcal{H}$ be a subset of the unit sphere of $L_2(\mu)$, i.e.,*

$$\mathcal{H} \subseteq \mathbb{S}_{L_2} = \left\{ h \ : \ \left\|\left|h\right|\right\|_{L_2} = \sqrt{\int_{\Omega} h^2(x) dx} = 1 \right\} \ , \tag{2.45}$$

Assume that $\sup_{h \in \mathcal{H}} \ \left\|\left|h\right|\right\|_{\psi_2} \leq \kappa$. Then, for any $\beta > 0$ and $n \geq 1$ satisfying

$$C_1 \kappa \cdot \gamma_2(\mathcal{H}, \left\|\left|\cdot\right|\right\|_{\psi_2}) \leq \beta \sqrt{n} \ , \tag{2.46}$$

with probability at least $1 - \exp\left(-\frac{C_2 \beta^2 n}{\kappa^4}\right)$,

$$\sup_{h \in \mathcal{H}} \ \left| \frac{1}{n} \sum_{i=1}^{n} h^2(X_i) - \mathbb{E}\left[h^2\right] \right| \leq \beta \ . \tag{2.47}$$

The suprema in both Lemma 2 and 3 are characterized in terms of γ_2-functional, which is not easily computable. In order to further bound the γ_2-functional, one needs the so-called *majorizing measures theorem* [160].

Lemma 4 *Given any Gaussian process $\{Y_{\mathbf{t}}\}_{\mathbf{t} \in \mathcal{T}}$, define $s(\mathbf{u}, \mathbf{v}) = \sqrt{\mathbb{E}|Y_{\mathbf{u}} - Y_{\mathbf{v}}|^2}$ for $\mathbf{u}, \mathbf{v} \in \mathcal{T}$. Then $\gamma_2(\mathcal{T}, s)$ can be upper bounded by*

$$\gamma_2(\mathcal{T}, s) \leq C_0 \mathbb{E}\left[\sup_{\mathbf{t} \in \mathcal{T}} Y_{\mathbf{t}}\right] \ , \tag{2.48}$$

where C_0 is an absolute constant.

In the analysis, we usually instantiate this lemma by constructing the simple Gaussian process $\{Y_{\mathbf{t}} = \langle \mathbf{t}, \mathbf{g} \rangle\}_{\mathbf{t} \in \mathcal{T}}$ for any $\mathcal{T} \subseteq \mathbb{R}^p$, where $\mathbf{g}$ is a standard Gaussian random vector. Hence $s(\mathbf{u}, \mathbf{v}) = \sqrt{\mathbb{E}|Y_{\mathbf{u}} - Y_{\mathbf{v}}|^2} = \sqrt{\mathbb{E}|\langle \mathbf{u} - \mathbf{v}, \mathbf{g} \rangle|^2} = \|\mathbf{u} - \mathbf{v}\|_2$. It follows from

Lemma 4 that

$$\gamma_2\left(\mathcal{T}, \|\cdot\|_2\right) \leq C_0 \mathbb{E}\left[\sup_{\mathbf{t}\in\mathcal{T}} \langle \mathbf{t}, \mathbf{g}\rangle\right] = C_0 \cdot w(\mathcal{T}) \ , \tag{2.49}$$

which makes the connection between γ_2-functional and Gaussian width. For matrix setting, we can also get such connection by a similar construction of the Gaussian process,

$$\gamma_2\left(\mathcal{A}, \|\cdot\|_F\right) \leq C_0 \mathbb{E}\left[\sup_{\mathbf{Z}\in\mathcal{A}} \langle\langle \mathbf{G}, \mathbf{Z}\rangle\rangle\right] = C_0 \cdot w(\mathcal{A}) \ , \tag{2.50}$$

where the set $\mathcal{A} \in \mathbb{R}^{d\times p}$.

combining Lemma 2 and 4, we can get the following theorem, which is more amenable to some of the proofs.

Theorem 1 *Let* $\{Z_{\mathbf{t}}\}_{\mathbf{t}\in\mathcal{T}}$ *be a stochastic process indexed by* $\mathcal{T} \subseteq \mathbb{R}^p$*, which satisfies*

$$\sup_{\mathbf{t},\mathbf{t}'\in\mathcal{T}} \frac{|||Z_{\mathbf{t}} - Z_{\mathbf{t}'}|||_{\psi_2}}{\|\mathbf{t} - \mathbf{t}'\|_2} \leq K < +\infty$$

There exist absolute constants C_0 *and* C_1 *such that the following bound holds with probability at least* $1 - C_1 \exp\left(-\frac{w^2(\mathcal{T})}{\mathrm{diam}^2(\mathcal{T})}\right)$,

$$\sup_{\mathbf{t},\mathbf{t}'\in\mathcal{T}} |Z_{\mathbf{t}} - Z_{\mathbf{t}'}| \leq C_0 K \cdot w(\mathcal{T}) \ , \tag{2.51}$$

where $\mathrm{diam}\,(\mathcal{T}) = \sup_{\mathbf{t},\mathbf{t}'\in\mathcal{T}} \|\mathbf{t} - \mathbf{t}'\|_2$.

In the analysis, sometimes we need to bound product processes, which can be dealt with by the following theorem. The result is essentially a simplified form of Theorem 1.13 in [124]. The original theorem is stated in terms of a variant of the γ_2-functional defined above, and contains a few more tunable variables, both of which are not central to the core idea and thus have been hidden. The bound here is expressed using Gaussian width.

Theorem 2 *Let (Ω, μ) be a probability space, and $Z_1, Z_2, \ldots, Z_n$ be an i.i.d. sample distributed according to μ. Suppose that $\mathcal{F} = \{f_{\mathbf{a}}\}_{\mathbf{a}\in\mathcal{A}}$ and $\mathcal{H} = \{h_{\mathbf{b}}\}_{\mathbf{b}\in\mathcal{B}}$ are two function classes defined on (Ω, μ), which are indexed by $\mathcal{A} \subseteq \mathbb{R}^p$ and $\mathcal{B} \subseteq \mathbb{R}^q$ respectively. Assume that*

$$\sup_{f\in\mathcal{F}} |||f|||_{\psi_2} \leq R_{\mathcal{F}} < +\infty \ , \qquad \sup_{h\in\mathcal{H}} |||h|||_{\psi_2} \leq R_{\mathcal{H}} < +\infty \ ,$$
$$\sup_{\mathbf{a},\mathbf{a}'\in\mathcal{A}} \frac{|||f_{\mathbf{a}} - f_{\mathbf{a}'}|||_{\psi_2}}{\|\mathbf{a} - \mathbf{a}'\|_2} \leq K_{\mathcal{F}} < +\infty \ , \qquad \sup_{\mathbf{b},\mathbf{b}'\in\mathcal{B}} \frac{|||h_{\mathbf{b}} - h_{\mathbf{b}'}|||_{\psi_2}}{\|\mathbf{b} - \mathbf{b}'\|_2} \leq K_{\mathcal{H}} < +\infty \ ,$$

and denote

$$\varepsilon = \min\left\{\frac{K_{\mathcal{F}} \cdot w(\mathcal{A})}{R_{\mathcal{F}}}, \ \frac{K_{\mathcal{H}} \cdot w(\mathcal{B})}{R_{\mathcal{H}}}\right\} \ .$$

There exist absolute constants C_0, C_1 and C_2 such that if $n \geq C_0\varepsilon^2$, the following inequality holds with probability at least $1 - 2\exp\left(-C_1\varepsilon^2\right)$,

$$\sup_{f\in\mathcal{F}} \sup_{h\in\mathcal{H}} \left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i)h(Z_i) - \mathbb{E}\left[fh\right]\right| \ \leq \ C_2 \cdot \frac{R_{\mathcal{H}}K_{\mathcal{F}} \cdot w(\mathcal{A}) + R_{\mathcal{F}}K_{\mathcal{H}} \cdot w(\mathcal{B})}{\sqrt{n}} \tag{2.52}$$

The theorem above immediately leads to the following corollary, which is similar to Lemma 3 but more flexible in some situations.

Corollary 1 *Under the setting of Theorem 2, if $\mathcal{F} = \mathcal{H}$ and $\mathcal{A} = \mathcal{B}$, then there exist absolute constants C_0, C_1 and C_2 such that if $n \geq C_0\left(\frac{K_{\mathcal{F}}\cdot w(\mathcal{A})}{R_{\mathcal{F}}}\right)^2$, the following inequality holds with probability at least $1 - 2\exp\left(-C_1\left(\frac{K_{\mathcal{F}}\cdot w(\mathcal{A})}{R_{\mathcal{F}}}\right)^2\right)$,*

$$\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} f^2(Z_i) - \mathbb{E}\left[f^2\right]\right| \ \leq \ C_2 \cdot \frac{R_{\mathcal{F}}K_{\mathcal{F}} \cdot w(\mathcal{A})}{\sqrt{n}} \tag{2.53}$$

Chapter 3

Generalized Dantzig Selector

3.1 Introduction

The Dantzig Selector (DS) [21, 32] provides an alternative to regularized regression approaches such as Lasso [165, 188] for sparse linear estimation. While DS does not consider a regularized maximum likelihood approach, [21] has established clear similarities between the estimates from DS and Lasso. While norm regularized regression approaches have been generalized to more general norms, such as decomposable norms [127], the literature on DS has primarily focused on the sparse L_1 norm case, with a few notable exceptions which have considered extensions to sparse group-structured norms [112]. Here we consider linear models of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta}^* + \boldsymbol{\epsilon} \ , \tag{3.1}$$

where $\mathbf{y} \in \mathbb{R}^n$ is a set of observations, $\mathbf{X} \in \mathbb{R}^{n\times p}$ is a design matrix, and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is a noise vector of i.i.d. entries. For *any* given norm $\|\cdot\|$, the parameter $\boldsymbol{\theta}^*$ is assumed to be structured so that $\|\boldsymbol{\theta}^*\|$ is of small value. For this setting, we propose the following

Generalized Dantzig Selector (GDS) for parameter estimation:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\operatorname{argmin}} \ \|\boldsymbol{\theta}\| \quad \text{s.t.} \quad \left\| \frac{1}{n} \mathbf{X}^T \left(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\right) \right\|_* \leq \lambda_n \ , \tag{3.2}$$

where $\|\cdot\|_*$ is the *dual norm* of $\|\cdot\|$, and λ_n is a tuning parameter. If $\|\cdot\|$ is the L_1 norm, (3.2) reduces to standard DS [32]. A key novel aspect of GDS is that the constraint is in terms of the dual norm $\|\cdot\|_*$ of the original structure inducing norm $\|\cdot\|$. It is instructive to contrast GDS with the recently proposed atomic norm based estimation framework [40] which, unlike GDS, considers constraints based on the L_2 norm of the error $\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2$.

In this chapter, we consider both computational and statistical aspects of the GDS. For the L_1-norm Dantzig selector, [32] proposed a primal-dual interior point method since the optimization is a linear program. DASSO and its generalization proposed in [89, 90] focused on homotopy methods, which provide a piecewise linear solution path through a sequential simplex-like algorithm. However, none of the algorithms above can be immediately extended to our general formulation. In recent work, the alternating direction method of multipliers (ADMM) has been applied to the L_1 Dantzig selection problem [114,176], and the linearized version in [176] proved to be efficient. Motivated by such results for DS, we propose a general inexact ADMM [175] framework for GDS where the primal update steps, interestingly, turn out respectively to be proximal operators involving $\|\boldsymbol{\theta}\|$ and its convex conjugate, the indicator function of the norm ball. As a result, by Moreau decomposition, it suffices to develop efficient proximal update for either $\|\boldsymbol{\theta}\|$ or its conjugate. As a non-trivial example, we consider estimation using the recently proposed k-support norm [7, 120]. We show that proximal operators for k-support norm can be efficiently computed in $O(p \log p + \log k \log(p-k))$ time, and hence the estimation can be done efficiently. Note that existing work [7, 120] on k-support norm has focused on the proximal operator for the *square* of the k-support norm, which is not directly applicable in our setting.

On the statistical side, we establish non-asymptotic high-probability bounds on the estimation error $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2$. Interestingly, the bound depends on the *Gaussian width* of the unit norm ball of $\|\cdot\|$ as well as the Gaussian width of suitable set where the estimation error belongs [40, 138]. Besides, the maximum ratio between $\|\cdot\|$ and $\|\cdot\|_2$ over this set also plays a role, which is termed *restricted norm compatibility*.

The rest of the chapter is organized as follows. We propose general optimization method for GDS in Section 3.2, along with an efficient algorithm to compute the proximal operator for k-support norm. In Section 3.3, we present the L_2-error bounds for GDS. Experimental results are provided in Section 3.4.

3.2 Optimization Algorithm

The optimization problem in (3.2) is a convex program, and a suitable choice of λ_n ensures that the feasible set is not empty. We start with an inexact ADMM framework for solving problems of the form (3.2), and then present the algorithm for computing proximal operator for the k-support norms.

3.2.1 Inexact ADMM for GDS

In optimization, we temporarily drop the subscript n of λ_n for convenience. We let $\mathbf{A} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$, $\mathbf{u} = \frac{1}{n}\mathbf{X}^T\mathbf{y}$, and define the set $\Omega_\lambda^* = \{\mathbf{v} \ : \ \|\mathbf{v}\|_* \leq \lambda\}$ as the scaled ball of dual norm. Then the optimization problem is equivalent to

$$\min_{\boldsymbol{\theta}, \mathbf{v} \in \mathbb{R}^p} \ \|\boldsymbol{\theta}\| \quad \text{s.t.} \quad \mathbf{u} - \mathbf{A}\boldsymbol{\theta} = \mathbf{v}, \ \mathbf{v} \in \Omega_\lambda^* \ . \tag{3.3}$$

Due to the nonsmoothness of both $\|\cdot\|$ and $\|\cdot\|_*$, a generally applicable algorithm is alternating direction method of multipliers (ADMM), which we briefly reviewed in Section 2.2. The augmented Lagrangian for (3.3) is given as

$$L_\rho(\boldsymbol{\theta}, \mathbf{v}, \mathbf{z}) = \|\boldsymbol{\theta}\| + \langle \mathbf{z}, \mathbf{A}\boldsymbol{\theta} + \mathbf{v} - \mathbf{u}\rangle + \frac{\rho}{2}\|\mathbf{A}\boldsymbol{\theta} + \mathbf{v} - \mathbf{u}\|_2^2 \ , \tag{3.4}$$

in which $\mathbf{z}$ is the dual variable and ρ controls the penalty introduced by the quadratic term. The iterative updates of the variables $(\boldsymbol{\theta}, \mathbf{v}, \mathbf{z})$ in standard ADMM are given by

$$\boldsymbol{\theta}^{t+1} \leftarrow \operatorname{argmin}_{\boldsymbol{\theta}} L_\rho(\boldsymbol{\theta}, \mathbf{v}^t, \mathbf{z}^t) \ , \tag{3.5}$$

$$\mathbf{v}^{t+1} \leftarrow \operatorname{argmin}_{\mathbf{v}\in\Omega_\lambda^*} L_\rho(\boldsymbol{\theta}^{t+1}, \mathbf{v}, \mathbf{z}^t) \ , \tag{3.6}$$

$$\mathbf{z}^{t+1} \leftarrow \mathbf{z}^t + \rho(\mathbf{A}\boldsymbol{\theta}^{t+1} + \mathbf{v}^{t+1} - \mathbf{u}) \ . \tag{3.7}$$

Note that update (3.5) amounts to a regularized least squares problem of $\boldsymbol{\theta}$, which can be computationally expensive. Thus we use an inexact update for $\boldsymbol{\theta}$ instead, which can alleviate the computational cost and lead to a simple algorithm. Inspired by [176], we consider a simpler subproblem for the $\boldsymbol{\theta}$-update which minimizes

$$\begin{aligned}\tilde{L}_\rho^t(\boldsymbol{\theta}, \mathbf{v}^t, \mathbf{z}^t) = \|\boldsymbol{\theta}\| + \left\langle \mathbf{z}^t, \mathbf{A}\boldsymbol{\theta} + \mathbf{v}^t - \mathbf{u}\right\rangle + \frac{\rho}{2}\Big(\left\|\mathbf{A}\boldsymbol{\theta}^t + \mathbf{v}^t - \mathbf{u}\right\|_2^2 + \\ 2\left\langle \boldsymbol{\theta} - \boldsymbol{\theta}^t, \mathbf{A}^T(\mathbf{A}\boldsymbol{\theta}^t + \mathbf{v}^t - \mathbf{u})\right\rangle + \frac{\mu}{2}\left\|\boldsymbol{\theta} - \boldsymbol{\theta}^t\right\|_2^2 \Big) \ ,\end{aligned} \tag{3.8}$$

where μ is a user-defined parameter. $\tilde{L}_\rho^t(\boldsymbol{\theta}, \mathbf{v}^t, \mathbf{z}^t)$ can be viewed as an approximation of $L_\rho(\boldsymbol{\theta}, \mathbf{v}^t, \mathbf{z}^t)$ with the quadratic term linearized at $\boldsymbol{\theta}^t$. Then the update (3.5) is replaced by

$$\begin{aligned}\boldsymbol{\theta}^{t+1} &\leftarrow \operatorname{argmin}_{\boldsymbol{\theta}} \tilde{L}_\rho^t(\boldsymbol{\theta}, \mathbf{v}^t, \mathbf{z}^t) \\ &= \operatorname{argmin}_{\boldsymbol{\theta}} \left\{ \frac{2}{\rho\mu} \cdot \|\boldsymbol{\theta}\| + \frac{1}{2}\left\|\boldsymbol{\theta} - \big(\boldsymbol{\theta}^t - \frac{2}{\mu}\mathbf{A}^T(\mathbf{A}\boldsymbol{\theta}^t + \mathbf{v}^t - \mathbf{u} + \frac{\mathbf{z}^t}{\rho})\big)\right\|_2^2 \right\} \ .\end{aligned} \tag{3.9}$$

Algorithm 4 Inexact ADMM for Generalized Dantzig Selector

Input: $\mathbf{A} = \mathbf{X}^T\mathbf{X}$, $\mathbf{u} = \mathbf{X}^T\mathbf{y}$, $\rho > 0$, $\mu > 0$
Output: Minimizer $\hat{\boldsymbol{\theta}}$ of (3.2)
1: Initialize $(\boldsymbol{\theta}^0, \mathbf{v}^0, \mathbf{z}^0)$
2: **while** not converged **do**
3: $\quad \boldsymbol{\theta}^{t+1} \leftarrow \mathbf{prox}_{\frac{2\|\cdot\|}{\rho\mu}}\left(\boldsymbol{\theta}^t - \frac{2}{\mu}\mathbf{A}^T\left(\mathbf{A}\boldsymbol{\theta}^t + \mathbf{v}^t - \mathbf{u} + \frac{\mathbf{z}^t}{\rho}\right)\right)$
4: $\quad \mathbf{v}^{t+1} \leftarrow \mathbf{prox}_{\mathbb{I}_{\mathcal{C}_\lambda}}\left(\mathbf{u} - \mathbf{A}\boldsymbol{\theta}^{t+1} - \frac{\mathbf{z}^t}{\rho}\right)$
5: $\quad \mathbf{z}^{t+1} \leftarrow \mathbf{z}^t + \rho\left(\mathbf{A}\boldsymbol{\theta}^{t+1} + \mathbf{v}^{t+1} - \mathbf{u}\right)$
6: **end while**

Similarly the update of $\mathbf{v}$ in (3.6) can be recast as

$$\mathbf{v}^{t+1} \leftarrow \operatorname{argmin}_{\mathbf{v}\in\Omega^*_\lambda} L_\rho(\boldsymbol{\theta}^{t+1}, \mathbf{v}, \mathbf{z}^t) = \operatorname{argmin}_{\mathbf{v}\in\Omega^*_\lambda} \frac{1}{2}\left\|\mathbf{v} - \left(\mathbf{u} - \mathbf{A}\boldsymbol{\theta}^{t+1} - \frac{\mathbf{z}^t}{\rho}\right)\right\|_2^2 . \tag{3.10}$$

In fact, the updates of $\boldsymbol{\theta}$ and $\mathbf{v}$ correspond to $\mathbf{prox}_{\frac{2\|\cdot\|}{\rho\mu}}(\cdot)$ and $\mathbf{prox}_{\mathbb{I}_{\mathcal{C}_\lambda}}(\cdot)$, respectively, which are proximal operators introduced in Section 2.2. Algorithm 4 provides the general ADMM for our GDS. In order for the ADMM to work, we need two subroutines that can efficiently compute the proximal operators in Line 3 and 4. The simplicity of the proposed approach stems from the fact that we in fact need *only one* subroutine, for any one of the functions, since the functions are conjugates of each other.

Proposition 14 *Given $\beta > 0$ and a norm $\|\cdot\|$, the two functions, $f(\mathbf{x}) = \beta\|\mathbf{x}\|$ and $g(\mathbf{x}) = \mathbb{I}_{\mathcal{C}_\beta}(\mathbf{x})$ are convex conjugate to each other, thus giving the following identity,*

$$\mathbf{x} = \mathbf{prox}_f(\mathbf{x}) + \mathbf{prox}_g(\mathbf{x}) . \tag{3.11}$$

Proof: the proposition simply follows the definition of convex conjugate and dual norm, and (3.11) is just *Moreau decomposition* provided in [133]. ■

The decomposition enables conversion of the two types of proximal operator to each other at negligible cost (i.e., vector subtraction). Thus we have the flexibility in

Algorithm 4 to focus on the proximal operator that is easier to compute, and the other can be simply obtained through (3.11).

Remark on convergence: Note that Algorithm 4 is a special case of inexact Bregman ADMM proposed in [175], which matches the case of linearizing quadratic penalty term by using $B_{\varphi'_{\boldsymbol{\theta}}}(\boldsymbol{\theta}, \boldsymbol{\theta}_t) = \frac{1}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_t\|_2^2$ as Bregman divergence. In order to converge, the algorithm requires $\frac{\mu}{2}$ to be larger than the spectral radius of $\mathbf{A}^T\mathbf{A}$, and the convergence rate is $O(1/T)$ according to Theorem 2 in [175].

3.2.2 Proximal Operator for k-Support Norm

We first introduce some notations. Given any $\boldsymbol{\theta} \in \mathbb{R}^p$, let $|\boldsymbol{\theta}|$ denote its absolute-valued counterpart and $\boldsymbol{\theta}^{\downarrow}$ denote the permutation of $\boldsymbol{\theta}$ with its elements arranged in decreasing order. In previous work [7, 120], the *k-support norm* is defined as

$$\|\boldsymbol{\theta}\|_k^{sp} = \min\left\{ \sum_{\mathcal{I}\in\mathcal{G}^{(k)}} \|\mathbf{v}_{\mathcal{I}}\|_2 \;\middle|\; \operatorname{supp}(\mathbf{v}_{\mathcal{I}}) \subseteq \mathcal{I}, \; \sum_{\mathcal{I}\in\mathcal{G}^{(k)}} \mathbf{v}_{\mathcal{I}} = \boldsymbol{\theta} \right\}, \tag{3.12}$$

where $\mathcal{G}^{(k)}$ denotes the set that includes all subsets of $\{1, \ldots, p\}$ of cardinality at most k. The unit ball of this norm is the set $\Omega_k^{sp} = \operatorname{conv}\left(\{\boldsymbol{\theta} \in \mathbb{R}^p \mid \|\boldsymbol{\theta}\|_0 \le k, \|\boldsymbol{\theta}\|_2 \le 1\}\right)$. The dual norm of the k-support norm is given by

$$\|\boldsymbol{\theta}\|_{k*}^{sp} = \max\left\{ \|\boldsymbol{\theta}_{\mathcal{I}}\|_2 \;\middle|\; \mathcal{I} \in \mathcal{G}^{(k)} \right\} = \left(\sum_{i=1}^{k} |\boldsymbol{\theta}|_i^{\downarrow 2} \right)^{\frac{1}{2}} . \tag{3.13}$$

Solving GDS with k-support norm $\|\cdot\|_k^{sp}$ requires that either $\mathbf{prox}_{\lambda\|\cdot\|_k^{sp}}(\cdot)$ or $\mathbf{prox}_{\mathbb{I}_{\mathcal{C}_\lambda}}(\cdot)$ for $\|\cdot\|_{k*}^{sp}$ is efficiently computable. Existing methods [7, 120] are inapplicable to our scenario since they compute the proximal operator for squared k-support norm, from which $\mathbf{prox}_{\mathbb{I}_{\mathcal{C}_\lambda}}(\cdot)$ cannot be directly obtained. In Theorem 3, we show that $\mathbf{prox}_{\mathbb{I}_{\mathcal{C}_\lambda}}(\cdot)$ can be efficiently computed, and thus Algorithm 4 is applicable.

Theorem 3 *Given* $\lambda > 0$ *and* $\mathbf{x} \in \mathbb{R}^p$*, if* $\|\mathbf{x}\|_{k*}^{sp} \leq \lambda$*, then* $\mathbf{w}^* = \mathbf{prox}_{\mathbb{I}_{\mathcal{C}_\lambda}}(\mathbf{x}) = \mathbf{x}$*. If* $\|\mathbf{x}\|_{k*}^{sp} > \lambda$*, define* $A_{sr} = \sum_{i=s+1}^{r} |\mathbf{x}|_i^{\downarrow}$*,* $B_s = \sum_{i=1}^{s}(|\mathbf{x}|_i^{\downarrow})^2$*, in which* $0 \leq s < k$ *and* $k \leq r \leq p$*, and construct the nonlinear equation of* β*,*

$$(k-s)A_{sr}^2 \left[\frac{1+\beta}{r-s+(k-s)\beta}\right]^2 - \lambda^2(1+\beta)^2 + B_s = 0 \ . \tag{3.14}$$

Let β_{sr} *be given by*

$$\beta_{sr} = \begin{cases} \text{nonnegative root of (3.14)} & \text{if } s > 0 \text{ and the root exists} \\ 0 & \text{otherwise} \end{cases} \ . \tag{3.15}$$

Then the proximal operator $\mathbf{w}^* = \mathbf{prox}_{\mathbb{I}_{\mathcal{C}_\lambda}}(\mathbf{x})$ *is given by*

$$|\mathbf{w}^*|_i^{\downarrow} = \begin{cases} \frac{1}{1+\beta_{s^*r^*}}|\mathbf{x}|_i^{\downarrow} & \text{if } 1 \leq i \leq s^* \\ \sqrt{\frac{\lambda^2 - B_{s^*}}{k-s^*}} & \text{if } s^* < i \leq r^* \text{ and } \beta_{s^*r^*} = 0 \\ \frac{A_{s^*r^*}}{r^*-s^*+(k-s^*)\beta_{s^*r^*}} & \text{if } s^* < i \leq r^* \text{ and } \beta_{s^*r^*} > 0 \\ |\mathbf{x}|_i^{\downarrow} & \text{if } r^* < i \leq p \end{cases} , \tag{3.16}$$

where the indices s^* *and* r^* *with computed* $|\mathbf{w}^*|^{\downarrow}$ *make the following two inequalities hold,*

$$|\mathbf{w}^*|_{s^*}^{\downarrow} > |\mathbf{w}^*|_k^{\downarrow} \ , \tag{3.17}$$

$$|\mathbf{x}|_{r^*+1}^{\downarrow} \leq |\mathbf{w}^*|_k^{\downarrow} < |\mathbf{x}|_{r^*}^{\downarrow} \ . \tag{3.18}$$

There might be multiple pairs of (s, r) *satisfying the inequalities* (3.17)-(3.18)*, and we choose the pair with the smallest* $\||\mathbf{x}|^{\downarrow} - |\mathbf{w}|^{\downarrow}\|_2$*. Finally,* $\mathbf{w}^*$ *is obtained by sign-changing* *and reordering* $|\mathbf{w}^*|^{\downarrow}$ *to conform to* $\mathbf{x}$*.*

Remark: The nonlinear equation (3.14) is quartic, for which we can use general formula to get all the roots [155]. In addition, if it exists, the nonnegative root is unique, as we show in the proof.

Theorem 3 indicates that computing $\mathbf{prox}_{\mathbb{I}_{\mathcal{C}_\lambda}}(\cdot)$ requires sorting of entries in $|\mathbf{x}|$ and a two-dimensional grid search of s^* and r^*. Hence the total time complexity is $O(p \log p + k(p-k))$. However, a more careful observation can particularly reduce the search complexity from $O(k(p-k))$ to $O(\log k \log(p-k))$, which is motivated by Theorem 4.

Theorem 4 *In search of* (s^*, r^*) *defined in Theorem 3, there can be only one* $\tilde{r}$ *for a given candidate* $\tilde{s}$ *of* s^**, such that the inequality* (3.18) *is satisfied. Moreover if such* $\tilde{r}$ *exists, then for any* $r < \tilde{r}$*, the associated* $|\tilde{\mathbf{w}}|_k^{\downarrow}$ *violates the first part of* (3.18)*, and for* $r > \tilde{r}$*,* $|\tilde{\mathbf{w}}|_k^{\downarrow}$ *violates the second part of* (3.18)*. On the other hand, based on the* $\tilde{r}$*, we have following assertion of* s^**,*

$$s^* \begin{cases} > \tilde{s} & \text{if } \tilde{r} \text{ does not exist} \\ \geq \tilde{s} & \text{if } \tilde{r} \text{ exists and the corresponding } |\tilde{\mathbf{w}}|_k^{\downarrow} \text{ satisfies (3.17)} \\ < \tilde{s} & \text{if } \tilde{r} \text{ exists but the corresponding } |\tilde{\mathbf{w}}|_k^{\downarrow} \text{ violates (3.17)} \end{cases} . \tag{3.19}$$

Based on Theorem 4, the accelerated search procedure of (s^*, r^*) is to execute a two-dimensional *binary search*, and Algorithm 5 gives the details. Therefore the overall time complexity becomes $O(p \log p + \log k \cdot \log(p-k))$. Compared with previous proximal operators for squared k-support norm, this complexity is better than that in [7], and roughly the same as the most recent one in [120].

Algorithm 5 Algorithm for computing $\mathbf{prox}_{\mathbb{I}_{\mathcal{C}_\lambda}}(\cdot)$ of $\|\cdot\|_{k*}^{sp}$

Input: $\mathbf{x}$, k, λ
Output: $\mathbf{w}^* = \mathbf{prox}_{\mathbb{I}_{\mathcal{C}_\lambda}}(\mathbf{x})$

if $\|\mathbf{x}\|_{k*}^{sp} \leq \lambda$ **then**
  $\mathbf{w}^* := \mathbf{x}$
else
  $l := 0$, $u := k-1$, and sort $|\mathbf{x}|$ to get $|\mathbf{x}|^{\downarrow}$
  while $l \leq u$ **do**
    $\tilde{s} := \lfloor (l+u)/2 \rfloor$, and binary search for $\tilde{r}$ that satisfies (3.18) and compute $\tilde{\mathbf{w}}$ based on (3.16)
    if $\tilde{r}$ does not exist **then**
      $l := \tilde{s} + 1$
    else if $\tilde{r}$ exists and (3.17) is satisfied **then**
      $\mathbf{w}^* := \tilde{\mathbf{w}}$, $l := \tilde{s} + 1$
    else if $\tilde{r}$ exists but (3.17) is not satisfied **then**
      $u := \tilde{s} - 1$
    end if
  end while
end if

3.3 Statistical Analysis

3.3.1 Deterministic Error Bound

Our goal is to provide error bounds on $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2$ between the population parameter $\boldsymbol{\theta}^*$ and the GDS estimate $\hat{\boldsymbol{\theta}}$. Let the *error vector* be defined as $\boldsymbol{\delta} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$. First we have the definitions for *error cone* and *error spherical cap*.

Definition 12 (error cone/spherical cap) The *error cone* of $\boldsymbol{\theta}^*$ for norm $\|\cdot\|$ is defined as

$$\mathcal{T} = \operatorname{cone}\left\{\mathbf{u} \in \mathbb{R}^p \;\middle|\; \|\boldsymbol{\theta}^* + \mathbf{u}\| \leq \|\boldsymbol{\theta}^*\|\right\} \tag{3.20}$$

The *error spherical cap* is the intersection of error cone and the unit sphere, i.e.

$$\mathcal{C} = \mathcal{T} \cap \mathbb{S}^{p-1} = \operatorname{cone}\left\{\mathbf{u} \in \mathbb{R}^p \;\middle|\; \|\boldsymbol{\theta}^* + \mathbf{u}\| \leq \|\boldsymbol{\theta}^*\|\right\} \cap \mathbb{S}^{p-1} \tag{3.21}$$

Note that error cone contains a restricted set of directions and does not in general span the entire space of $\mathbb{R}^p$. One relevant notation to error cone and spherical cap is the *restricted norm compatibility*, which is the largest quotient of $\|\cdot\|$ and $\|\cdot\|_2$ over all the directions in error cone.

Definition 13 (restricted norm compatibility) The *restricted norm compatibility* for a norm $\|\cdot\|$ is defined as

$$\Psi \triangleq \sup_{\mathbf{v}\in\mathcal{T}} \frac{\|\mathbf{v}\|}{\|\mathbf{v}\|_2} = \sup_{\mathbf{v}\in\mathcal{C}} \|\mathbf{v}\| \tag{3.22}$$

In the rest of the thesis, the notions introduced in Definition 12 and 13 will be frequently used. For specific norms, we may add subscripts or superscripts to $\mathcal{T}$, $\mathcal{C}$ and Ψ for clarity. The deterministic L_2-error of $\hat{\boldsymbol{\theta}}$ depends on the following two conditions.

Definition 14 (restricted eigenvalue (RE) condition) The design matrix $\mathbf{X}$ satisfies the *restricted eigenvalue* (RE) condition for a set $\mathcal{C} \subseteq \mathbb{S}^{p-1}$ with parameter $\alpha > 0$, if

$$\inf_{\mathbf{v}\in\mathcal{C}} \frac{1}{n} \|\mathbf{X}\mathbf{v}\|_2^2 \geq \alpha \tag{3.23}$$

Definition 15 (admissible tuning parameter) The tuning parameter λ_n of (3.2) is said to be *admissible* if it satisfies that

$$\left\| \frac{1}{n}\mathbf{X}^T \left(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}^*\right) \right\|_* = \left\| \frac{1}{n}\mathbf{X}^T \boldsymbol{\epsilon} \right\|_* \leq \lambda_n \tag{3.24}$$

An admissible λ_n essentially guarantees that the true parameter $\boldsymbol{\theta}^*$ is inside the feasible set of GDS, which further explains why $\mathcal{T}$ is called error cone. Since $\boldsymbol{\theta}^*$ is feasible, the norm $\|\hat{\boldsymbol{\theta}}\|$ must be less than or equal to that of $\boldsymbol{\theta}^*$, which translates to $\|\boldsymbol{\delta} + \boldsymbol{\theta}^*\| \leq \|\boldsymbol{\theta}^*\|$. Thus error cone encompasses all directions that the error $\boldsymbol{\delta}$ could point towards. The

RE condition ensures sufficient curvature of $\|\mathbf{X}\mathbf{v}\|_2^2$ along the error cone, which help us confine the magnitude of $\boldsymbol{\delta}$ if $\|\mathbf{X}\boldsymbol{\delta}\|_2^2$ is known to be small. The next lemma bounds the deterministic L_2-error of GDS.

Lemma 5 *Suppose that the RE condition* (3.23) *is satisfied by* $\mathbf{X}$ *for the error spherical cap defined in* (3.21)*, and the parameter* λ_n *is chosen to be admissible. Then GDS* $\hat{\boldsymbol{\theta}}$ *given by* (3.2) *satisfies*

$$\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_2 \leq 2\Psi \cdot \frac{\lambda_n}{\alpha} \tag{3.25}$$

Proof: Under the admissibility of λ_n and the optimality of $\hat{\boldsymbol{\theta}}$ for (3.2), we have

$$\left\|\frac{1}{n}\mathbf{X}^T\left(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}^*\right)\right\|_* \leq \lambda_n \ , \qquad \left\|\frac{1}{n}\mathbf{X}^T\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}}\right)\right\|_* \leq \lambda_n \ ,$$
$$\|\hat{\boldsymbol{\theta}}\| = \|\boldsymbol{\delta} + \boldsymbol{\theta}^*\| \leq \|\boldsymbol{\theta}^*\| \qquad \Longrightarrow \qquad \boldsymbol{\delta} \in \mathcal{T}$$

Adding the first two inequalities and applying triangular inequality, we obtain

$$\left\|\frac{1}{n}\mathbf{X}^T\mathbf{X}\boldsymbol{\delta}\right\|_* \leq \left\|\frac{1}{n}\mathbf{X}^T\left(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}^*\right)\right\|_* + \left\|\frac{1}{n}\mathbf{X}^T\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}}\right)\right\|_* \leq 2\lambda_n$$
$$\Longrightarrow \qquad \frac{1}{n}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 = \left\langle \boldsymbol{\delta}, \frac{1}{n}\mathbf{X}^T\mathbf{X}\boldsymbol{\delta}\right\rangle \leq \|\boldsymbol{\delta}\| \left\|\frac{1}{n}\mathbf{X}^T\mathbf{X}\boldsymbol{\delta}\right\|_* \leq 2\lambda_n\|\boldsymbol{\delta}\| \ ,$$

which follows from Hölder's inequality. By $\boldsymbol{\delta} \in \mathcal{T}$ and RE condition for $\mathcal{C}$, we have

$$\frac{1}{n}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 \geq \|\boldsymbol{\delta}\|_2^2 \cdot \inf_{\mathbf{v}\in\mathcal{C}} \frac{1}{n}\|\mathbf{X}\mathbf{v}\|_2^2 = \alpha\|\boldsymbol{\delta}\|_2^2$$

Combining the results above, we have

$$\alpha\|\boldsymbol{\delta}\|_2^2 \leq 2\|\boldsymbol{\delta}\|\lambda_n \qquad \Longrightarrow \qquad \|\boldsymbol{\delta}\|_2 \leq 2 \cdot \frac{\|\boldsymbol{\delta}\|}{\|\boldsymbol{\delta}\|_2} \cdot \frac{\lambda_n}{\alpha} \leq 2\Psi \cdot \frac{\lambda_n}{\alpha} \ ,$$

which completes the proof. ∎

3.3.2 Error Bound with Random Design and Noise

The deterministic bound (3.25) gives a clear characterization of L_2-error of GDS. When we consider the randomness of the design $\mathbf{X}$ and noise $\boldsymbol{\epsilon}$, there are two terms remain to be resolved in the deterministic bound. First we need to find the parameter α for RE condition. Under the sub-Gaussianity of $\mathbf{X}$, we obtain the following result for RE condition.

Theorem 5 *Let the rows of $\mathbf{X} \in \mathbb{R}^{n \times p}$ be i.i.d. copies of an isotropic sub-Gaussian random vector $\mathbf{x} \in \mathbb{R}^p$ with $\|\|\mathbf{x}\|\|_{\psi_2} \leq \kappa$. With probability at least $1 - \exp\left(-C_1 w^2(\mathcal{C})\right)$, we have*

$$\inf_{\mathbf{v} \in \mathcal{C}} \frac{1}{n} \|\mathbf{X}\mathbf{v}\|_2^2 \geq 1 - C_0 \kappa^2 \cdot \frac{w(\mathcal{C})}{\sqrt{n}} \ , \tag{3.26}$$

where C_0 and C_1 are absolute constants.

Based on Theorem 5, we immediately have the corollary below.

Corollary 2 *Under the setting of Theorem 5, if sample size $n \geq 4C_0^2 \kappa^4 w^2(\mathcal{C})$, then with probability at least $1 - \exp\left(-C_1 w^2(\mathcal{C})\right)$, the RE condition holds for $\mathcal{C}$ with parameter $\alpha = \frac{1}{2}$.*

Second, we have to choose the smallest admissible λ_n so that the upper bound is as tight as possible, which requires an estimation of the random quantity $\|\frac{1}{n}\mathbf{X}^T\boldsymbol{\epsilon}\|_*$.

Theorem 6 *Let the rows of $\mathbf{X} \in \mathbb{R}^{n \times p}$ be i.i.d. copies of an isotropic sub-Gaussian random vector $\mathbf{x} \in \mathbb{R}^p$ with $\|\|\mathbf{x}\|\|_{\psi_2} \leq \kappa$, and the entries of $\boldsymbol{\epsilon} \in \mathbb{R}^n$ be i.i.d. copies of a sub-Gaussian random variable ϵ with $\|\|\epsilon\|\|_{\psi_2} \leq \tau$. The following inequality holds with probability at least $1 - \exp(-c_1 n) - c_2 \exp\left(-\frac{w^2(\Omega)}{c_3^2 \rho^2}\right)$,*

$$\left\|\mathbf{X}^T \boldsymbol{\epsilon}\right\|_* \leq c_0 \kappa \tau \cdot \sqrt{n} w\left(\Omega\right) \ , \tag{3.27}$$

in which Ω is the unit ball of $\|\cdot\|$, $\rho = \sup_{\mathbf{v}\in\Omega} \|\mathbf{v}\|_2$, and c_0, c_1, c_2 and c_3 are all absolute constants.

Theorem 6 directly yields an "safe" choice of λ_n, which is admissible with high probability.

Corollary 3 *Under the setting of Theorem 6, $\lambda_n = \frac{c_0\kappa\tau\cdot w(\Omega)}{\sqrt{n}}$ is admissible with probability at least $1 - \exp(-c_1 n) - c_2 \exp\left(-\frac{w^2(\Omega)}{c_3^2\rho^2}\right)$.*

Combining Corollary 2 and 3, the L_2-error bound is given in the theorem below.

Theorem 7 *Let the rows of $\mathbf{X} \in \mathbb{R}^{n\times p}$ be i.i.d. copies of an isotropic sub-Gaussian random vector $\mathbf{x} \in \mathbb{R}^p$ with $|||\mathbf{x}|||_{\psi_2} \leq \kappa$, and the entries of $\boldsymbol{\epsilon} \in \mathbb{R}^n$ be i.i.d. copies of a sub-Gaussian random variable ϵ with $|||\epsilon|||_{\psi_2} \leq \tau$. if sample size $n \geq 4C_0^2\kappa^4 w^2(\mathcal{C})$, with probability at least $1 - \exp(-c_1 n) - c_2 \exp\left(-\frac{w^2(\Omega)}{c_3^2\rho^2}\right) - \exp\left(-C_1 w^2\left(\mathcal{C}\right)\right)$, the L_2-error of GDS satisfies*

$$\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_2 \leq c\kappa\tau \cdot \frac{\Psi \cdot w\left(\Omega\right)}{\sqrt{n}} \tag{3.28}$$

Proof: the error bound is a direct result of Lemma 5 and Corollary 2 and 3. ∎

Remark: In the above theorem, other than some constants and sub-Gaussian parameters, the error bound for $\hat{\boldsymbol{\theta}}$ essentially depends on there quantities regarding the structure of $\boldsymbol{\theta}^*$, the Gaussian width $w(\Omega)$ of the unit norm ball, the Gaussian width $w(\mathcal{C})$ of the error spherical cap, and the restricted norm compatibility Ψ. We call these *geometric measures*, since they rely on the geometry of $\boldsymbol{\theta}^*$ and the norm $\|\cdot\|$.

3.4 Experimental Results

On optimization side, our ADMM framework is concentrated on its generality, and its efficiency has been shown in [176] for the special case of L_1 norm. Hence we focus on the efficiency of different proximal operators related to k-support norm. On statistical side, we concentrate on the behavior and performance of GDS with k-support norm.

3.4.1 Efficiency of Proximal Operator

We tested four proximal operators related to k-support norm, which are our normal $\mathbf{prox}_{\mathbb{I}_{\mathcal{C}_\lambda}}(\cdot)$ and its accelerated version, $\mathbf{prox}_{\frac{1}{2\beta}(\|\cdot\|_k^{sp})^2}(\cdot)$ in [7], and $\mathbf{prox}_{\frac{\lambda}{2}\|\cdot\|_\Theta^2}(\cdot)$ in [120]. The dimension p of vector in experiment varied from 1000 to 10000, and the ratio $p/k = \{200, 100, 50, 20\}$. As illustrated in Figure 3.1, in general, the speedup of accelerated $\mathbf{prox}_{\mathbb{I}_{\mathcal{C}_\lambda}}(\cdot)$ is considerable when compared with the normal $\mathbf{prox}_{\mathbb{I}_{\mathcal{C}_\lambda}}(\cdot)$ and $\mathbf{prox}_{\frac{1}{2\beta}(\|\cdot\|_k^{sp})^2}(\cdot)$. Empirically it is also slightly better than the $\mathbf{prox}_{\frac{\lambda}{2}\|\cdot\|_\Theta^2}(\cdot)$.

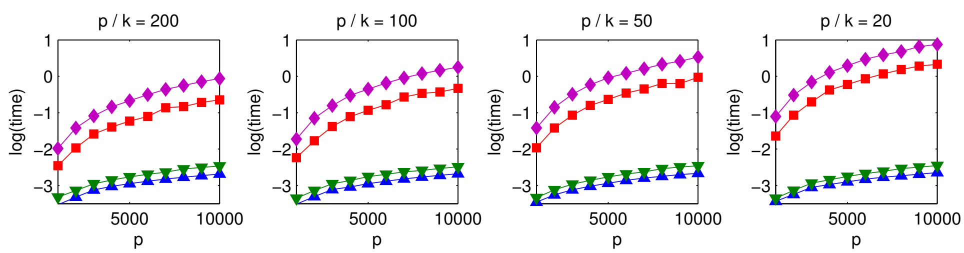
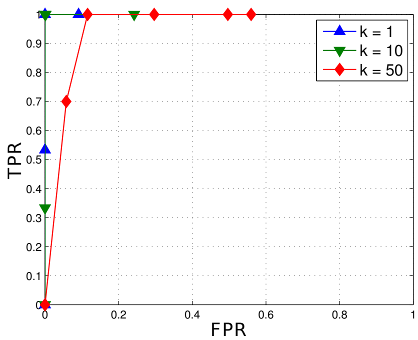


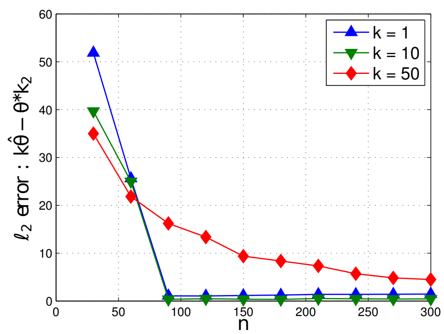
Figure 3.1: Efficiency of proximal operators for k-support norm. Diamond: normal $\mathbf{prox}_{\mathbb{I}_{\mathcal{C}_\lambda}}(\cdot)$, Square: $\mathbf{prox}_{\frac{1}{2\beta}(\|\cdot\|_k^{sp})^2}(\cdot)$, Downward-pointing triangle: $\mathbf{prox}_{\frac{\lambda}{2}\|\cdot\|_\Theta^2}(\cdot)$, Upward-pointing triangle: accelerated $\mathbf{prox}_{\mathbb{I}_{\mathcal{C}_\lambda}}(\cdot)$. For each $(p,\, k)$, 200 vectors are randomly generated for testing.

3.4.2 Statistical Recovery

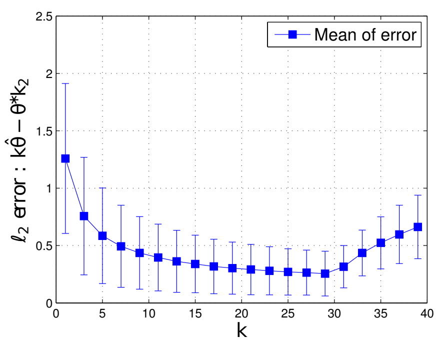
Data generation: We fix $p = 600$, and $\boldsymbol{\theta}^* = [\underbrace{10, \ldots, 10}_{10}, \underbrace{10, \ldots, 10}_{10}, \underbrace{10, \ldots, 10}_{10}, \underbrace{0, \ldots, 0}_{570}]^T$ throughout the experiment, in which nonzero entries are divided equally into three groups. The design matrix $\mathbf{X}$ are generated from a normal distribution such that the entries in the same group have the same mean sampled from $\mathcal{N}(0, 1)$. $\mathbf{X}$ is normalized afterwards. The response vector $\mathbf{y}$ is given by $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}^* + 0.01 \times \mathcal{N}(0, 1)$. The number of samples n is specified later.



(a) ROC curves



(b) L_2-error vs. n



(c) L_2-error vs. k

Figure 3.2: Statistical recovery of GDS with k-support norm. (a) The true positive rate reaches 1 quite early for $k = 1, 10$. When $k = 50$, the ROC gets worse due to the strong smoothing effect introduced by large k. (b) For each k, the L_2-error is large when the sample is inadequate. As n increases, the error decreases dramatically for $k = 1, 10$ and becomes stable afterwards, while the decrease is not that significant for $k = 50$ and the error remains relatively large. (c) Both mean and standard deviation of L_2-error are decreasing as k increases until it exceeds the number of nonzero entries in $\boldsymbol{\theta}^*$, and then the error goes up for larger k, which matches our analysis quite well. The result also shows that the k-support-norm GDS with suitable k outperforms the L_1 DS when correlated variables present in data (Note that $k = 1$ corresponds to standard DS).

ROC curves with different k: We fix $n = 400$ to obtain the ROC plot for $k = \{1, 10, 50\}$ as shown in Figure 3.2(a). λ_n ranged from 10^{-2} to 10^3.

L_2-error vs. n: We investigate how the L_2-error $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2$ of Dantzig selector changes as the number of samples increases, where $k = \{1, 10, 50\}$ and $n = \{30, 60, 90, \ldots, 300\}$. The plot is shown in Figure 3.2(b).

L_2-error vs. k: We also look at the L_2-error with different k. We again fix $n = 400$ and vary k from 1 to 39. For each k, we repeat the experiment 100 times, and obtained the mean and standard deviation plot in Figure 3.2(c).

Appendix

Appendix 3.A Proof of Proximal Operator for k-Support Norm

3.A.1 Proof of Theorem 3

Proof: Let $\mathbf{w}^* = \mathbf{prox}_{\mathbb{I}_{\Omega^*_\lambda}}(\mathbf{x}) = \operatorname{argmin}_{\mathbf{w} \in \Omega^*_\lambda} \frac{1}{2}\|\mathbf{x} - \mathbf{w}\|_2^2$. For simplicity, we drop the constant $\frac{1}{2}$ in later discussion. Given a vector $\mathbf{x}$, we use the notation $\mathbf{x}_{i:j}$ to denote its subvector $[\mathbf{x}_i, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_j]^T$. We consider the following two cases.

Case 1: if $\|\mathbf{x}\|_{k*}^{sp} \leq \lambda$, it is trivial that $\mathbf{w}^* = \mathbf{x}$, which is also the global minimizer of $\|\mathbf{x} - \mathbf{w}\|_2^2$ without the constraint $\mathbf{x} \in \Omega^*_\lambda$.

Case 2: if $\|\mathbf{x}\|_{k*}^{sp} > \lambda$, first we start by noting that given $\mathbf{x}$ and $\mathbf{w}$, the following inequality holds

$$\begin{aligned}\|\mathbf{x} - \mathbf{w}\|_2^2 &= \|\mathbf{x}\|_2^2 - 2\langle \mathbf{x}, \mathbf{w} \rangle + \|\mathbf{w}\|_2^2 \\ &\geq \|\mathbf{x}\|_2^2 - 2\langle |\mathbf{x}|^{\downarrow}, |\mathbf{w}|^{\downarrow} \rangle + \|\mathbf{w}\|_2^2 \ ,\end{aligned}$$

which implies that $\mathbf{w}^*$ should achieve this lower bound by conforming with the signs and orders of elements in $\mathbf{x}$. Without loss of generality, we are simply focused on the case where $\mathbf{x} = |\mathbf{x}|^{\downarrow}$.

For $\mathbf{w}^*$ to be the optimal, $\mathbf{w}^*_{k:p}$ should be chosen such that $\mathbf{w}^*_{k:r} = [\mathbf{w}^*_k, \mathbf{w}^*_k, \ldots, \mathbf{w}^*_k]^T$ and $\mathbf{w}^*_{r+1:p} = \mathbf{x}^*_{r+1:p}$, where r satisfies

$$\mathbf{x}_r > \mathbf{w}^*_k \geq \mathbf{x}_{r+1} \ ,$$

otherwise either the decreasing order of $\mathbf{w}^*$ will be violated or the $\|\mathbf{x}_{k:p} - \mathbf{w}_{k:p}\|_2$ is not minimized. As for $\mathbf{w}^*_{1:k-1}$, we similarly assume $\mathbf{w}^*_{s+1:k-1} = [\mathbf{w}^*_k, \mathbf{w}^*_k, \ldots, \mathbf{w}^*_k]^T$ for some $0 \leq s \leq k-1$, then $\mathbf{w}^*_{1:s}$ should be chosen to minimize $\|\mathbf{x}_{1:s} - \mathbf{w}_{1:s}\|_2$ such that

$$\|\mathbf{w}_{1:s}\|_2^2 = \|\mathbf{w}^*_{1:k}\|_2^2 - \|\mathbf{w}^*_{s+1:k}\|_2^2 \leq \lambda^2 - (k-s)(\mathbf{w}^*_k)^2.$$

By Cauchy-Schwarz inequality, we have

$$\|\mathbf{x}_{1:s} - \mathbf{w}_{1:s}\|_2^2 \geq \|\mathbf{x}_{1:s}\|_2^2 - 2\|\mathbf{x}_{1:s}\|_2\|\mathbf{w}_{1:s}\|_2 + \|\mathbf{w}_{1:s}\|_2^2 \ ,$$

where the equality holds when $\mathbf{w}^*_{1:s}$ follows the form of $\mathbf{w}^*_{1:s} = \frac{1}{1+\beta_{sr}}\mathbf{x}_{1:s}$, and $\beta_{sr} \geq 0$ satisfies the constraint $\frac{B_s}{(1+\beta_{sr})^2} = \lambda^2 - (k-s)(\mathbf{w}_k)^2$.

So far we have figured out the structure of $\mathbf{w}^* = [\mathbf{w}^*_{1:s}, \mathbf{w}^*_{s+1:r}, \mathbf{w}^*_{r+1:p}]^T$, in which the three subvectors, compared with $\mathbf{x}$, are shrunk by a common factor $1+\beta_{sr}$, constant $\mathbf{w}^*_k$, or unchanged. Next we need to determine the value of β_{sr} and $\mathbf{w}^*_k$. By optimality,

$\|\mathbf{x}-\mathbf{w}\|_2^2 = \|\mathbf{x}_{1:r}-\mathbf{w}_{1:r}\|_2^2$ must be minimized at $\mathbf{w}^*$, so we have the following problem,

$$\begin{aligned}\min_{\beta,\mathbf{w}_k} \ \|\mathbf{x}_{1:r}-\mathbf{w}_{1:r}\|_2^2 &= \|\mathbf{x}_{1:s}-\mathbf{w}_{1:s}\|_2^2 + \|\mathbf{x}_{s+1:r}-\mathbf{w}_{s+1:r}\|_2^2 \\ &= \left(\frac{\beta}{1+\beta}\right)^2 B_s + \sum_{i=s+1}^{r} (\mathbf{x}_i - \mathbf{w}_k)^2 \end{aligned} \tag{3.29}$$

$$\text{s.t.} \qquad (\|\mathbf{w}\|_{k*}^{sp})^2 = \frac{B_s}{(1+\beta)^2} + (k-s)(\mathbf{w}_k)^2 = \lambda^2 \tag{3.30}$$

Replacing $\mathbf{w}_k$ in (3.29) with $\mathbf{w}_k = \sqrt{\frac{\lambda^2 - \frac{B_s}{(1+\beta)^2}}{k-s}}$ obtained from (3.30), we express $\|\mathbf{x}_{1:r} - \mathbf{w}_{1:r}\|_2^2$ as a function of β,

$$\Phi_{sr}(\beta) = \left(\frac{\beta}{1+\beta}\right)^2 B_s + \sum_{i=s+1}^{r}\left(\mathbf{x}_i - \sqrt{\frac{\lambda^2 - \frac{B_s}{(1+\beta)^2}}{k-s}}\right)^2 \tag{3.31}$$

Set derivative of $\Phi_{sr}(\beta)$ to be zero, we have

$$\frac{d}{d\beta}\Phi_{sr}(\beta) = \frac{d}{d\beta}\left[(\frac{\beta}{1+\beta})^2 B_s + \sum_{i=s+1}^{r}\big(\mathbf{x}_i - \sqrt{\frac{\lambda^2 - \frac{B_s}{(1+\beta)^2}}{k-s}}\big)^2\right] \tag{3.32}$$

$$= \frac{2\beta}{(1+\beta)^3}B_s - \frac{2A_{sr}B_s}{(1+\beta)^3(k-s)\sqrt{\frac{\lambda^2-\frac{B_s}{(1+\beta)^2}}{k-s}}} + \frac{2(r-s)B_s}{(k-s)(1+\beta)^3} \tag{3.33}$$

$$= \frac{2B_s}{(k-s)(1+\beta)^3}\left[(k-s)\beta - \frac{A_{sr}}{\sqrt{\frac{\lambda^2-\frac{B_s}{(1+\beta)^2}}{k-s}}} + (r-s)\right] = 0 \tag{3.34}$$

If $s > 0$, then $B_s > 0$ and (3.34) is equivalent to (3.14). And we can see that the quantity inside the bracket of (3.34) is monotonically increasing when $\beta \geq \max\left\{0, \frac{\sqrt{B_s}-\lambda}{\lambda}\right\}$, thus ensuring the nonnegative root β_{sr} is unique if it exists. If the nonnegative root exists, the expression for $\mathbf{w}^*_{s+1:r}$ can be obtained from (3.34), whose entries are all equal to $\mathbf{w}^*_k$.

If $s > 0$ and a nonnegative root of (3.34) is nonexistent, the derivative is always positive when $\beta \geq 0$, which means that $\Phi_{sr}(\beta)$ is increasing. Hence the minimizer of $\Phi_{sr}(\beta)$ is $\beta_{sr} = 0$. If $s = 0$, we actually do not care about the value of β_{sr} because the problem defined by (3.29) and (3.30) is independent of β, and we set it to be 0 for simplicity. According to (3.30), both cases of $\beta_{sr} = 0$ lead to the same expression for $\mathbf{w}^*_{s+1:r}$ in (3.16).

As we do not know beforehand which s and r to choose, we need to search for s^* and r^* that give the smallest $\||\mathbf{x}|^{\downarrow} - |\mathbf{w}|^{\downarrow}\|_2$, and also need to check whether the $\mathbf{w}^*$ obtained by (3.16) is in decreasing order, which are the conditions (3.17) and (3.18) presented in Theorem 3. ∎

3.A.2 Proof of Theorem 4

To prove Theorem 4, we first need the following lemma derived from the proof of Theorem 3.

Lemma 6 *When* $\beta \geq \max\left\{0, \frac{\sqrt{B_s}-\lambda}{\lambda}\right\}$, $\Phi_{sr}(\beta)$ *defined in* (3.31) *is decreasing when* $\beta < \beta_{sr}$, *and increasing when* $\beta > \beta_{sr}$. *Equivalently,* $\Phi_{sr}(\beta) = \|\mathbf{x}_{1:r} - \mathbf{w}_{1:r}\|_2^2$, *when treated as function of* $\mathbf{w}_k$, *is decreasing when* $\mathbf{w}_k < \mathbf{w}_k^*$ *and increasing when* $\mathbf{w}_k > \mathbf{w}_k^*$.

Proof: The first part simply follows the monotonicity of $\frac{d}{d\beta}\Phi_{sr}(\beta)$ mentioned in the proof of Theorem 3, which implies that $\frac{d}{d\beta}\Phi_{sr}(\beta)$ is negative when $\beta < \beta_{sr}$, and positive when $\beta > \beta_{sr}$. The constraint (3.30) implies that $\mathbf{w}_k$ increases as β increases. So $\|\mathbf{x}_{1:r} - \mathbf{w}_{1:r}\|_2^2$, as a function of $\mathbf{w}_k$, has the same monotonicity w.r.t. $\mathbf{w}_k$. ∎

Proof of Theorem 4: It suffices to just focus on the case where $\mathbf{x} = |\mathbf{x}|^{\downarrow}$. First we show by contradiction that for a given $\tilde{s}$, the $\tilde{r}$ that satisfies (3.18) can be at most one.

Suppose there are two indices, say r_1 and r_2, which satisfy that condition with the same $\tilde{s}$. Without loss of generality, let $r_1 < r_2$, we know that their corresponding $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$ should minimize $\|\mathbf{x}_{1:r_1} - \mathbf{w}_{1:r_1}\|_2^2$ and $\|\mathbf{x}_{1:r_2} - \mathbf{w}_{1:r_2}\|_2^2$, respectively. As $r_1 < r_2$, then $\mathbf{w}_k^{(1)} \geq \mathbf{x}_{r_2} > \mathbf{w}_k^{(2)}$ according to (3.18). Construct

$$\mathbf{w}' = \left[\underbrace{\frac{\mathbf{x}_1}{1+\beta'}, \ldots, \frac{\mathbf{x}_{\tilde{s}}}{1+\beta'}}_{\tilde{s}}, \; \underbrace{\mathbf{x}_{r_2}, \ldots, \mathbf{x}_{r_2}}_{r_2-\tilde{s}}, \; \mathbf{x}_{r_2+1}, \ldots, \mathbf{x}_p \right]^T ,$$

where β' is chosen to satisfy the constraint (3.30) with $\mathbf{w}'_k = \mathbf{x}_{r_2}$, and $\|\mathbf{x}_{1:r_2} - \mathbf{w}^{(2)}_{1:r_2}\|_2^2$ can be decomposed as

$$\begin{aligned} \left\|\mathbf{x}_{1:r_2} - \mathbf{w}^{(2)}_{1:r_2}\right\|_2^2 &= \left\|\mathbf{x}_{1:r_1} - \mathbf{w}^{(2)}_{1:r_1}\right\|_2^2 + \left\|\mathbf{x}_{r_1+1:r_2} - \mathbf{w}^{(2)}_{r_1+1:r_2}\right\|_2^2 \\ &> \left\|\mathbf{x}_{1:r_1} - \mathbf{w}'_{1:r_1}\right\|_2^2 + \left\|\mathbf{x}_{r_1+1:r_2} - \mathbf{w}'_{r_1+1:r_2}\right\|_2^2 \\ &= \left\|\mathbf{x}_{1:r_2} - \mathbf{w}'_{1:r_2}\right\|_2^2 \end{aligned}$$

which contradicts that $\mathbf{w}^{(2)}_{1:r_2}$ minimizes $\|\mathbf{x}_{1:r_2} - \mathbf{w}_{1:r_2}\|_2^2$. Note that $\|\mathbf{x}_{1:r_1} - \mathbf{w}^{(2)}_{1:r_1}\|_2^2 > \|\mathbf{x}_{1:r_1} - \mathbf{w}'_{1:r_1}\|_2^2$ simply follows Lemma 6 as $\mathbf{w}_k^{(1)} \geq \mathbf{x}_{r_2} = \mathbf{w}'_k > \mathbf{w}_k^{(2)}$, and $\|\mathbf{x}_{r_1+1:r_2} - \mathbf{w}^{(2)}_{r_1+1:r_2}\|_2^2 > \|\mathbf{x}_{r_1+1:r_2} - \mathbf{w}'_{r_1+1:r_2}\|_2^2$ is due to the fact that $\mathbf{x}_{r_1+1} \geq \ldots \geq \mathbf{x}_{r_2} = \mathbf{w}'_k > \mathbf{w}_k^{(2)}$.

Next we show by contradiction that if $\tilde{r}$ exists for given $\tilde{s}$, then any $r < \tilde{r}$ violates the first part of (3.18), and any $r > \tilde{r}$ violates the second part.

Let $\tilde{\mathbf{w}}$ denote the minimizer of $\|\mathbf{x}_{1:\tilde{r}} - \mathbf{w}_{1:\tilde{r}}\|_2^2$. Suppose $r < \tilde{r}$ and the first part of (3.18) is not violated, then its second part must be violated due to the uniqueness of $\tilde{r}$. Then we can construct new

$$\mathbf{w}' = \left[\underbrace{\frac{\mathbf{x}_1}{1+\beta'}, \ldots, \frac{\mathbf{x}_{\tilde{s}}}{1+\beta'}}_{\tilde{s}}, \; \underbrace{\mathbf{x}_{\tilde{r}}, \ldots, \mathbf{x}_{\tilde{r}}}_{\tilde{r}-\tilde{s}}, \; \mathbf{x}_{\tilde{r}+1}, \ldots, \mathbf{x}_p \right]^T ,$$

where β' is again chosen to satisfy the constraint (3.30) with $\mathbf{w}'_k = \mathbf{x}_{\tilde{r}}$. This by the same argument for proving the uniqueness of $\tilde{r}$ make the following inequality hold,

$$\begin{aligned}
\|\mathbf{x}_{1:\tilde{r}} - \tilde{\mathbf{w}}_{1:\tilde{r}}\|_2^2 &= \|\mathbf{x}_{1:r} - \tilde{\mathbf{w}}_{1:r}\|_2^2 + \|\mathbf{x}_{r+1:\tilde{r}} - \tilde{\mathbf{w}}_{r+1:\tilde{r}}\|_2^2 \\
&> \|\mathbf{x}_{1:r} - \mathbf{w}'_{1:r}\|_2^2 + \|\mathbf{x}_{r+1:\tilde{r}} - \mathbf{w}'_{r+1:\tilde{r}}\|_2^2 \\
&= \|\mathbf{x}_{1:\tilde{r}} - \mathbf{w}'_{1:\tilde{r}}\|_2^2 \ .
\end{aligned}$$

This contradicts that $\tilde{\mathbf{w}}$ is the minimizer of $\|\mathbf{x}_{1:\tilde{r}} - \mathbf{w}_{1:\tilde{r}}\|_2^2$. Similar argument applies to the case when $r > \tilde{r}$. Let β'' satisfy (3.30) together with $\mathbf{w}''_k = \mathbf{x}_{r+1}$, and we construct

$$\mathbf{w}'' = \left[\underbrace{\frac{\mathbf{x}_1}{1+\beta''}, \ldots, \frac{\mathbf{x}_s}{1+\beta''}}_{\tilde{s}},\ \underbrace{\mathbf{x}_{r+1}, \ldots, \mathbf{x}_{r+1}}_{r-\tilde{s}},\ \mathbf{x}_{r+1}, \ldots, \mathbf{x}_p\right]^T ,$$

which gives smaller $\|\mathbf{x}_{1:r} - \mathbf{w}_{1:r}\|_2^2$ than any $\mathbf{w}$ with $\mathbf{w}_k < \mathbf{x}_{r+1}$. Therefore it is impossible for $r > \tilde{r}$ to violate the first inequality.

Finally we show the assertion (3.19) for s^*. We note that given $\tilde{s}$, finding solution to the proximal operator can be viewed as minimization of (3.29) under the constraint $\|\mathbf{w}_{1:k}\|_2 \leq \lambda$ and $\mathbf{w}_k = \mathbf{w}_{k-1} = \ldots = \mathbf{w}_{\tilde{s}+1}$. So for $s < \tilde{s}$, the minimization problem is equivalent to the one for $\tilde{s}$ under additional constraint $\mathbf{w}_{\tilde{s}+1} = \mathbf{w}_{\tilde{s}} = \ldots = \mathbf{w}_{s+1}$. If the $\tilde{r}$ does not exist, for $s < \tilde{s}$, $\tilde{r}$ is nonexistent either, thus $s^* > \tilde{s}$. If the $\tilde{r}$ exists and (3.17) is satisfied, then $s^* \geq \tilde{s}$ because $s < \tilde{s}$ considers a more restricted problem and is unable to obtain a smaller $\|\mathbf{x} - \mathbf{w}\|_2$.

For the situation in which $\tilde{r}$ exists for $\tilde{s}$ but the associated $\tilde{\mathbf{w}}_k$ violates (3.17), we show by contradiction that for any $s > \tilde{s}$, (3.17) is also violated. Assume that $\mathbf{w}'$ (different from the previously used) satisfies both (3.17) and (3.18) for $s' = \tilde{s} + 1$ and the corresponding r'. It is not difficult to see that $\mathbf{w}'_k < \tilde{\mathbf{w}}_k$ and $r' \geq \tilde{r}$, otherwise

$\|\mathbf{w}'_{1:k}\|_2 > \lambda$. By the violation we have shown for r, the minimizer of (3.29) for $(s', \tilde{r})$, denoted by $\mathbf{w}''$, satisfies $\mathbf{w}''_k \leq \mathbf{w}'_k$ (Note that $\mathbf{w}'$ is the minimizer of (3.29) for (s', r') and $r' \geq \tilde{r}$). Combined with $\mathbf{w}'_k < \tilde{\mathbf{w}}_k$, this indicates by Lemma 6 that $\Phi_{s'\tilde{r}}(\cdot)$ is increasing on the interval $[\mathbf{w}''_k, \tilde{\mathbf{w}}_k]$. Then we consider two sequential modifications on $\tilde{\mathbf{w}}$,

1. Replacing the $\tilde{\mathbf{w}}_{1:s'}$ in $\tilde{\mathbf{w}}$ with $\frac{\|\tilde{\mathbf{w}}_{1:s'}\|_2}{\|\mathbf{x}_{1:s'}\|_2} \cdot \mathbf{x}_{1:s'}$,
2. Decreasing $\tilde{\mathbf{w}}_{s'+1:\tilde{r}}$ by certain amount and amplifying the new $\tilde{\mathbf{w}}_{1:s'}$ by some factor, such that (3.30) still holds for s' and $\tilde{\mathbf{w}}_{s'+1} = \tilde{\mathbf{w}}_{s'}$.

Note that the two modifications both decrease $\|\mathbf{x}_{1:\tilde{r}} - \tilde{\mathbf{w}}_{1:\tilde{r}}\|_2$. Decrease in Modification 1 is the result of Cauchy-Schwarz Inequality, and decrease in Modification 2 is due to the monotonicity of $\Phi_{s'\tilde{r}}(\cdot)$ we mentioned upfront. The modified $\tilde{\mathbf{w}}$ satisfies $\tilde{\mathbf{w}}_{\tilde{s}+1} = \tilde{\mathbf{w}}_{\tilde{s}+2} = \ldots = \tilde{\mathbf{w}}_k$, thus contradicting that the old $\tilde{\mathbf{w}}$ is the minimizer of (3.29) for $(\tilde{s}, \tilde{r})$. Hence, by induction, we conclude that for any $s' > \tilde{s}$, its solution also violates (3.17).

Assembling the conclusions above, we complete the proof of (3.19) for s^*. ■

Appendix 3.B Proof of Statistical Guarantees

3.B.1 Proof of Theorem 5

Proof: Let (Ω, μ) be the probability space that $\mathbf{x}$ is defined on, and construct

$$\mathcal{H} = \{h = \langle \cdot, \mathbf{v} \rangle \mid \mathbf{v} \in \mathcal{C}\} .$$

$|||\mathbf{x}|||_{\psi_2} \leq \kappa$ immediately implies that $\sup_{h \in \mathcal{H}} |||h|||_{\psi_2} \leq \kappa$. As $\mathbf{x}$ is isotropic, i.e., $\mathbb{E}[\langle \mathbf{x}, \mathbf{v} \rangle^2] = 1$ for any $\mathbf{v} \in \mathcal{C} \subseteq \mathbb{S}^{p-1}$, thus we have $\mathcal{H} \subseteq \mathbb{S}_{L_2(\mu)}$ and $\mathbb{E}[h^2] = 1$ for any $h \in \mathcal{H}$. Given $h_1 = \langle \cdot, \mathbf{v}_1 \rangle, h_2 = \langle \cdot, \mathbf{v}_2 \rangle \in \mathcal{H}$, where $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{C}$, the metric induced by ψ_2 norm satisfies

$$|||h_1 - h_2|||_{\psi_2} = |||\langle \mathbf{x}, \mathbf{v}_1 - \mathbf{v}_2 \rangle|||_{\psi_2} \le \kappa \|\mathbf{v}_1 - \mathbf{v}_2\|_2 \ .$$

Using (2.39) - (2.41) and Lemma 4, we have

$$\gamma_2(\mathcal{H}, |||\cdot|||_{\psi_2}) \ \le \ \kappa \gamma_2(\mathcal{C}, \|\cdot\|_2) \ \le \ \kappa c_4 w(\mathcal{C}) \ ,$$

where c_4 is an absolute constant. Hence, by choosing $\beta = \frac{c_1 c_4 \kappa^2 w(\mathcal{C})}{\sqrt{n}}$, we can guarantee that condition $c_1 \kappa \gamma_2(\mathcal{H}, |||\cdot|||_{\psi_2}) \le \beta \sqrt{n}$ holds for $\mathcal{H}$. Applying Lemma 3 to this $\mathcal{H}$, with probability at least $1 - \exp(-c_2 c_1^2 c_4^2 w^2(\mathcal{C}))$, we have

$$\sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} h^2(\mathbf{x}_i) - 1 \right| \le \beta \ ,$$

which implies $\inf_{\mathbf{v} \in \mathcal{C}} \frac{1}{n} \|\mathbf{X}\mathbf{v}\|_2^2 \ge 1 - \beta$. Letting $C_0 = c_1 c_4$ and $C_1 = c_2 c_1^2 c_4^2$, we complete the proof. ∎

3.B.2 Proof of Theorem 6

Proof: We first bound the magnitude of the error vector $\boldsymbol{\epsilon}$. For each entry in $\boldsymbol{\epsilon}$, we have

$$\sqrt{\mathbb{E}[\epsilon_i^2]} \le \sqrt{2} |||\epsilon_i|||_{\psi_2} = \sqrt{2} \tau \ ,$$
$$\left|\left|\left| \epsilon_i^2 - \mathbb{E}[\epsilon_i^2] \right|\right|\right|_{\psi_1} \le 2 \left|\left|\left| \epsilon_i^2 \right|\right|\right|_{\psi_1} \le 4 |||\epsilon_i|||_{\psi_2}^2 \le 4\tau^2 \ ,$$

where we use the definition of ψ_2-norm and Proposition 13. By Bernstein-type inequality, we get

$$\mathbb{P}\left(\|\boldsymbol{\epsilon}\|_2^2 - 2\tau^2 \geq t\right) \leq \mathbb{P}\left(\|\boldsymbol{\epsilon}\|_2^2 - \mathbb{E}[\|\boldsymbol{\epsilon}\|_2^2] \geq t\right) \leq \exp\left(-c_1 \min\left(\frac{t^2}{16\tau^4 n}, \frac{t}{4\tau^2}\right)\right) .$$

Taking $t = 4\tau^2 n$, we have

$$\mathbb{P}\left(\|\boldsymbol{\epsilon}\|_2 \geq \tau\sqrt{6n}\right) \leq \exp\left(-c_1 n\right) .$$

Next we bound the quantity $\|\mathbf{X}^T\mathbf{u}\|_*$ for any fixed unit vector $\mathbf{u}$. For any fixed $\mathbf{u} \in \mathbb{S}^{n-1}$, we have $\left|\!\left|\!\left|\mathbf{X}^T\mathbf{u}\right|\!\right|\!\right|_{\psi_2} \leq c\kappa$ since

$$\left|\!\left|\!\left|\langle\mathbf{X}^T\mathbf{u}, \mathbf{v}\rangle\right|\!\right|\!\right|_{\psi_2} = |\!|\!|\langle\mathbf{u}, \mathbf{X}\mathbf{v}\rangle|\!|\!|_{\psi_2} \leq |\!|\!|\mathbf{X}\mathbf{v}|\!|\!|_{\psi_2} \leq \ c\kappa \ \text{ for any } \mathbf{v} \in \mathbb{S}^{p-1} ,$$

where the last inequality is obtained by noting that $\mathbf{X}\mathbf{v}$ consists of i.i.d. sub-Gaussian entries with ψ_2-norm bounded by κ. Fixing $\mathbf{u} \in \mathbb{S}^{n-1}$, we construct the stochastic process $\{Z_{\mathbf{v}} = \langle\mathbf{X}^T\mathbf{u}, \mathbf{v}\rangle\}_{\mathbf{v}\in\Omega}$, and note that any $Z_{\mathbf{v}_1}$ and $Z_{\mathbf{v}_2}$ from this process satisfy

$$\mathbb{P}\left(|Z_{\mathbf{v}_1} - Z_{\mathbf{v}_2}| \geq t\right) = \mathbb{P}\left(\left|\langle\mathbf{X}^T\mathbf{u}, \mathbf{v}_1 - \mathbf{v}_2\rangle\right| \geq t\right) \leq e \cdot \exp\left(-\frac{Ct^2}{\kappa^2\|\mathbf{v}_1 - \mathbf{v}_2\|_2^2}\right) ,$$

which implies that $\{Z_{\mathbf{v}}\}$ has sub-Gaussian incremental w.r.t. the metric $s(\mathbf{v}_1, \mathbf{v}_2) = \kappa\|\mathbf{v}_1 - \mathbf{v}_2\|_2$. Moreover, as Ω is symmetric, it follows that

$$\sup_{\mathbf{v}_1,\mathbf{v}_2\in\Omega} |Z_{\mathbf{v}_1} - Z_{\mathbf{v}_2}| = 2\sup_{\mathbf{v}\in\Omega} Z_{\mathbf{v}}$$

$$\sup_{\mathbf{v}_1,\mathbf{v}_2\in\Omega} \|\mathbf{v}_1 - \mathbf{v}_2\|_2 = 2\sup_{\mathbf{v}\in\Omega} \|\mathbf{v}\|_2 = 2\rho$$

Using Lemma 2, we have

$$\mathbb{P}\left(2\sup_{\mathbf{v}\in\Omega} Z_{\mathbf{v}} \geq c_4\kappa\left(\gamma_2(\Omega,\|\cdot\|_2)+t\cdot 2\rho\right)\right) \leq c_2\exp\left(-t^2\right)\ ,$$

where c_2 and c_4 are absolute constant. By the definition of dual norm and Lemma 4, there exist constants c_3 and c_5 such that

$$\mathbb{P}\left(2\left\|\mathbf{X}^T\mathbf{u}\right\|_* \geq c_5\kappa\left(w\left(\Omega\right)+t\right)\right) = \mathbb{P}\left(2\sup_{\mathbf{v}\in\Omega} Z_{\mathbf{v}} \geq c_5\kappa\left(w\left(\Omega\right)+t\right)\right) \leq c_2\exp\left(-\frac{t^2}{c_3^2\rho^2}\right).$$

Letting $t = w(\Omega)$, we have for any fixed $\mathbf{u}\in\mathbb{S}^{n-1}$

$$\mathbb{P}\left(\left\|\mathbf{X}^T\mathbf{u}\right\|_* \geq c_5\kappa w\left(\Omega\right)\right) \leq c_2\exp\left(-\frac{w^2\left(\Omega\right)}{c_3^2\rho^2}\right)\ .$$

Combining this with the bound for $\|\boldsymbol{\epsilon}\|_2$ and letting $c_0 = \sqrt{6}c_5$, by union bound and the independence between $\mathbf{X}$ and $\boldsymbol{\epsilon}$, we have

$$\begin{aligned}\mathbb{P}\left(\left\|\mathbf{X}^T\boldsymbol{\epsilon}\right\|_* \geq c_0\kappa\tau\sqrt{n}w\left(\Omega\right)\right) &\leq \mathbb{P}\left(\frac{\|\mathbf{X}^T\boldsymbol{\epsilon}\|_*}{\|\boldsymbol{\epsilon}\|_2} \geq c_5\kappa w\left(\Omega\right)\right) + \mathbb{P}\left(\|\boldsymbol{\epsilon}\|_2 \geq \tau\sqrt{6n}\right)\\ &\leq \sup_{\mathbf{u}\in\mathbb{S}^{n-1}} \mathbb{P}\left(\|\mathbf{X}^T\mathbf{u}\|_* \geq c_5\kappa w\left(\Omega\right)\right) + \mathbb{P}\left(\|\boldsymbol{\epsilon}\|_2 \geq \tau\sqrt{6n}\right)\\ &\leq c_2\exp\left(-\frac{w^2\left(\Omega\right)}{c_3^2\rho^2}\right) + \exp\left(-c_1 n\right)\ ,\end{aligned}$$

which completes the proof. ■

Chapter 4

Geometric Measures with Atomic Norms

4.1 Introduction

In Chapter 3, we propose the generalized Dantzig selector (GDS) (3.2) for structured linear models, where structure of the parameter $\boldsymbol{\theta}^*$ is captured by a general norm $\|\dot{\|}$. In particular, we show that the L_2-error $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2$ of the estimate $\hat{\boldsymbol{\theta}}$ given by GDS are determined by three *geometric measures*: (i) $w(\Omega)$, the Gaussian width of the unit norm ball, (ii) $w(\mathcal{C})$, the Gaussian width of the error spherical cap $\mathcal{C}$, and (iii) Ψ, the restricted norm compatibility, where $\|\cdot\|$ is the norm used in GDS to capture the structure of $\boldsymbol{\theta}^*$. To be specific, if sample size $n \geq O(w^2(\mathcal{C}))$ and the tuning parameter satisfies $\lambda_n = O\left(\frac{w(\Omega)}{\sqrt{n}}\right)$, then the following error bound hold with high probability,

$$\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_2 \leq O\left(\frac{\Psi \cdot w(\Omega)}{\sqrt{n}}\right) \ . \tag{4.1}$$

In order to make use of this result, the geometric measures should be easy to compute or upper bound, which otherwise will render the bound meaningless. For the simple case like L_1 norm, accurate characterization of all three measures exists [40, 127]. However, for more general norms, the literature is rather limited. For $w(\Omega)$, the characterization is often reduced to comparison with either $w(\mathcal{C})$ [14] or known results on other norm balls [54]. While $w(\mathcal{C})$ has been investigated for certain decomposable norms [4, 36, 40], little is known about general non-decomposable norms. One general approach for upper bounding $w(\mathcal{C})$ is via the *statistical dimension* [4, 40, 132], which computes the expected squared distance between a Gaussian random vector and the *polar cone* of $\mathcal{T}$. To specify the polar, one need full information of the subdifferential $\partial\|\boldsymbol{\theta}^*\|$, which could be difficult to obtain for non-decomposable norms. A notable bound for (overlapping) $L_{2,1}$ norms is presented in [138], which yields tight bounds for mildly non-overlapping cases, but is loose for highly overlapping ones. For Ψ, the restricted norm compatibility, results are only available for decomposable norms [14, 127].

In this chapter, we consider the class of atomic norms $\|\cdot\|_{\mathcal{A}}$ that are *invariant under sign-changes*, i.e., the norm of a vector stays unchanged if any entry changes only by flipping its sign. The class is quite general, and covers most of the popular norms used in practical applications, e.g., L_1 norm, ordered weighted L_1 (OWL) norm [22] and k-support norm [7]. For such atomic norms, we confirm the practicability of error bound (4.1) by presenting a set of general bounds for their Gaussian width $w(\Omega_{\mathcal{A}})$, $w(\mathcal{C}_{\mathcal{A}})$, and the restricted norm compatibility $\Psi_{\mathcal{A}}$. Specifically we show that sharp bounds on $w(\Omega_{\mathcal{A}})$ can be obtained using simple calculation based on a decomposition inequality from [118]. To upper bound $w(\mathcal{C}_{\mathcal{A}})$ and $\Psi_{\mathcal{A}}$, instead of a full specification of $\mathcal{T}_{\mathcal{A}}$, we only require some information regarding the subgradient of $\|\boldsymbol{\theta}^*\|_{\mathcal{A}}$, which is often readily accessible. The key insight is that bounding statistical dimension often ends up computing the expected distance from Gaussian vector to a single point rather than to the whole polar cone, thus the full information on $\partial\|\boldsymbol{\theta}^*\|_{\mathcal{A}}$ is unnecessary. In addition, we derive the corresponding lower bounds to show the tightness of our results on $w(\mathcal{C}_{\mathcal{A}})$ and $\Psi_{\mathcal{A}}$. As examples, we illustrate the bounds for L_1 and OWL norms [22]. Finally, we give sharp bounds for the recently proposed k-support norm [7], for which existing analysis is incomplete.

The rest of the chapter is organized as follows. In Section 4.2, we introduce the general upper bounds for the geometric measures. We discuss the corresponding lower bounds in Section 4.3. Section 4.4 is dedicated to the example of k-support norm.

4.2 General Upper Bounds

In this section, we present detailed analysis of the general bounds for the geometric measures. In general, knowing the atomic set $\mathcal{A}$ is sufficient for bounding $w(\Omega_{\mathcal{A}})$. For $w(\mathcal{C}_{\mathcal{A}})$ and $\Psi_{\mathcal{A}}$, we only need a single subgradient of $\|\boldsymbol{\theta}^*\|_{\mathcal{A}}$ and some simple additional calculations.

4.2.1 Gaussian Width of Unit Norm Ball

Although the atomic set $\mathcal{A}$ may contain uncountably many vectors, we assume that $\mathcal{A}$ can be decomposed as a union of M "simple" sets, $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2 \cup \ldots \cup \mathcal{A}_M$. By "simple", we mean the Gaussian width of each $\mathcal{A}_i$ is easy to compute/bound. Such a decomposition assumption is often satisfied by commonly used atomic norms, e.g., L_1, $L_{2,1}$, OWL, k-support norm. The Gaussian width of the unit norm ball of $\|\cdot\|_{\mathcal{A}}$ can be easily obtained using the following lemma, which is essentially the Lemma 2 in [118].

Lemma 7 *Let $M > 4$, $\mathcal{A}_1, \cdots, \mathcal{A}_M \subset \mathbb{R}^p$, and $\mathcal{A} = \cup_m \mathcal{A}_m$. The Gaussian width of* *unit norm ball of* $\|\cdot\|_{\mathcal{A}}$ *satisfies*

$$w(\Omega_{\mathcal{A}}) \leq \max_{1\leq m\leq M} w(\mathcal{A}_m) + 2\sup_{\mathbf{z}\in\mathcal{A}} \|\mathbf{z}\|_2 \sqrt{\log M} \tag{4.2}$$

Proof: This inequality is a direct result of Lemma 1 and the properties of atomic norm and Gaussian width, $w(\Omega_{\mathcal{A}}) = w(\mathrm{conv}(\mathcal{A})) = w(\mathcal{A})$. ■

Next we illustrate application of this result to bounding the Gaussian width of the unit norm ball of L_1 and OWL norm.

Example 1 ($\boldsymbol{w(\Omega_{\mathcal{A}})}$ for $\boldsymbol{L_1}$ norm): Recall that the L_1 norm can be viewed as the atomic norm induced by the set $\mathcal{A}_{L_1} = \{\pm\mathbf{e}_i \ : \ 1 \leq i \leq p\}$, where $\{\mathbf{e}_i\}_{i=1}^p$ is the canonical basis of $\mathbb{R}^p$. Since the Gaussian width of a singleton is 0, if we treat $\mathcal{A}$ as the union of individual $\{+\mathbf{e}_i\}$ and $\{-\mathbf{e}_i\}$, we have

$$w(\Omega_{L_1}) \leq 0 + 2\sqrt{\log 2p} = O(\sqrt{\log p}) \ . \tag{4.3}$$

Example 2 ($\boldsymbol{w(\Omega_{\mathcal{A}})}$ for OWL norm): A recent variant of L_1 norm is the so-called *ordered weighted* L_1 *(OWL)* norm [22, 54, 185] defined as $\|\boldsymbol{\theta}\|_{\text{owl}} = \sum_{i=1}^p w_i |\boldsymbol{\theta}|_i^{\downarrow}$, where $w_1 \geq w_2 \geq \ldots \geq w_p \geq 0$ are pre-specified ordered weights, and $|\boldsymbol{\theta}|^{\downarrow}$ is the permutation of $|\boldsymbol{\theta}|$ with entries sorted in decreasing order. In [185], the OWL norm is proved to be an atomic norm with atomic set

$$\mathcal{A}_{\text{owl}} = \bigcup_{1\leq i\leq p} \mathcal{A}_i = \bigcup_{1\leq i\leq p} \ \bigcup_{|\,\mathrm{supp}(\mathcal{S})|=i} \left\{ \mathbf{u}\in\mathbb{R}^p \ : \ \mathbf{u}_{\mathcal{S}^c} = \mathbf{0}, \mathbf{u}_{\mathcal{S}} = \frac{\mathbf{v}_{\mathcal{S}}}{\sum_{j=1}^i w_j}, \mathbf{v}\in\{\pm 1\}^p \right\} \ .$$

We first apply Lemma 7 to each set $\mathcal{A}_i$, and note that each $\mathcal{A}_i$ contains $2^i\binom{p}{i}$ atomic vectors.

$$w(\mathcal{A}_i) \leq 0 + 2\sqrt{\frac{i}{(\sum_{j=1}^{i} w_j)^2}}\sqrt{\log 2^i \binom{p}{i}} \leq \frac{2i}{\sum_{j=1}^{i} w_j}\sqrt{2 + \log\left(\frac{p}{i}\right)} \leq \frac{2}{\bar{w}}\sqrt{2 + \log\left(\frac{p}{i}\right)} ,$$

where $\bar{w}$ is the average of $w_1, \ldots, w_p$. Then we apply the lemma again to $\mathcal{A}_{\text{owl}}$ and obtain

$$w(\Omega_{\text{owl}}) = w(\mathcal{A}_{\text{owl}}) \leq \frac{2}{\bar{w}}\sqrt{2 + \log p} + \frac{2}{\bar{w}}\sqrt{\log p} = O\left(\frac{\sqrt{\log p}}{\bar{w}}\right) , \tag{4.4}$$

which matches the result in [54].

4.2.2 Gaussian Width of Error Spherical Cap

In this subsection, we consider the computation of general $w(\mathcal{C}_{\mathcal{A}})$. Using the definition of dual norm, we can write $\|\boldsymbol{\theta}^*\|_{\mathcal{A}}$ as $\|\boldsymbol{\theta}^*\|_{\mathcal{A}} = \sup_{\|\mathbf{u}\|_{\mathcal{A}}^* \leq 1} \langle \mathbf{u}, \boldsymbol{\theta}^* \rangle$, where $\|\cdot\|_{\mathcal{A}}^*$ denotes the dual norm of $\|\cdot\|_{\mathcal{A}}$. The $\mathbf{u}^*$ for which $\langle \mathbf{u}^*, \boldsymbol{\theta}^* \rangle = \|\boldsymbol{\theta}^*\|_{\mathcal{A}}$, is a subgradient of $\|\boldsymbol{\theta}^*\|_{\mathcal{A}}$. One can obtain $\mathbf{u}^*$ by simply solving the so-called *polar operator* [187] for the dual norm $\|\cdot\|_{\mathcal{A}}^*$,

$$\mathbf{u}^* \in \underset{\|\mathbf{u}\|_{\mathcal{A}}^* \leq 1}{\operatorname{argmax}} \ \langle \mathbf{u}, \boldsymbol{\theta}^* \rangle \ . \tag{4.5}$$

Based on polar operator, we start with the Lemma 8, which plays a key role in our analysis of $w(\mathcal{C}_{\mathcal{A}})$.

Lemma 8 *Let $\mathbf{u}^*$ be a solution to the polar operator* (4.5)*, and define the weighted L_1 semi-norm $\|\cdot\|_{\mathbf{u}^*}$ as $\|\mathbf{v}\|_{\mathbf{u}^*} = \sum_{i=1}^{p} |u_i^*| \cdot |v_i|$. Then the following relation holds*

$$\mathcal{T}_{\mathcal{A}} \subseteq \mathcal{T}_{\mathbf{u}^*} \ ,$$

where $\mathcal{T}_{\mathbf{u}^*} = \operatorname{cone}\{\mathbf{v} \in \mathbb{R}^p \mid \|\boldsymbol{\theta}^* + \mathbf{v}\|_{\mathbf{u}^*} \leq \|\boldsymbol{\theta}^*\|_{\mathbf{u}^*}\}$.

Proof: As both $\mathcal{T}_{\mathbf{u}^*}$ and $\mathcal{T}_{\mathbf{u}^*}$ are cones, it is sufficient to show that $\{\mathbf{v} \mid \|\mathbf{v}\|_{\mathcal{A}} \leq \|\boldsymbol{\theta}^*\|_{\mathcal{A}}\} \subseteq \{\mathbf{v} \mid \|\mathbf{v}\|_{\mathbf{u}^*} \leq \|\boldsymbol{\theta}^*\|_{\mathbf{u}^*}\}$. Since $\|\boldsymbol{\theta}^*\|_{\mathbf{u}^*} = \|\boldsymbol{\theta}^*\|_{\mathcal{A}}$, it also suffices to show that $\{\mathbf{v} \mid \|\mathbf{v}\|_{\mathcal{A}} \leq 1\} \subseteq \{\mathbf{v} \mid \|\mathbf{v}\|_{\mathbf{u}^*} \leq 1\}$, i.e., the $\|\mathbf{v}\|_{\mathcal{A}} \geq \|\mathbf{v}\|_{\mathbf{u}^*}$ for $\mathbf{v} \in \mathbb{R}^p$. Using the dual norm definition and sign-change invariance of $\|\cdot\|_{\mathcal{A}}^*$, we obtain

$$\|\mathbf{v}\|_{\mathcal{A}} = \sup_{\|\mathbf{a}\|_{\mathcal{A}}^* \leq 1} \langle \mathbf{a}, \mathbf{v} \rangle \geq \langle \mathrm{sign}(\mathbf{v}) \odot |\mathbf{u}^*|, \mathbf{v} \rangle = \langle |\mathbf{u}^*|, |\mathbf{v}| \rangle = \|\mathbf{v}\|_{\mathbf{u}^*} \ ,$$

thus $\mathcal{T}_{\mathcal{A}} \subseteq \mathcal{T}_{\mathbf{u}^*}$. ■

Lemma 8 finds a superset of the error cone through $\|\cdot\|_{\mathbf{u}^*}$, which has simpler structures that can be utilized in subsequent analysis. Note that the solution to (4.5) may not be unique. A good criterion for choosing $\mathbf{u}^*$ is to avoid zeros in $\mathbf{u}^*$, as any $u_i^* = 0$ will lead to the unboundedness of unit ball of $\|\cdot\|_{\mathbf{u}^*}$, which could potentially increase the size of $\mathcal{T}_{\mathbf{u}^*}$. Next we present the upper bound for $w(\mathcal{C}_{\mathcal{A}})$.

Theorem 8 *Suppose that $\mathbf{u}^*$ is one of the solutions to* (4.5)*, and define the following sets,*

$$\mathcal{Q} = \{i \mid u_i^* = 0\}, \qquad \mathcal{S} = \{i \mid u_i^* \neq 0,\ \theta_i^* \neq 0\}, \qquad \mathcal{R} = \{i \mid u_i^* \neq 0,\ \theta_i^* = 0\} \ .$$

The Gaussian width $w(\mathcal{C}_{\mathcal{A}})$ is upper bounded by

$$w(\mathcal{C}_{\mathcal{A}}) \leq \begin{cases} \sqrt{p} \ , & \text{if } \mathcal{R} \text{ is empty} \\ \\ \sqrt{m + \frac{3}{2}s + \frac{2\kappa_{\max}^2}{\kappa_{\min}^2} s \log\left(\frac{p-m}{s}\right)} \ , & \text{if } \mathcal{R} \text{ is nonempty} \end{cases} \ , \tag{4.6}$$

where $m = |\mathcal{Q}|$, $s = |\mathcal{S}|$, $\kappa_{\min} = \min_{i \in \mathcal{R}} |u_i^*|$ and $\kappa_{\max} = \max_{i \in \mathcal{S}} |u_i^*|$.

Suppose that $\boldsymbol{\theta}^*$ is a s-sparse vector. We illustrate the above bound on the Gaussian width of the error spherical cap using L_1 norm and OWL norm as examples.

Example 3 ($\boldsymbol{w(\mathcal{C}_{\mathcal{A}})}$ for $\boldsymbol{L_1}$ norm): The dual norm of L_1 is L_∞ norm, and its easy to verify that $\mathbf{u}^* = [1, 1, \ldots, 1]^T \in \mathbb{R}^p$ is a solution to (4.5). Applying Theorem 8 to $\mathbf{u}^*$, we have

$$w(\mathcal{C}_{L_1}) \leq \sqrt{\frac{3}{2}s + 2s\log\left(\frac{p}{s}\right)} = O\left(\sqrt{s + s\log\left(\frac{p}{s}\right)}\right) . \tag{4.7}$$

Example 4 ($\boldsymbol{w(\mathcal{C}_{\mathcal{A}})}$ for OWL norm): For OWL, its dual norm is given by $\|\mathbf{u}\|_{\mathrm{owl}}^* = \max_{\mathbf{b}\in\mathcal{A}_{\mathrm{owl}}}\langle \mathbf{b}, \mathbf{u}\rangle$. W.l.o.g. we assume $\boldsymbol{\theta}^* = |\boldsymbol{\theta}^*|^{\downarrow}$, and a solution to (4.5) is given by $\mathbf{u}^* = [w_1, \ldots, w_s, \tilde{w}, \tilde{w}, \ldots, \tilde{w}]^T$, in which $\tilde{w}$ is the average of $w_{s+1}, \ldots, w_p$. If all w_i's are nonzero, the Gaussian width satisfies

$$w(\mathcal{C}_{\mathrm{owl}}) \leq \sqrt{\frac{3}{2}s + \frac{2w_1^2}{\tilde{w}^2}s\log\left(\frac{p}{s}\right)} . \tag{4.8}$$

4.2.3 Restricted Norm Compatibility

The next theorem gives general upper bounds for the restricted norm compatibility $\Psi_{\mathcal{A}}$.

Theorem 9 *Assume that $\|\mathbf{u}\|_{\mathcal{A}} \leq \max\{\beta_1\|\mathbf{u}\|_1, \beta_2\|\mathbf{u}\|_2\}$ for all $\mathbf{u} \in \mathbb{R}^p$. Under the setting of Theorem 8, the restricted norm compatibility $\Psi_{\mathcal{A}}$ is upper bounded by*

$$\Psi_{\mathcal{A}} \leq \begin{cases} \Phi\ , \quad \textit{if } \mathcal{R} \textit{ is empty} \\ \Phi_{\mathcal{Q}} + \max\left\{\beta_2,\ \beta_1\left(1 + \frac{\kappa_{\max}}{\kappa_{\min}}\right)\sqrt{s}\right\}\ , \quad \textit{if } \mathcal{R} \textit{ is nonempty} \end{cases} , \tag{4.9}$$

where $\Phi = \sup_{\mathbf{u}\in\mathbb{R}^p} \frac{\|\mathbf{u}\|_{\mathcal{A}}}{\|\mathbf{u}\|_2}$ and $\Phi_{\mathcal{Q}} = \sup_{\mathrm{supp}(\mathbf{u})\subseteq\mathcal{Q}} \frac{\|\mathbf{u}\|_{\mathcal{A}}}{\|\mathbf{u}\|_2}$.

Proof: As analyzed in the proof of Theorem 8, $\mathbf{v}_{\mathcal{Q}}$ for $\mathbf{v} \in \mathcal{T}_{\mathbf{u}^*}$ can be arbitrary, and the $\mathbf{v}_{\mathcal{S}\cup\mathcal{R}} = \mathbf{v}_{\mathcal{Q}^c}$ satisfies

$$\|\mathbf{v}_{\mathcal{Q}^c} + \boldsymbol{\theta}^*_{\mathcal{Q}^c}\|_{\mathbf{u}^*} \leq \|\boldsymbol{\theta}^*_{\mathcal{Q}^c}\|_{\mathbf{u}^*} \quad \Longrightarrow \quad \sum_{i\in\mathcal{S}} |\theta_i^* + v_i||u_i^*| + \sum_{j\in\mathcal{R}} |v_j||u_j^*| \leq \sum_{i\in\mathcal{S}} |\theta_i^*||u_i^*|$$
$$\Longrightarrow \quad \sum_{i\in\mathcal{S}} \left(|\theta_i^*| - |v_i|\right)|u_i^*| + \sum_{j\in\mathcal{R}} |v_j||u_j^*| \leq \sum_{i\in\mathcal{S}} |\theta_i^*||u_i^*| \quad \Longrightarrow \quad \kappa_{\min}\|\mathbf{v}_{\mathcal{R}}\|_1 \leq \kappa_{\max}\|\mathbf{v}_{\mathcal{S}}\|_1$$

If $\mathcal{R}$ is empty, by Lemma 8, we obtain

$$\Psi_{\mathcal{A}} \leq \Psi_{\mathbf{u}^*} \triangleq \sup_{\mathbf{v}\in\mathcal{T}_{\mathbf{u}^*}} \frac{\|\mathbf{v}\|_{\mathcal{A}}}{\|\mathbf{v}\|_2} \leq \sup_{\mathbf{v}\in\mathbb{R}^p} \frac{\|\mathbf{v}\|_{\mathcal{A}}}{\|\mathbf{v}\|_2} = \Phi \ .$$

If $\mathcal{R}$ is nonempty, we have

$$\begin{aligned}
\Psi_{\mathcal{A}} \leq \Psi_{\mathbf{u}^*} &\leq \sup_{\mathbf{v}\in\mathcal{T}_{\mathbf{u}^*}} \frac{\|\mathbf{v}_{\mathcal{Q}}\|_{\mathcal{A}} + \|\mathbf{v}_{\mathcal{Q}^c}\|_{\mathcal{A}}}{\|\mathbf{v}\|_2} \\
&\leq \sup_{\substack{\mathrm{supp}(\mathbf{v})\subseteq\mathcal{Q},\ \mathrm{supp}(\mathbf{v}')\subseteq\mathcal{Q}^c \\ \kappa_{\min}\|\mathbf{v}'_{\mathcal{R}}\|_1\leq\kappa_{\max}\|\mathbf{v}'_{\mathcal{S}}\|_1}} \frac{\|\mathbf{v}\|_{\mathcal{A}} + \|\mathbf{v}'\|_{\mathcal{A}}}{\|\mathbf{v}+\mathbf{v}'\|_2} \\
&\leq \sup_{\mathrm{supp}(\mathbf{v})\subseteq\mathcal{Q}} \frac{\|\mathbf{v}\|_{\mathcal{A}}}{\|\mathbf{v}\|_2} + \sup_{\substack{\mathrm{supp}(\mathbf{v}')\subseteq\mathcal{Q}^c \\ \kappa_{\min}\|\mathbf{v}'_{\mathcal{R}}\|_1\leq\kappa_{\max}\|\mathbf{v}'_{\mathcal{S}}\|_1}} \frac{\max\{\beta_1\|\mathbf{v}'\|_1, \beta_2\|\mathbf{v}'\|_2\}}{\|\mathbf{v}'\|_2} \\
&\leq \Phi_{\mathcal{Q}} + \max\left\{\beta_2, \sup_{\mathrm{supp}(\mathbf{v}')\subseteq\mathcal{S}} \frac{\beta(1+\frac{\kappa_{\max}}{\kappa_{\min}})\|\mathbf{v}'\|_1}{\|\mathbf{v}'\|_2}\right\} \\
&\leq \Phi_{\mathcal{Q}} + \max\left\{\beta_2, \beta_1\left(1+\frac{\kappa_{\max}}{\kappa_{\min}}\right)\sqrt{s}\right\} \ ,
\end{aligned}$$

in which the last inequality in the first line uses the property of $\mathcal{T}_{\mathbf{u}^*}$. ■

Remark: We call Φ the *unrestricted norm compatibility*, and $\Phi_{\mathcal{Q}}$ the *subspace norm compatibility*, both of which are often easier to compute than $\Psi_{\mathcal{A}}$. The β_1 and β_2 in the assumption of $\|\cdot\|_{\mathcal{A}}$ can have multiple choices, and one has the flexibility to choose the one that yields the tightest bound.

Example 5 ($\boldsymbol{\Psi_{\mathcal{A}}}$ for $\boldsymbol{L_1}$ norm): To apply the Theorem 9 to L_1 norm, we can choose $\beta_1 = 1$ and $\beta_2 = 0$. We recall the $\mathbf{u}^*$ for L_1 norm, whose $\mathcal{Q}$ is empty while $\mathcal{R}$ is nonempty. So we have for s-sparse $\boldsymbol{\theta}^*$

$$\Psi_{L_1} \leq 0 + \max\left\{0, \left(1 + \frac{1}{1}\right)\sqrt{s}\right\} = 2\sqrt{s} \ .$$

Example 6 ($\boldsymbol{\Psi_{\mathcal{A}}}$ for OWL norm): For OWL, note that $\|\cdot\|_{\text{owl}} \leq w_1\|\cdot\|_1$. Hence we choose $\beta_1 = w_1$ and $\beta_2 = 0$. As a result, we similarly have for s-sparse $\boldsymbol{\theta}^*$

$$\Psi_{\text{owl}} \leq 0 + \max\left\{0, w_1\left(1 + \frac{w_1}{\tilde{w}}\right)\sqrt{s}\right\} \leq \frac{2w_1^2}{\tilde{w}}\sqrt{s} \ .$$

4.3 General Lower Bounds

So far we have shown that the geometric measures can be upper bounded for general atomic norms. One might wonder how tight the bounds in Section 4.2 are for these measures. For $w(\Omega_{\mathcal{A}})$, as the result from [118] depends on the decomposition of $\mathcal{A}$ for the ease of computation, it might be tricky to discuss its tightness in general. Hence we will focus on the other two, $w(\mathcal{C}_{\mathcal{A}})$ and $\Psi_{\mathcal{A}}$.

To characterize the tightness, we need to compare the lower bounds of $w(\mathcal{C}_{\mathcal{A}})$ and $\Psi_{\mathcal{A}}$, with their upper bounds determined by $\mathbf{u}^*$. While there can be multiple $\mathbf{u}^*$, it is easy to see that any convex combination of them is also a solution to (4.5). Therefore we can always find a $\mathbf{u}^*$ that has the largest support, i.e., $\text{supp}(\mathbf{u}') \subseteq \text{supp}(\mathbf{u}^*)$ for any other solution $\mathbf{u}'$. We will use such $\mathbf{u}^*$ to generate the lower bounds. First we need the following lemma for the cone $\mathcal{T}_{\mathcal{A}}$.

Lemma 9 *Consider a solution* $\mathbf{u}^*$ *to* (4.5)*, which satisfies* $\text{supp}(\mathbf{u}') \subseteq \text{supp}(\mathbf{u}^*)$ *for any other solution* $\mathbf{u}'$*. Under the setting of notations in Theorem 8, we define an additional* *set of coordinates* $\mathcal{P} = \{i \mid u_i^* = 0,\ \theta_i^* = 0\}$*. Then the tangent cone* $\mathcal{T}_{\mathcal{A}}$ *satisfies*

$$\mathcal{T}_1 \oplus \mathcal{T}_2 \subseteq \mathrm{cl}(\mathcal{T}_{\mathcal{A}}) \ , \tag{4.10}$$

where $\oplus$ *denotes the direct (Minkowski) sum operation,* $\mathrm{cl}(\cdot)$ *denotes the closure,* $\mathcal{T}_1 = \{\mathbf{v} \in \mathbb{R}^p \mid v_i = 0 \text{ for } i \notin \mathcal{P}\}$ *is a* $|\mathcal{P}|$*-dimensional subspace, and* $\mathcal{T}_2 = \{\mathbf{v} \in \mathbb{R}^p \mid \mathrm{sign}(v_i) = -\mathrm{sign}(\theta_i^*) \text{ for } i \in \mathrm{supp}(\boldsymbol{\theta}^*),\ v_i = 0 \text{ for } i \notin \mathrm{supp}(\boldsymbol{\theta}^*)\}$ *is a* $|\mathrm{supp}(\boldsymbol{\theta}^*)|$*-dimensional orthant.*

The following theorem gives us the lower bound for $w(\mathcal{C}_{\mathcal{A}})$ and $\Psi_{\mathcal{A}}$.

Theorem 10 *Under the setting of Theorem 8 and Lemma 9, the following lower bounds hold,*

$$w(\mathcal{C}_{\mathcal{A}}) \ \geq \ O(\sqrt{m+s}) \ , \tag{4.11}$$

$$\Psi_{\mathcal{A}} \ \geq \ \Phi_{\mathcal{Q}\cup\mathcal{S}} \ . \tag{4.12}$$

Proof: To lower bound $w(\mathcal{C}_{\mathcal{A}})$, we use Lemma 9 and the relation between Gaussian width and statistical dimension (Proposition 10.2 in [4]),

$$w(\mathcal{T}_{\mathcal{A}}) \geq w(\mathcal{T}_1 \oplus \mathcal{T}_2 \cap \mathbb{S}^{p-1}) \geq \sqrt{\mathbb{E}\left[\inf_{\mathbf{z}\in\mathcal{N}_{\mathcal{T}_1\oplus\mathcal{T}_2}} \|\mathbf{z}-\mathbf{g}\|_2^2\right] - 1} \quad (*) \ ,$$

where the normal cone $\mathcal{N}_{\mathcal{T}_1\oplus\mathcal{T}_2}$ of $\mathcal{T}_1 \oplus \mathcal{T}_2$ is given by $\mathcal{N}_{\mathcal{T}_1\oplus\mathcal{T}_2} = \{\mathbf{z} \ : z_i = 0 \ \text{ for } \ i \in \mathcal{P},\ \mathrm{sign}(z_i) = \mathrm{sign}(\theta_i^*) \ \text{ for } \ i \in \mathrm{supp}(\boldsymbol{\theta}^*)\}$. Hence we have

$$\begin{aligned}(*) \ &= \sqrt{\mathbb{E}\left[\sum_{i\in\mathcal{P}} g_i^2 + \sum_{j\in\mathrm{supp}(\boldsymbol{\theta}^*)} g_j^2 \mathbb{I}_{\{g_j\theta_j^*<0\}}\right] - 1} \\ &= \sqrt{|\mathcal{P}| + \frac{|\mathrm{supp}(\boldsymbol{\theta}^*)|}{2} - 1} = O(\sqrt{m+s}) \ ,\end{aligned}$$

where the last equality follows the fact that $\mathcal{P} \cup \text{supp}(\boldsymbol{\theta}^*) = \mathcal{Q} \cup \mathcal{S}$. This completes proof of (4.11). To prove (4.12), we again use Lemma 9 and the fact $\mathcal{P} \cup \text{supp}(\boldsymbol{\theta}^*) = \mathcal{Q} \cup \mathcal{S}$. Noting that $\|\cdot\|_{\mathcal{A}}$ is invariant under sign-changes, we get

$$\Psi_{\mathcal{A}} = \sup_{\mathbf{v} \in \mathcal{T}_{\mathcal{A}}} \frac{\|\mathbf{v}\|_{\mathcal{A}}}{\|\mathbf{v}\|_2} \geq \sup_{\mathbf{v} \in \mathcal{T}_1 \oplus \mathcal{T}_2} \frac{\|\mathbf{v}\|_{\mathcal{A}}}{\|\mathbf{v}\|_2} = \sup_{\text{supp}(\mathbf{v}) \subseteq \mathcal{P} \cup \text{supp}(\boldsymbol{\theta}^*)} \frac{\|\mathbf{v}\|_{\mathcal{A}}}{\|\mathbf{v}\|_2} = \Phi_{\mathcal{Q} \cup \mathcal{S}} . \qquad \blacksquare$$

Remark: We compare the lower bounds (4.11) (4.12) with the upper bounds (4.6) (4.9). If $\mathcal{R}$ is empty, $m + s = p$, and the lower bounds actually match the upper bounds up to a constant factor for both $w(\mathcal{C}_{\mathcal{A}})$ and $\Psi_{\mathcal{A}}$. If $\mathcal{R}$ is nonempty, the lower and upper bounds of $w(\mathcal{C}_{\mathcal{A}})$ differ by a multiplicative factor $\frac{2\kappa_{\max}^2}{\kappa_{\min}^2} \log(\frac{p-m}{s})$, which can be small in practice. For $\Psi_{\mathcal{A}}$, as $\Phi_{\mathcal{Q} \cup \mathcal{S}} \geq \Phi_{\mathcal{Q}}$, we usually have at most an additive $O(\sqrt{s})$ term in upper bound, since the assumption on $\|\cdot\|_{\mathcal{A}}$ often holds with a constant β_1 and $\beta_2 = 0$ for most norms.

4.4 Application to k-Support Norm

In this section, we apply our general results on geometric measures to a non-trivial example, k-support norm [7], which has been proved effective for sparse recovery [41,42, 120]. The definitions of k-support norm and its dual have been given in (3.12) and (3.13). The k-support norm can be viewed as an atomic norm, for which $\mathcal{A} = \{\mathbf{a} \in \mathbb{R}^p \mid \|\mathbf{a}\|_0 \leq k,\ \|\mathbf{a}\|_2 \leq 1\}$. Suppose that all the subsets of coordinates $\{1, 2, \ldots, p\}$ with cardinality k can be listed as $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_{\binom{p}{k}}$. Then $\mathcal{A}$ can be written as $\mathcal{A} = \mathcal{A}_1 \cup \ldots \cup \mathcal{A}_{\binom{p}{k}}$, where each $\mathcal{A}_i = \{\mathbf{a} \in \mathbb{R}^p \mid \text{supp}(\mathbf{a}) \subseteq \mathcal{S}_i,\ \|\mathbf{a}\|_2 \leq 1\}$. It is not difficult to see that $w(\mathcal{A}_i) = \mathbb{E}\left[\sup_{\mathbf{a} \in \mathcal{A}_i} \langle \mathbf{a}, \mathbf{g} \rangle\right] = \mathbb{E}\|\mathbf{g}_{\mathcal{S}_i}\|_2 \leq \sqrt{\mathbb{E}\|\mathbf{g}_{\mathcal{S}_i}\|_2^2} \leq \sqrt{k}$. Using Lemma 7, we know the Gaussian width of the unit ball of k-support norm

$$w(\Omega_k^{sp}) \leq \sqrt{k} + 2\sqrt{\log \binom{p}{k}} \leq \sqrt{k} + 2\sqrt{k \log \left(\frac{p}{k}\right) + k} = O\left(\sqrt{k \log \left(\frac{p}{k}\right) + k}\right) . \quad (4.13)$$

Now we turn to the calculation of $w(\mathcal{C}_k^{sp})$ and Ψ_k^{sp}. As we have seen in the general analysis, the solution $\mathbf{u}^*$ to the polar operator (4.5) is important in characterizing the two geometric measures. We first present a simple procedure in Algorithm 6 for solving the polar operator for $\|\cdot\|_{k*}^{sp}$. The time complexity is only $O(p \log p + k)$. This procedure can be utilized to compute the k-support norm, or be applied to estimation with $\|\cdot\|_k^{sp^*}$ using *generalized conditional gradient* method [187], which requires solving the polar operator in each iteration.

Algorithm 6 Solving polar operator for $\|\cdot\|_{k*}^{sp}$

Input: $\boldsymbol{\theta}^* \in \mathbb{R}^p$, positive integer k

Output: Solution $\mathbf{u}^*$ to the polar operator (4.5)

```
z = |θ*|↓,  t = 0
for i = 1 to k do
    γ1 = ‖z_{1:i−1}‖_2,  γ2 = ‖z_{i:p}‖_1,  d = k − i + 1,  β = γ2 / sqrt(γ2² d + γ1² d²),  α = γ1 / (2 sqrt(1 − β² d)),
    w = z_{1:i−1} / (2α)
    if γ1²/(2α) + βγ2 > t  and  β < w_{i−1} then
        t = γ1²/(2α) + βγ2,  u* = [w, β1]^T    (1 is (p − i + 1)-dimensional vector with all ones)
    end if
end for
change the sign and order of u* to conform with θ*
return  u*
```

$\mathbf{z} = |\boldsymbol{\theta}^*|^{\downarrow}$, $t = 0$; $\gamma_1 = \|\mathbf{z}_{1:i-1}\|_2$, $\gamma_2 = \|\mathbf{z}_{i:p}\|_1$, $d = k - i + 1$, $\beta = \frac{\gamma_2}{\sqrt{\gamma_2^2 d + \gamma_1^2 d^2}}$, $\alpha = \frac{\gamma_1}{2\sqrt{1-\beta^2 d}}$, $\mathbf{w} = \frac{\mathbf{z}_{1:i-1}}{2\alpha}$; condition $\frac{\gamma_1^2}{2\alpha} + \beta\gamma_2 > t$ and $\beta < w_{i-1}$; update $t = \frac{\gamma_1^2}{2\alpha} + \beta\gamma_2$, $\mathbf{u}^* = [\mathbf{w}, \beta\mathbf{1}]^T$.

Theorem 11 *For a given $\boldsymbol{\theta}^*$, Algorithm 6 returns a solution to polar operator* (4.5) *for $\|\cdot\|_{k*}^{sp}$.*

Now we consider $w(\mathcal{C}_k^{sp})$ and Ψ_k^{sp} for s-sparse $\boldsymbol{\theta}^*$ (here s-sparse $\boldsymbol{\theta}^*$ means $|\operatorname{supp}(\boldsymbol{\theta}^*)| = s$) in three scenarios: (i) over-specified k, where $s < k$, (ii) exactly specified k, where $s = k$, and (iii) under-specified k, where $s > k$. The bounds are given in Theorem 12.

Theorem 12 *For given s-sparse $\boldsymbol{\theta}^* \in \mathbb{R}^p$, the Gaussian width $w(\mathcal{C}_k^{sp})$ and the restricted* *norm compatibility* Ψ_k^{sp} *for a specified* k *are given by*

$$w(\mathcal{C}_k^{sp}) \leq \begin{cases} \sqrt{p} \ , \ \textit{if} \ \ s < k \\ \\ \sqrt{\frac{3}{2}s + \frac{2\theta^{*2}_{\max}}{\theta^{*2}_{\min}} s \log\left(\frac{p}{s}\right)} \ , \ \textit{if} \ \ s = k \\ \\ \sqrt{\frac{3}{2}s + \frac{2\kappa^2_{\max}}{\kappa^2_{\min}} s \log\left(\frac{p}{s}\right)} \ , \ \textit{if} \ \ s > k \end{cases} , \tag{4.14}$$

$$\Psi_k^{sp} \leq \begin{cases} \sqrt{\frac{2p}{k}} \ , \ \textit{if} \ \ s < k \\ \\ \sqrt{2}(1 + \frac{\theta^*_{\max}}{\theta^*_{\min}}) \ , \ \textit{if} \ \ s = k \\ \\ (1 + \frac{\kappa_{\max}}{\kappa_{\min}})\sqrt{\frac{2s}{k}} \ , \ \textit{if} \ \ s > k \end{cases} , \tag{4.15}$$

where $\theta^*_{\max} = \max_{i \in \mathrm{supp}(\boldsymbol{\theta}^*)} |\theta^*_i|$ *and* $\theta^*_{\min} = \min_{i \in \mathrm{supp}(\boldsymbol{\theta}^*)} |\theta^*_i|$.

Proof: For $s < k$, we note that $\|\boldsymbol{\theta}^*\|_k^{sp} = \|\boldsymbol{\theta}^*\|_2$, and $\mathbf{u}^*$ can be obtained in a closed-form $\mathbf{u}^* = \frac{\boldsymbol{\theta}^*}{\|\boldsymbol{\theta}^*\|_2}$. Applying Theorem 8, we find that the set $\mathcal{R}$ is empty, and thus the Gaussian width $w(\mathcal{C}_k^{sp}) = \sqrt{p}$. For $s = k$, $\mathbf{u}^*$ is in closed-form as well,

$$u_i^* = \begin{cases} \frac{\theta^*_i}{\|\boldsymbol{\theta}^*\|_2} \ , & \text{if} \ \ i \in \mathrm{supp}(\boldsymbol{\theta}^*) \\ \frac{|\boldsymbol{\theta}^*|^{\downarrow}_k}{\|\boldsymbol{\theta}^*\|_2} = \frac{\theta^*_{\min}}{\|\boldsymbol{\theta}^*\|_2} \ , & \text{if otherwise} \end{cases} .$$

In this case, $\mathcal{Q}$ is empty, $\mathcal{R}$ is nonempty, and $|\mathcal{S}| = s = k$. Hence Theorem 8 implies the corresponding Gaussian width, and $\frac{\kappa_{\max}}{\kappa_{\min}} = \frac{\theta^*_{\max}}{\theta^*_{\min}}$. For $s > k$, the closed-form solution is generally unavailable, but we can see from Algorithm 6 that β should be nonzero, thus $\mathcal{Q}$ is empty and $\mathcal{R}$ is nonempty, which gives us the corresponding Gaussian width.

Base on the analysis of $\boldsymbol{\theta}^*$, $\mathcal{Q}$, $\mathcal{R}$ and $\mathcal{S}$, and the fact that $\|\cdot\|_k^{sp} \leq \|\cdot\|_1$, the restricted norm compatibility constant for $s \geq k$ directly follows Theorem 9. For $s < k$, we need to compute the unrestricted norm compatibility constant Φ. As $\|\cdot\|_k^{sp} < \sqrt{2}\max\{\|\cdot\|_2, \frac{\|\cdot\|_1}{\sqrt{k}}\}$ shown in [7], we have

$$\Phi = \sup_{\mathbf{u}\in\mathbb{R}^p} \frac{\|\mathbf{u}\|_k^{sp}}{\|\mathbf{u}\|_2} \leq \sup_{\mathbf{u}\in\mathbb{R}^p} \frac{\sqrt{2}\max\{\|\mathbf{u}\|_2, \frac{\|\mathbf{u}\|_1}{\sqrt{k}}\}}{\|\mathbf{u}\|_2} \leq \max\{\sqrt{2}, \sqrt{\frac{2p}{k}}\} = \sqrt{\frac{2p}{k}}\ . \qquad \blacksquare$$

Remark: Previously Ψ_k^{sp} is unknown and the bound on $w(\mathcal{C}_k^{sp})$ given in [41] is loose, as it used the result in [138]. Based on Theorem 12, we note that the choice of k can affect the recovery guarantees. Over-specified k leads to a direct dependence on the dimensionality p for $w(\mathcal{C}_k^{sp})$ and Ψ_k^{sp}, resulting in a weak error bound. The bounds are sharp for exactly specified or under-specified k. Thus, it is better to under-specify k in practice. where the estimation error satisfies

$$\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_2 \leq O\left(\sqrt{\frac{s + s\log\left(\frac{p}{k}\right)}{n}}\right) \tag{4.16}$$

Appendix

Appendix 4.A Supplementary Proofs

4.A.1 Proof of Theorem 8

Proof: By Lemma 8, we have $w(\mathcal{C}_{\mathcal{A}}) \leq w(\mathcal{T}_{\mathbf{u}^*} \cap \mathbb{S}^{p-1}) \triangleq w(\mathcal{C}_{\mathbf{u}^*})$. Hence we can focus on bounding $w(\mathcal{C}_{\mathbf{u}^*})$. We first analyze the structure of $\mathbf{v}$ that satisfies $\|\boldsymbol{\theta}^* + \mathbf{v}\|_{\mathbf{u}^*} \leq \|\boldsymbol{\theta}^*\|_{\mathbf{u}^*}$. For the coordinates $\mathcal{Q} = \{i \mid u_i^* = 0\}$, the corresponding entries v_i's can be arbitrary since it does not affect the value of $\|\boldsymbol{\theta}^* + \mathbf{v}\|_{\mathbf{u}^*}$. Thus all possible $\mathbf{v}_{\mathcal{Q}}$ form a m-dimensional subspace, where $m = |\mathcal{Q}|$. For $\mathcal{S}\cup\mathcal{R} = \{i \mid u_i^* \neq 0\}$, we define $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^*_{\mathcal{S}\cup\mathcal{R}}$

and $\tilde{\mathbf{v}} = \mathbf{v}_{\mathcal{S}\cup\mathcal{R}}$, and $\tilde{\mathbf{v}}$ needs to satisfy

$$\|\tilde{\mathbf{v}} + \tilde{\boldsymbol{\theta}}\|_{\mathbf{u}^*} \leq \|\tilde{\boldsymbol{\theta}}\|_{\mathbf{u}^*} ,$$

which is similar to the L_1-norm error cone except that coordinates are weighted by $|\mathbf{u}^*|$. Therefore we use the techniques for proving the Proposition 3.10 in [40]. Based on the structure of $\mathbf{v}$, The normal cone at $\boldsymbol{\theta}^*$ for $\mathcal{T}_{\mathbf{u}^*}$ is given by

$$\begin{aligned}\mathcal{N} &= \{\mathbf{z} \;:\; \langle \mathbf{z}, \mathbf{v}\rangle \leq 0 \;\; \forall\, \mathbf{v} \;\; \text{s.t.} \;\; \|\mathbf{v} + \boldsymbol{\theta}^*\|_{\mathbf{u}^*} \leq \|\boldsymbol{\theta}^*\|_{\mathbf{u}^*}\} \\ &= \Big\{\mathbf{z} \;: z_i = 0 \;\; \text{for} \;\; i \in \mathcal{Q}, \; z_i = |u_i^*|\text{sign}(\tilde{\theta}_i)t \;\; \text{for} \;\; i \in \mathcal{S}, \\ &\qquad |z_i| \leq |u_i^*|t \;\; \text{for} \;\; i \in \mathcal{R}, \; \text{for any } t \geq 0\Big\} \; .\end{aligned}$$

Given a standard Gaussian random vector $\mathbf{g}$, using the relation between Gaussian width and statistical dimension (Proposition 2.4 and 10.2 in [4]), we have

$$\begin{aligned} w^2(\mathcal{C}_{\mathbf{u}^*}) &\leq \mathbb{E}\left[\inf_{\mathbf{z}\in\mathcal{N}} \|\mathbf{z} - \mathbf{g}\|_2^2\right] = \mathbb{E}\left[\inf_{\mathbf{z}\in\mathcal{N}} \sum_{i\in\mathcal{Q}} g_i^2 + \sum_{j\in\mathcal{S}} (z_j - g_j)^2 + \sum_{k\in\mathcal{R}} (z_k - g_k)^2\right] \\ &= |\mathcal{Q}| + \mathbb{E}\left[\inf_{\mathbf{z}_{\mathcal{S}\cup\mathcal{R}}\in\mathcal{N}} \sum_{j\in\mathcal{S}} (|u_j^*|\text{sign}(\tilde{\theta}_j)t - g_j)^2 + \sum_{k\in\mathcal{R}} (z_k - g_k)^2\right] \\ &\leq |\mathcal{Q}| + t^2 \sum_{j\in\mathcal{S}} |u_j^*|^2 + |\mathcal{S}| + \mathbb{E}\left[\sum_{k\in\mathcal{R}} \inf_{|z_k|\leq|u_k^*|t} (z_k - g_k)^2\right] \\ &\leq |\mathcal{Q}| + t^2 \sum_{j\in\mathcal{S}} |u_j^*|^2 + |\mathcal{S}| + \sum_{k\in\mathcal{R}} \frac{2}{\sqrt{2\pi}} \left(\int_{|u_k^*|t}^{+\infty} (g_k - |u_k^*|t)^2 \exp(\frac{-g_k^2}{2}) dg_k\right) \\ &\leq |\mathcal{Q}| + t^2 \sum_{j\in\mathcal{S}} |u_j^*|^2 + |\mathcal{S}| + \sum_{k\in\mathcal{R}} \frac{2}{\sqrt{2\pi}} \frac{1}{|u_k^*|t} \exp\left(-\frac{|u_k^*|^2 t^2}{2}\right) \qquad (*) \; . \end{aligned}$$

The details for the derivation above can be found in Appendix C of [40]. If $\mathcal{R}$ is empty, by taking $t = 0$, we have

$$(*) \ \leq |\mathcal{Q}| + t^2 \sum_{j \in \mathcal{S}} |u_j^*|^2 + |\mathcal{S}| = |\mathcal{Q}| + |\mathcal{S}| = p \ .$$

If $\mathcal{R}$ is nonempty, we denote $\kappa_{\min} = \min_{i \in \mathcal{R}} |u_i^*|$ and $\kappa_{\max} = \max_{i \in \mathcal{S}} |u_i^*|$. Taking $t = \frac{1}{\kappa_{\min}} \sqrt{2 \log \left(\frac{|\mathcal{S} \cup \mathcal{R}|}{|\mathcal{S}|} \right)}$, we obtain

$$\begin{aligned} (*) \ &\leq |\mathcal{Q}| + |\mathcal{S}|(\kappa_{\max}^2 t^2 + 1) + \frac{2|\mathcal{R}| \exp\left(-\frac{\kappa_{\min}^2 t^2}{2} \right)}{\sqrt{2\pi} \kappa_{\min} t} \\ &= |\mathcal{Q}| + |\mathcal{S}| \left(\frac{2\kappa_{\max}^2}{\kappa_{\min}^2} \log \left(\frac{|\mathcal{S} \cup \mathcal{R}|}{|\mathcal{S}|} \right) + 1 \right) + \frac{|\mathcal{R}||\mathcal{S}|}{|\mathcal{S} \cup \mathcal{R}| \sqrt{\pi \log \left(\frac{|\mathcal{S} \cup \mathcal{R}|}{|\mathcal{S}|} \right)}} \\ &\leq |\mathcal{Q}| + \frac{2\kappa_{\max}^2}{\kappa_{\min}^2} |\mathcal{S}| \log \left(\frac{|\mathcal{S} \cup \mathcal{R}|}{|\mathcal{S}|} \right) + \frac{3}{2} |\mathcal{S}| \ . \end{aligned}$$

Substituting $|\mathcal{Q}| = m$, $|\mathcal{S}| = s$ and $|\mathcal{S} \cup \mathcal{R}| = p - m$ into the last inequality completes the proof. ■

4.A.2 Proof of Lemma 9

Proof: For any fixed $\boldsymbol{\theta}^* \in \mathbb{R}^p$ and its $\mathcal{P}$, we define a vector sequence $\{\mathbf{v}^{(k)} = \delta^{(k)} \mathbf{w}\}$ based on a given $\mathbf{w} \in \mathbb{R}^p$ and a monotonically decreasing positive scalar sequence $\{\delta^{(k)}\}$ with $\delta^{(1)} < \min_{i \in \mathrm{supp}(\boldsymbol{\theta}^*)} |\theta_i^*|$ and $\lim_{k \to +\infty} \delta^{(k)} = 0$. $\mathbf{w}$ satisfies

$$w_i = \begin{cases} 0 \ , \quad \text{if} \ \ i \notin \mathcal{P} \cup \mathrm{supp}(\boldsymbol{\theta}^*) \\ -\mathrm{sign}(\theta_i^*) \ , \quad \text{if} \ \ i \in \mathrm{supp}(\boldsymbol{\theta}^*) \\ \text{arbitrary} \ , \quad \text{if} \ \ i \in \mathcal{P} \end{cases} .$$

Let $\mathbf{u}^{(k)}$ be one solution to the polar operator for $\boldsymbol{\theta}^* + \mathbf{v}^{(k)}$, and form another sequence $\{\mathbf{u}^{(k)}\}$. Note that $\text{sign}(\theta_i^* + v_i^{(k)}) = \text{sign}(\theta_i^* - \text{sign}(\theta_i^*)\delta^{(k)}) = \text{sign}(\theta_i^*) = \text{sign}(u_i^{(k)})$ for $i \in \text{supp}(\boldsymbol{\theta}^*)$. Then we have

$$\begin{aligned}
\|\boldsymbol{\theta}^* + \mathbf{v}^{(k)}\|_{\mathcal{A}} - \|\boldsymbol{\theta}^*\|_{\mathcal{A}} &\leq \langle \boldsymbol{\theta}^* + \mathbf{v}^{(k)}, \mathbf{u}^{(k)} \rangle - \langle \boldsymbol{\theta}^*, \mathbf{u}^{(k)} \rangle = \langle \mathbf{v}^{(k)}, \mathbf{u}^{(k)} \rangle \\
&\leq -\delta^{(k)} \|\mathbf{u}^{(k)}_{\text{supp}(\boldsymbol{\theta}^*)}\|_1 + \delta^{(k)} \langle \mathbf{w}_{\mathcal{P}}, \mathbf{u}^{(k)}_{\mathcal{P}} \rangle \\
&\leq -\delta^{(k)} (\|\mathbf{u}^{(k)}_{\text{supp}(\boldsymbol{\theta}^*)}\|_1 - \|\mathbf{w}_{\mathcal{P}}\|_\infty \|\mathbf{u}^{(k)}_{\mathcal{P}}\|_1)
\end{aligned}$$

As $\delta^{(k)}$ approaches 0, $\boldsymbol{\theta}^* + \mathbf{v}^{(k)}$ converges to $\boldsymbol{\theta}^*$, and a subsequence $\{\mathbf{u}^{(k_i)}\}$ of $\{\mathbf{u}^{(k)}\}$ will converge to a solution $\mathbf{u}'$ to the polar operator for $\boldsymbol{\theta}^*$. Hence $\lim_{i \to +\infty} \|\mathbf{u}^{(k_i)}_{\text{supp}(\boldsymbol{\theta}^*)}\|_1 = \|\mathbf{u}'_{\text{supp}(\boldsymbol{\theta}^*)}\|_1 > 0$, $\lim_{i \to +\infty} \|\mathbf{u}^{(k_i)}_{\mathcal{P}}\|_1 = \|\mathbf{u}'_{\mathcal{P}}\|_1 = 0$, and for large enough k_i, we have

$$\|\boldsymbol{\theta}^* + \mathbf{v}^{(k_i)}\|_{\mathcal{A}} - \|\boldsymbol{\theta}^*\|_{\mathcal{A}} \leq -\delta^{(k_i)} (\|\mathbf{u}^{(k_i)}_{\text{supp}(\boldsymbol{\theta}^*)}\|_1 - \|\mathbf{w}_{\mathcal{P}}\|_\infty \|\mathbf{u}^{(k_i)}_{\mathcal{P}}\|_1) \leq 0,$$

thus $\mathbf{v}^{(k_i)}$ belongs to $\mathcal{T}_{\mathcal{A}}$. Since $\mathbf{v}^{(k)} = \delta^{(k)}\mathbf{w}$, $\mathbf{w}$ also belongs to $\mathcal{T}_{\mathcal{A}}$.

Now we show $\mathcal{T}_1 \subseteq \text{cl}(\mathcal{T}_{\mathcal{A}})$. For any $\mathbf{a} \in \mathcal{T}_1 = \{\mathbf{v} \in \mathbb{R}^p \mid v_i = 0 \text{ for } i \notin \mathcal{P}\}$ and arbitrarily small $\xi > 0$, we construct $\mathbf{w}$ such that $w_i = \frac{a_i}{\xi}$ for $i \in \mathcal{P}$. Based on the argument above, this $\mathbf{w}$ is in $\mathcal{T}_{\mathcal{A}}$. Therefore $\mathbf{a}' \triangleq \xi\mathbf{w} \in \mathcal{T}_{\mathcal{A}}$, and $\|\mathbf{a} - \mathbf{a}'\|_2 \leq \sqrt{|\text{supp}(\boldsymbol{\theta}^*)|}\xi$, which can be arbitrarily close to 0. Therefore taking the closure of $\mathcal{T}_{\mathcal{A}}$ gives us $\mathcal{T}_1 \subseteq \text{cl}(\mathcal{T}_{\mathcal{A}})$.

Next we show $\mathcal{T}_2 \subseteq \mathcal{T}_{\mathcal{A}}$. For any coordinate $i \in \text{supp}(\boldsymbol{\theta}^*)$, construct $\mathbf{v} \in \mathbb{R}^p$ such that $v_i = -\theta_i^*$ and $v_j = 0$ for $j \neq i$, and $\boldsymbol{\theta}' \in \mathbb{R}^p$ such that $\theta_i' = -\theta_i^*$ and $\theta_j' = \theta_j^*$ for $j \neq i$. As the norm $\|\cdot\|_{\mathcal{A}}$ is invariant under sign-changes, we can verify that

$$\|\boldsymbol{\theta}^* + \mathbf{v}\|_{\mathcal{A}} = \|\frac{\boldsymbol{\theta}^* + \boldsymbol{\theta}'}{2}\| \leq \frac{1}{2}\|\boldsymbol{\theta}^*\|_{\mathcal{A}} + \frac{1}{2}\|\boldsymbol{\theta}'\|_{\mathcal{A}} = \|\boldsymbol{\theta}^*\|_{\mathcal{A}} \ .$$

Thus $\mathbf{v} \in \mathcal{T}_{\mathcal{A}}$. Repeat the construction of $\mathbf{v}$ for each $i \in \text{supp}(\boldsymbol{\theta}^*)$, and then the conic combination of these $\mathbf{v}$'s forms $\mathcal{T}_2$. Therefore we have $\mathcal{T}_2 \subseteq \mathcal{T}_\mathcal{A}$, which together with $\mathcal{T}_1 \subseteq \text{cl}(\mathcal{T}_\mathcal{A})$ implies $\mathcal{T}_1 \oplus \mathcal{T}_2 \subseteq \text{cl}(\mathcal{T}_\mathcal{A})$. ■

4.A.3 Proof of and Theorem 11

Proof: The polar operator for 2-k symmetric gauge norm is essentially

$$\mathbf{u}^* = \text{argmax } \langle \mathbf{u}, \boldsymbol{\theta}^* \rangle \quad \text{s.t.} \quad \|\mathbf{u}^*\|_{k*}^{sp} \leq 1 \ .$$

As 2-k symmetric gauge norm is sign and permutation invariant, $\mathbf{u}^*$ should conform with the sign and order of $\boldsymbol{\theta}^*$ in order to achieve maxima, i.e., $\langle \mathbf{u}^*, \boldsymbol{\theta}^* \rangle \leq \langle |\mathbf{u}^*|^{\downarrow}, |\boldsymbol{\theta}^*|^{\downarrow} \rangle$. W.l.o.g, we assume $\boldsymbol{\theta}^* = |\boldsymbol{\theta}^*|^{\downarrow} \triangleq \mathbf{z}$. Now we analyze the structure of the solution $\mathbf{u}^*$, whose entries should be nonnegative and sorted in descending order. Assume that u_k^* takes certain fixed but unknown value β. It is easy that the entries in $\mathbf{u}_{k+1:p}^*$ can take the value of β, as it will always maximize $\langle \mathbf{u}_{k+1:p}^*, \boldsymbol{\theta}_{k+1:p}^* \rangle$ without violating the constraint $\|\mathbf{u}^*\|_{(k)} \leq 1$. Generally we also assume that $\mathbf{u}_{i:k}^*$ take the value of β and $u_{i-1}^* > u_i^*$. Then the maximization problem becomes

$$\begin{aligned} &\max_{\mathbf{u}_{1:i-1}, \beta} \ \langle \mathbf{u}_{1:i-1}, \mathbf{z}_{1:i-1} \rangle + \beta \|\mathbf{z}_{i:p}\|_1 \\ &\text{s.t.} \quad \|\mathbf{u}_{1:i-1}\|_2^2 \leq 1 - (k-i+1)\beta^2, \ u_j > \beta \ \text{ for } \ 1 \leq j < i \ . \end{aligned}$$

Then we let $\mathbf{w} = \mathbf{u}_{1:i-1}$ and introduce the Lagrange multiplier $\boldsymbol{\lambda} \in \mathbb{R}^{i-1}$ and $\alpha \in \mathbb{R}$. Using strong duality, we have the equivalent problem

$$\min_{\boldsymbol{\lambda} \succeq \mathbf{0}, \alpha \geq 0} \max_{\beta, \mathbf{w}} \ \langle \mathbf{w}, \mathbf{z}_{1:i-1} \rangle + \beta \|\mathbf{z}_{i:p}\|_1 + \langle \boldsymbol{\lambda}, \mathbf{w} - \mathbf{b} \rangle - \alpha((k-i+1)\beta^2 + \|\mathbf{w}\|_2^2 - 1) \ ,$$

where $\mathbf{b} = [\beta, \beta, \ldots, \beta]^T \in \mathbb{R}^{i-1}$. By complementary slackness, we know $\boldsymbol{\lambda} = \mathbf{0}$ for the optimal solution if it is feasible. Taking the gradient of the objective function w.r.t β

and $\mathbf{w}$, we obtain

$$\|\mathbf{z}_{i:p}\|_1 - \sum_i \lambda_i - 2\alpha\beta(k-i+1) = \|\mathbf{z}_{i:p}\|_1 - 2\alpha\beta(k-i+1) = 0 \tag{4.17}$$

$$\mathbf{z}_{1:i-1} + \boldsymbol{\lambda} - 2\alpha\mathbf{w} = \mathbf{z}_{1:i-1} - 2\alpha\mathbf{w} = 0\ . \tag{4.18}$$

It is also not difficult to see that the optimal solution will make the constraint $\|\mathbf{u}_{1:i-1}\|_2^2 \leq 1-(k-i+1)\beta^2$ hold with equality, i.e.,

$$\|\mathbf{w}\|_2^2 = 1-(k-i+1)\beta^2 \tag{4.19}$$

Combining the Equation (4.17) (4.18) (4.19), we solve β and α and $\mathbf{w}$

$$\begin{aligned}
\beta &= \frac{\|\mathbf{z}_{i:p}\|_1}{\sqrt{\|\mathbf{z}_{i:p}\|_1^2(k-i+1) + \|\mathbf{z}_{1:i-1}\|_2^2(k-i+1)^2}}\ , \\
\alpha &= \frac{\|\mathbf{z}_{1:i-1}\|_2}{2\sqrt{1-(k-i+1)\beta^2}}\ , \\
\mathbf{w} &= \frac{\mathbf{z}_{1:i-1}}{2\alpha}\ ,
\end{aligned}$$

which is essentially the Line 3 in Algorithm 6. As we do not know the i beforehand, we have to check every possible $1 \leq i \leq k$ to find the one that achieves the maxima without violating the constraint, which corresponds to the loop and if-then statement in Algorithm 6. Since the optimal $\mathbf{w}$ is proportional to $\mathbf{z}_{1:i-1}$, which is sorted in descending order, we only need to ensure $\beta < w_{i-1}$. ■

Chapter 5

Structure Matrix Recovery via Generalized Dantzig Selector

5.1 Introduction

In Chapter 3 and 4, we have studied the estimation of structured linear models for vector setting. In this Chapter, we extend the results obtained there to the matrix setting, with an emphasis on general structures as well. Structured matrix recovery has found a wide spectrum of applications in real world, e.g., recommender systems [96], face recognition [33], etc. In the context of matrix estimation, the linear model has the form

$$y = \langle\langle \mathbf{\Theta}^*, \mathbf{X} \rangle\rangle + \omega \ , \tag{5.1}$$

where $\mathbf{\Theta}^* \in \mathbb{R}^{d\times p}$ is the *unknown* matrix to be recovered, $\mathbf{X} \in \mathbb{R}^{d\times p}$ is the measurement matrix, y is the response and ω is the additive noise. $\langle\langle \cdot, \cdot \rangle\rangle$ denotes the matrix inner product. Our goal is to recovery the matrix $\mathbf{\Theta}^*$ given n i.i.d. copies of $(\mathbf{X}, y)$, denoted by $\{(\mathbf{X}_i, y_i)\}_{i=1}^n$. In the literature, various types of measurement matrices $\mathbf{X}$

has been investigated, for example, Gaussian ensemble where $\mathbf{X}$ consists of i.i.d. standard Gaussian entries [40], rank-one projection model where $\mathbf{X}$ is randomly generated with constraint $\text{rank}(\mathbf{X}) = 1$ [30]. A special case of rank-one projection is the matrix completion model [31], in which $\mathbf{X}$ has a single entry equal to one with all the rest set to zero, i.e., y takes the value of one entry from $\mathbf{\Theta}^*$ at each measurement. Other measurement models include row-and-column affine measurement [192], exponential family matrix completion [67, 68], and so on.

Like the vector scenario, previous works have shown that low-complexity structure of $\mathbf{\Theta}^*$ can significantly benefit its recovery [40, 127]. For instance, one of the popular structures of $\mathbf{\Theta}^*$ is low-rank, which can be approximated by a small value of trace norm (a.k.a. nuclear norm) $\|\cdot\|_{\text{tr}}$. Under the low-rank assumption of $\mathbf{\Theta}^*$, recovery guarantees have been established for different measurement matrices using convex programs, e.g., trace-norm regularized least-square estimator [35, 68, 127, 141],

$$\min_{\mathbf{\Theta} \in \mathbb{R}^{d \times p}} \quad \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \langle\langle \mathbf{X}_i, \mathbf{\Theta} \rangle\rangle\right)^2 + \lambda_n \|\mathbf{\Theta}^*\|_{\text{tr}} \ , \tag{5.2}$$

and constrained trace-norm minimization estimators [30, 35, 40, 67, 141], such as

$$\min_{\mathbf{\Theta} \in \mathbb{R}^{d \times p}} \|\mathbf{\Theta}\|_{\text{tr}} \quad \text{s.t.} \quad \left\| \sum_{i=1}^{n} \left(\langle\langle \mathbf{X}_i, \mathbf{\Theta} \rangle\rangle - y_i\right) \mathbf{X}_i \right\|_{\text{op}} \leq \lambda_n \ , \tag{5.3}$$

where λ_n is a tuning parameter, and $\|\cdot\|_{\text{op}}$ denotes the operator (spectral) norm. Among the convex approaches, the exact recovery guarantee of a matrix-form basis-pursuit [48] estimator was analyzed for the noiseless setting in [141], under certain matrix-form restricted isometry property (RIP). In the presence of noise, [35] also used matrix RIP to establish the recovery error bound for both regularized and constraint estimators, i.e., (5.2) and (5.3). In [30], a variant of estimator (5.3) was proposed and its recovery guarantee was built on a so-called restricted uniform boundedness

(RUB) condition, which is more suitable for the rank-one projection based measurement model. Despite the fact that the low-rank structure has been well studied, only a few works extend to more general structures. In [127], the regularized estimator (5.2) was generalized by replacing the trace norm with a decomposable norm $\|\cdot\|$ for other structures. [40] extended the estimator in [141] with $\|\cdot\|_{\text{tr}}$ replaced by an atomic norm, but the consistency of the estimator is only available when the noise vector is bounded.

In this work, we first present a general framework for estimation of structured matrices via the generalized Dantzig sector (GDS) [28, 41] as follows

$$\hat{\mathbf{\Theta}} = \underset{\mathbf{\Theta} \in \mathbb{R}^{d \times p}}{\operatorname{argmin}} \quad \|\mathbf{\Theta}\| \quad \text{s.t.} \quad \left\| \sum_{i=1}^{n} \left(\langle\langle \mathbf{X}_i, \Theta \rangle\rangle - y_i \right) \mathbf{X}_i \right\|_* \leq \lambda_n \ , \tag{5.4}$$

in which $\|\cdot\|$ can be any norm and its dual norm is $\|\cdot\|_*$. Computationally the matrix GDS can be solved using the inexact ADMM algorithm proposed in Chapter 3. By assuming sub-Gaussian $\mathbf{X}$ and ω, we can bound the statistical error using the matrix counterpart of the geometric measures in Chapter 3. This result can be extended to heavy tailed measurement and noise, following recent advances [150]. Second, by extending the idea in Chapter 4, we further bound those geometric measures for the structures captured by the class of *unitarily invariant* norms, which include the widely-used trace norm, spectral norm and Frobenius norm. We also illustrate concrete versions of the bounds using the trace norm and the recently proposed spectral k-support norm [121].

The rest of the chapter is organized as follows: we first provide the deterministic analysis in Section 5.2. In Section 5.3, we present the probabilistic analysis for sub-Gaussian measurement matrices and noise, along with the general bounds of the geometric measures for unitarily invariant norms. Section 5.4 is dedicated to the examples for the application of general bounds. Throughout the chapter, the symbols

c, C, c_0, C_0, etc., are reserved for universal constants, which may be different at each occurrence. We also introduce some notations before proceeding with the analysis. We denote by $\boldsymbol{\sigma}(\boldsymbol{\Theta}) \in \mathbb{R}^d$ the vector of singular values (sorted in descending order) of matrix $\boldsymbol{\Theta} \in \mathbb{R}^{d\times p}$, and may use the shorthand $\boldsymbol{\sigma}^*$ for $\boldsymbol{\sigma}(\boldsymbol{\Theta}^*)$. For any $\boldsymbol{\theta} \in \mathbb{R}^d$, we define the corresponding $|\boldsymbol{\theta}|^{\downarrow}$ by arranging the absolute values of elements of $\boldsymbol{\theta}$ in descending order. Given any matrix $\boldsymbol{\Theta} \in \mathbb{R}^{d\times p}$ and subspace $\mathcal{M} \subseteq \mathbb{R}^{d\times p}$, we denote by $\boldsymbol{\Theta}_{\mathcal{M}}$ the orthogonal projection of $\boldsymbol{\Theta}$ onto $\mathcal{M}$. Besides we let colsp($\boldsymbol{\Theta}$) (rowsp($\boldsymbol{\Theta}$)) be the subspace spanned by columns (rows) of $\boldsymbol{\Theta}$. The notation $\mathbb{S}^{dp-1}$ represents the unit sphere of $\mathbb{R}^{d\times p}$, i.e., the set $\{\boldsymbol{\Theta} | \|\boldsymbol{\Theta}\|_F = 1\}$. The unit ball of norm $\|\cdot\|$ is denoted by $\Omega = \{\boldsymbol{\Theta} \mid \|\boldsymbol{\Theta}\| \leq 1\}$.

5.2 Deterministic Analysis

5.2.1 Deterministic Error Bound

To evaluate the performance of GDS (5.4), we focus on the Frobenius-norm error, i.e., $\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F$. Throughout the analysis, w.l.o.g. we assume that $d \leq p$. In the following theorem, we provide a deterministic bound for $\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F$ under some standard assumptions on λ_n and $\mathbf{X}$.

Theorem 13 *Define the error cone*

$$\mathcal{T} = \operatorname{cone}\{\ \boldsymbol{\Delta} \in \mathbb{R}^{d\times p} \mid \|\boldsymbol{\Delta} + \boldsymbol{\Theta}^*\| \leq \|\boldsymbol{\Theta}^*\|\}\ . \tag{5.5}$$

Assume that

$$\lambda_n \geq \left\| \sum_{i=1}^{n} \omega_i \mathbf{X}_i \right\|_* , \tag{5.6}$$

$$\sum_{i=1}^{n} \frac{\langle\langle \mathbf{X}_i, \boldsymbol{\Delta}\rangle\rangle^2}{\|\boldsymbol{\Delta}\|_F^2} \geq \alpha > 0, \quad \forall\ \boldsymbol{\Delta} \in \mathcal{T}\ . \tag{5.7}$$

Then the estimation $\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F$ *error satisfies*

$$\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F \leq \frac{2\Psi \cdot \lambda_n}{\alpha} \ , \tag{5.8}$$

where Ψ *is the restricted norm compatibility defined as*

$$\Psi = \sup_{\mathbf{\Delta} \in \mathcal{T}} \frac{\|\mathbf{\Delta}\|}{\|\mathbf{\Delta}\|_F} \ . \tag{5.9}$$

Proof: Since λ_n satisfies the condition (5.6) and $\omega_i = y_i - \langle\langle \mathbf{X}_i, \mathbf{\Theta}^* \rangle\rangle$, we have

$$\left\| \sum_{i=1}^{n} (\langle\langle \mathbf{X}_i, \mathbf{\Theta}^* \rangle\rangle - y_i)\, \mathbf{X}_i \right\|_* \leq \lambda_n \ ,$$

which indicates that the constraint set in (5.4) is feasible, thus

$$\left\| \sum_{i=1}^{n} \left(\langle\langle \mathbf{X}_i, \hat{\mathbf{\Theta}} \rangle\rangle - y_i\right) \mathbf{X}_i \right\|_* \leq \lambda_n \ .$$

Using triangular inequality, one has

$$\left\| \sum_{i=1}^{n} \langle\langle \mathbf{X}_i, \hat{\mathbf{\Theta}} - \mathbf{\Theta}^* \rangle\rangle \cdot \mathbf{X}_i \right\|_* \leq 2\lambda_n \ .$$

Denote $\hat{\mathbf{\Theta}} - \mathbf{\Theta}^*$ by $\mathbf{\Delta}$, and by the definition of dual norm, we get

$$\begin{aligned} \sum_{i=1}^{n} \langle\langle \mathbf{X}_i, \mathbf{\Delta} \rangle\rangle^2 &= \left\langle\left\langle \mathbf{\Delta}, \sum_{i=1}^{n} \langle\langle \mathbf{X}_i, \mathbf{\Delta} \rangle\rangle \cdot \mathbf{X}_i \right\rangle\right\rangle \\ &\leq \|\mathbf{\Delta}\| \cdot \left\| \sum_{i=1}^{n} \langle\langle \mathbf{X}_i, \hat{\mathbf{\Theta}} - \mathbf{\Theta}^* \rangle\rangle \cdot \mathbf{X}_i \right\|_* \leq 2\lambda_n \|\mathbf{\Delta}\| \ . \end{aligned}$$

On the other hand, the objective function in (5.4) implies that $\|\hat{\mathbf{\Theta}}\| \leq \|\mathbf{\Theta}^*\|$. Therefore the error vector $\mathbf{\Delta}$ must belong to the set $\mathcal{T}$. Using condition (5.7), we obtain

$$\alpha\|\mathbf{\Delta}\|_F^2 \leq \sum_{i=1}^{n} \langle\langle \mathbf{X}_i, \mathbf{\Delta}\rangle\rangle^2 \leq 2\lambda_n \|\mathbf{\Delta}\| \qquad \Longrightarrow$$
$$\|\mathbf{\Delta}\|_F \leq \frac{2\lambda_n}{\alpha}\frac{\|\mathbf{\Delta}\|}{\|\mathbf{\Delta}\|_F} \leq \frac{2\Psi \cdot \lambda_n}{\alpha} \ ,$$

which complete the proof. ■

In this work, we are particularly interested in the norm $\|\cdot\|$ from the class of *unitarily invariant* matrix norm, which essentially satisfies the following property, $\|\mathbf{\Theta}\| = \|\mathbf{U}\mathbf{\Theta}\mathbf{V}\|$ for any $\mathbf{\Theta} \in \mathbb{R}^{d\times p}$ and unitary matrices $\mathbf{U} \in \mathbb{R}^{d\times d}$, $\mathbf{V} \in \mathbb{R}^{p\times p}$. A useful result for such norms is given in Lemma 10 (see [19, 102] for details).

Lemma 10 *Suppose that the singular values of a matrix $\mathbf{\Theta} \in \mathbb{R}^{d\times p}$ are given by $\boldsymbol{\sigma} = [\sigma_1, \sigma_2, \ldots, \sigma_d]^T$. A unitarily invariant norm $\|\cdot\| : \mathbb{R}^{d\times p} \mapsto \mathbb{R}$ can be characterized by some symmetric gauge function*[1] *$f : \mathbb{R}^d \mapsto \mathbb{R}$ as $\|\mathbf{\Theta}\| = f(\boldsymbol{\sigma})$, and its dual norm is given by $\|\mathbf{\Theta}\| = f_*(\boldsymbol{\sigma})$, in which f_* is the dual norm of f.*

As the sparsity of $\boldsymbol{\sigma}$ equals the rank of $\mathbf{\Theta}$, the class of unitarily invariant matrix norms is useful in structured low-rank matrix recovery and includes many widely used norms, e.g., trace norm with $f(\cdot) = \|\cdot\|_1$, Frobenius norm with $f(\cdot) = \|\cdot\|_2$, Schatten p-norm with $f(\cdot) = \|\cdot\|_p$, Ky Fan k-norm when $f(\cdot)$ is the L_1 norm of the largest k elements in magnitude, etc.

In the rest of our analysis, we will frequently use the so-called ordered weighted L_1 (OWL) norm for $\mathbb{R}^d$ [55], which is defined as $\|\boldsymbol{\theta}\|_{\mathbf{w}} \triangleq \langle|\boldsymbol{\theta}|^{\downarrow}, |\mathbf{w}|^{\downarrow}\rangle$, where $\mathbf{w} \in \mathbb{R}^d$ is a predefined weight vector. Noting that the OWL norm is a symmetric gauge, we define

[1] Symmetric gauge function is a norm that is invariant under sign-changes and permutations of the elements.

the *spectral OWL norm* for $\boldsymbol{\Theta}$ as: $\|\boldsymbol{\Theta}\|_{\mathbf{w}} \triangleq \|\boldsymbol{\sigma}(\boldsymbol{\Theta})\|_{\mathbf{w}}$, i.e., applying the OWL norm on $\boldsymbol{\sigma}(\boldsymbol{\Theta})$.

5.2.2 Bounding Restricted Norm Compatibility

Given the definition of restricted norm compatibility in Theorem 13, Ψ involves no randomness and purely depends on the $\|\cdot\|$ and the geometry of $\mathcal{T}$. Hence we directly work on its upper bound for unitarily invariant norms. In general, characterizing the error cone $\mathcal{T}$ is difficult, especially for non-decomposable norm. To address the issue, we first define the seminorm below.

Definition 16 (subspace spectral OWL seminorm) Given two orthogonal subspaces $\mathcal{M}_1, \mathcal{M}_2 \subseteq \mathbb{R}^{d\times p}$ and two vectors $\mathbf{w}, \mathbf{z} \in \mathbb{R}^d$, the *subspace spectral OWL seminorm* for $\mathbb{R}^{d\times p}$ is defined as

$$\|\boldsymbol{\Theta}\|_{\mathbf{w},\mathbf{z}} \triangleq \|\boldsymbol{\Theta}_{\mathcal{M}_1}\|_{\mathbf{w}} + \|\boldsymbol{\Theta}_{\mathcal{M}_2}\|_{\mathbf{z}} \ , \tag{5.10}$$

where $\boldsymbol{\Theta}_{\mathcal{M}_1}$ and $\boldsymbol{\Theta}_{\mathcal{M}_2}$ are the orthogonal projections of $\boldsymbol{\Theta}$ onto $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively.

Next we will construct such a seminorm based on a subgradient $\boldsymbol{\theta}^*$ of the symmetric gauge f associated with $\|\cdot\|$ at $\boldsymbol{\sigma}^*$, which can be obtained by solving the *polar operator* [187]

$$\boldsymbol{\theta}^* \in \underset{\mathbf{x}: f_*(\mathbf{x})\leq 1}{\operatorname{argmax}} \ \langle \mathbf{x}, \boldsymbol{\sigma}^* \rangle \ . \tag{5.11}$$

Given that $\boldsymbol{\sigma}^*$ is sorted, w.l.o.g. we may assume that $\boldsymbol{\theta}^*$ is nonnegative and sorted because $\langle \boldsymbol{\sigma}^*, \boldsymbol{\theta}^* \rangle \leq \langle \boldsymbol{\sigma}^*, |\boldsymbol{\theta}^*|^{\downarrow} \rangle$ and $f_*(\boldsymbol{\theta}^*) = f_*(|\boldsymbol{\theta}^*|^{\downarrow})$. Also, we denote by $\theta^*_{\max}$ ($\theta^*_{\min}$) the largest (smallest) element of the $\boldsymbol{\theta}^*$, and define $\rho = \theta^*_{\max}/\theta^*_{\min}$ (if $\theta^*_{\min} = 0$, we define $\rho = +\infty$). As shown in the lemma below, a constructed seminorm based on $\boldsymbol{\theta}^*$ will induce a set $\mathcal{T}'$ that contains $\mathcal{T}$ and is considerably easier to work with.

Lemma 11 *Assume that* $\text{rank}(\mathbf{\Theta}^*) = r$ *and its compact SVD is given by* $\mathbf{\Theta}^* = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, *where* $\mathbf{U} \in \mathbb{R}^{d\times r}$, $\mathbf{\Sigma} \in \mathbb{R}^{r\times r}$ *and* $\mathbf{V} \in \mathbb{R}^{p\times r}$. *Let* $\boldsymbol{\theta}^*$ *be any subgradient of* $f(\boldsymbol{\sigma}^*)$, $\mathbf{w} = [\theta_1^*, \theta_2^*, \dots, \theta_r^*, 0, \dots, 0]^T \in \mathbb{R}^d$, $\mathbf{z} = [\theta_{r+1}^*, \theta_{r+2}^*, \dots, \theta_d^*, 0, \dots, 0]^T \in \mathbb{R}^d$, $\mathcal{U} = \text{colsp}(\mathbf{U})$ *and* $\mathcal{V} = \text{rowsp}(\mathbf{V}^T)$, *and define* $\mathcal{M}_1$, $\mathcal{M}_2$ *as*

$$\mathcal{M}_1 = \{\mathbf{\Theta} \mid \text{colsp}(\mathbf{\Theta}) \subseteq \mathbf{U}, \text{rowsp}(\mathbf{\Theta}) \subseteq \mathbf{V}\} \ , \tag{5.12}$$

$$\mathcal{M}_2 = \{\mathbf{\Theta} \mid \text{colsp}(\mathbf{\Theta}) \subseteq \mathbf{U}^\perp, \text{rowsp}(\mathbf{\Theta}) \subseteq \mathbf{V}^\perp\} \ , \tag{5.13}$$

where $\mathbf{U}^\perp$, $\mathbf{V}^\perp$ *are orthogonal complements of* $\mathbf{U}$ *and* $\mathbf{V}$ *respectively. Then the specified subspace spectral OWL seminorm* $\|\cdot\|_{\mathbf{w},\mathbf{z}}$ *satisfies*

$$\mathcal{T} \subseteq \mathcal{T}' \triangleq \text{cone}\{\mathbf{\Delta} \mid \|\mathbf{\Delta} + \mathbf{\Theta}^*\|_{\mathbf{w},\mathbf{z}} \le \|\mathbf{\Theta}^*\|_{\mathbf{w},\mathbf{z}}\} \tag{5.14}$$

The proof is given in the appendix. Base on the superset $\mathcal{T}'$, we are able to bound the restricted norm compatibility for unitarily invariant norms by the following theorem.

Theorem 14 *Assume there exist* η_1 *and* η_2 *such that the symmetric gauge* f *for* $\|\cdot\|$ *satisfies*

$$f(\boldsymbol{\delta}) \le \max\{\eta_1\|\boldsymbol{\delta}\|_1, \ \eta_2\|\boldsymbol{\delta}\|_2\} \quad \textit{for any } \boldsymbol{\delta} \in \mathbb{R}^d. \tag{5.15}$$

Then given a rank-r $\mathbf{\Theta}^*$, *the restricted norm compatibility* Ψ *is upper bounded by*

$$\Psi \le 2\Phi_f(r) + \max\left\{\eta_2, \eta_1(1+\rho)\sqrt{r}\right\} \ , \tag{5.16}$$

where $\rho = \frac{\theta^*_{\max}}{\theta^*_{\min}}$, *and* $\Phi_f(r) = \sup_{\|\boldsymbol{\delta}\|_0 \le r} \frac{f(\boldsymbol{\delta})}{\|\boldsymbol{\delta}\|_2}$ *is called sparse norm compatibility.*

Remark: The assumption for Theorem 14 might seem cumbersome at the first glance, but the different combinations of η_1 and η_2 give us more flexibility. In fact, it trivially covers two cases, $\eta_2 = 0$ along with $f(\boldsymbol{\delta}) \le \eta_1\|\boldsymbol{\delta}\|_1$ for any $\boldsymbol{\delta}$, and the other way around,

$\eta_1 = 0$ along with $f(\boldsymbol{\delta}) \leq \eta_2 \|\boldsymbol{\delta}\|_2$.

5.3 Probabilistic Analysis

For the probabilistic analysis, we assume that the measurement matrix $\mathbf{X}$ is sub-Gaussian with $|||\mathbf{X}|||_{\psi_2} \leq \kappa$ for a constant κ, i.e.,

$$|||\langle\langle \mathbf{X}, \mathbf{Z} \rangle\rangle|||_{\psi_2} \leq \kappa \ \text{ for any } \mathbf{Z} \in \mathbb{S}^{dp-1} \ .$$

The noise ω is also assumed to be sub-Gaussian with $\|\omega\|_{\psi_2} \leq \tau$ for a constant τ.

5.3.1 Bounding Restricted Convexity α

The RE condition in (5.7) is equivalent to

$$\sum_{i=1}^{n} \langle\langle \mathbf{X}_i, \boldsymbol{\Delta} \rangle\rangle^2 \geq \alpha > 0, \ \forall \ \boldsymbol{\Delta} \in \mathcal{T} \cap \mathbb{S}^{dp-1} \ . \tag{5.17}$$

In the following theorem, we express the restricted convexity α in terms of Gaussian width.

Theorem 15 *Assume that* $\mathbf{X}$ *is a centered isotropic sub-Gaussian random matrix* $\mathbf{X}$ *with* $|||\mathbf{X}|||_{\psi_2} \leq \kappa$*, and let the error spherical cap be*

$$\mathcal{C} = \mathcal{T} \cap \mathbb{S}^{dp-1} \ . \tag{5.18}$$

With probability at least $1 - \exp(-\zeta w^2(\mathcal{C}))$*, the following inequality holds with absolute constant* ζ *and* ξ*,*

$$\inf_{\boldsymbol{\Delta} \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} \langle\langle X_i, \boldsymbol{\Delta} \rangle\rangle^2 \ \geq \ 1 - \xi\kappa^2 \cdot \frac{w(\mathcal{C})}{\sqrt{n}} \ . \tag{5.19}$$

The bound (5.19) involves the Gaussian width of error spherical cap $\mathcal{C}$, i.e., the error cone intersecting with unit sphere. For unitarily invariant R, the theorem below provides a general way to bound $w(\mathcal{C})$.

Theorem 16 *Under the setting of Lemma 11, let* $\rho = \frac{\theta^*_{\max}}{\theta^*_{\min}}$ *and* $\mathrm{rank}(\boldsymbol{\Theta}^*) = r$*. The Gaussian width* $w(\mathcal{C})$ *satisfies*

$$w(\mathcal{C}) \leq \min\left\{\sqrt{dp}, \sqrt{(2\rho^2+1)\,(d+p-r)\,r}\right\} \ . \tag{5.20}$$

The proof of Theorem 16 is included in the appendix, which relies on a few specific properties of Gaussian random matrix [4, 40].

5.3.2 Bounding Regularization Parameter λ_n

In view of Theorem 13, we should choose the λ_n large enough to satisfy the condition in (5.6). Hence we need an upper bound for random quantity $\|\sum_{i=1}^n \omega_i \mathbf{X}_i\|_*$.

Theorem 17 *Assume that* $\mathbf{X}$ *is a centered isotropic sub-Gaussian random matrix* $\mathbf{X}$ *with* $|||\mathbf{X}|||_{\psi_2} \leq \kappa$*, and the noise* ω *is sub-Gaussian* $|||\omega|||_{\psi_2} \leq \tau$*. Let* Ω *be the unit ball of* $\|\cdot\|$ *and* $\eta = \sup_{\boldsymbol{\Delta}\in\Omega} \|\boldsymbol{\Delta}\|_F$*. With probability at least* $1 - \exp(-c_1 n) - c_2 \exp\left(-\frac{w^2(\Omega)}{c_3^2\eta^2}\right)$*, the following inequality holds*

$$\left\|\sum_{i=1}^n \omega_i \mathbf{X}_i\right\|_* \leq c_0 \kappa\tau \cdot \sqrt{n} w(\Omega) \ . \tag{5.21}$$

The theorem above shows that the lower bound of λ_n depends on the Gaussian width of the unit ball of $\|\cdot\|$. Next we give its general bound for the unitarily invariant matrix norm.

Theorem 18 *Suppose that the symmetric gauge f associated with $\|\cdot\|$ satisfies $f(\cdot) \geq \nu\|\cdot\|_1$. Then the Gaussian width $w(\Omega)$ is upper bounded by*

$$w(\Omega) \leq \frac{\sqrt{d} + \sqrt{p}}{\nu} \ . \tag{5.22}$$

Proof: As $f(\cdot) \geq \nu\|\cdot\|_1$, we have

$$\|\cdot\| \geq \nu\|\cdot\|_{\text{tr}} \qquad \Longrightarrow \qquad \Omega \subseteq \Omega_{\nu\|\cdot\|_{\text{tr}}} \ .$$

Hence it follows that

$$w\left(\Omega\right) \leq \frac{w\left(\Omega_{\text{tr}}\right)}{\nu} = \frac{\mathbb{E}\|\mathbf{G}\|_{\text{op}}}{\nu} \leq \frac{\sqrt{d} + \sqrt{p}}{\nu} \ ,$$

5.4 Examples

Combining results in Section 5.3, we have that if the number of measurements $n > O(w^2(\mathcal{C}))$, then the recovery error, with high probability, satisfies

$$\left\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}^*\right\|_F \leq O\left(\frac{\Psi \cdot w(\Omega)}{\sqrt{n}}\right) \ . \tag{5.23}$$

Here we give two examples based on the trace norm [35] and the recently proposed spectral k-support norm [121] to illustrate how to bound the geometric measures and obtain the error bound.

5.4.1 Trace Norm

Trace norm has been widely used in low-rank matrix recovery. The trace norm of $\mathbf{\Theta}^*$ is basically the L_1 norm of $\boldsymbol{\sigma}^*$, i.e., $f = \|\cdot\|_1$. Now we turn to the three geometric measures. Assuming that $\text{rank}(\mathbf{\Theta}^*) = r \ll d$, one subgradient of $\|\boldsymbol{\sigma}^*\|_1$ is $\boldsymbol{\theta}^* = [1, 1, \ldots, 1]^T$.

Restricted norm compatibility Ψ_{tr}: It is obvious that assumption in Theorem 14 will hold for f by choosing $\eta_1 = 1$ and $\eta_2 = 0$, and we have $\rho = 1$. The sparse compatibility constant $\Phi_{L_1}(r)$ is $\sqrt{r}$ because $\|\boldsymbol{\delta}\|_1 \leq \sqrt{r}\|\boldsymbol{\delta}\|_2$ for any r-sparse $\boldsymbol{\delta}$. Using Theorem 14, we have

$$\Psi_{\text{tr}} \leq 4\sqrt{r} \ . \tag{5.24}$$

Gaussian width $w(\mathcal{C}_{\text{tr}})$: As $\rho = 1$, Theorem 16 implies that

$$w\left(\mathcal{C}_{\text{tr}}\right) \leq \sqrt{3r(d+p-r)} \ . \tag{5.25}$$

Gaussian width $w(\Omega_{\text{tr}})$: Using Theorem 18 with $\nu = 1$, it is easy to see that

$$w\left(\Omega_{\text{tr}}\right) \leq \sqrt{d} + \sqrt{p} \ . \tag{5.26}$$

Putting all the results together, we have the following bound hold with high probability when $n > O(r(d+p-r))$

$$\left\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\right\|_F \leq O\left(\sqrt{\frac{rd}{n}} + \sqrt{\frac{rp}{n}}\right) \ , \tag{5.27}$$

which matches the bound in [31].

5.4.2 Spectral k-Support Norm

The k-support norm proposed in [7] is defined as

$$\|\boldsymbol{\theta}\|_k^{sp} \triangleq \inf\left\{\sum_i \|\mathbf{u}_i\|_2 \;\middle|\; \|\mathbf{u}_i\|_0 \leq k,\ \sum_i \mathbf{u}_i = \boldsymbol{\theta}\right\} \ , \tag{5.28}$$

and its dual norm is simply given by

$$\|\boldsymbol{\theta}\|_{k*}^{sp} = \left\| |\boldsymbol{\theta}|_{1:k}^{\downarrow} \right\|_2 \ . \tag{5.29}$$

Spectral k-support norm (denoted by $\|\cdot\|_{\mathrm{sk}}$) of $\boldsymbol{\Theta}^*$ is defined by applying the k-support norm on $\boldsymbol{\sigma}^*$, i.e., $f = \|\cdot\|_k^{sp}$, which has demonstrated better performance than trace norm in matrix completion task [121]. For simplicity, We assume that $\mathrm{rank}(\boldsymbol{\Theta}^*) = r = k$ and $\|\boldsymbol{\sigma}^*\|_2 = 1$. One subgradient of $\|\boldsymbol{\sigma}^*\|_k^{sp}$ is $\boldsymbol{\theta}^* = [\sigma_1^*, \sigma_2^*, \ldots, \sigma_r^*, \sigma_r^*, \ldots, \sigma_r^*]^T$.

Restricted norm compatibility Ψ_{sk}: The following relation has been shown for k-support norm in [7],

$$\max\left\{\|\cdot\|_2, \frac{\|\cdot\|_1}{\sqrt{k}}\right\} \le \|\cdot\|_k^{sp} \le \sqrt{2}\max\{\|\cdot\|_2, \frac{\|\cdot\|_1}{\sqrt{k}}\} \ . \tag{5.30}$$

Hence the assumption in Theorem 14 will hold for $\eta_1 = \sqrt{\frac{2}{k}}$ and $\eta_2 = \sqrt{2}$, and we have $\rho = \frac{\sigma_1^*}{\sigma_r^*}$. The sparse compatibility constant $\Phi_k^{sp}(r) = \Phi_k^{sp}(k) = 1$ because $\|\boldsymbol{\delta}\|_k^{sp} = \|\boldsymbol{\delta}\|_2$ for any k-sparse $\boldsymbol{\delta}$. Using Theorem 14, we have

$$\Psi_{\mathrm{sk}} \le 2\sqrt{2} + \sqrt{2}\left(1 + \frac{\sigma_1^*}{\sigma_r^*}\right) = \sqrt{2}\left(3 + \frac{\sigma_1^*}{\sigma_r^*}\right) . \tag{5.31}$$

Gaussian width $w(\mathcal{C}_{\mathrm{sk}})$: Theorem 16 implies

$$w\left(\mathcal{C}_{\mathrm{sk}}\right) \le \sqrt{r(d+p-r)\left[\frac{2\sigma_1^{*2}}{\sigma_r^{*2}} + 1\right]} \ . \tag{5.32}$$

Gaussian width $w(\Omega_{\mathrm{sk}})$: The relation above for k-support norm shown in [7] also implies that $\nu = \frac{1}{\sqrt{k}} = \frac{1}{\sqrt{r}}$. By Theorem 18, we get

$$w\left(\Omega_{\mathrm{sk}}\right) \le \sqrt{r}(\sqrt{d} + \sqrt{p}) \ . \tag{5.33}$$

Given the upper bounds for geometric measures, with high probability, we have

$$\left\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}^*\right\|_F \leq O\left(\sqrt{\frac{rd}{n}} + \sqrt{\frac{rp}{n}}\right) \tag{5.34}$$

when $n > O(r(d+p-r))$. The spectral k-support norm was first introduced in [121], in which no statistical results are provided. Although [67] investigated the statistical aspects of spectral k-support norm in matrix completion setting, the analysis was quite different from our setting. Hence this error bound is new in the literature.

Appendix

Appendix 5.A Proof of Deterministic Analysis

5.A.1 Proof of Lemma 11

Proof: Both $\mathcal{T}$ and $\mathcal{T}'$ are induced by scaled (semi)norm balls (i.e., Ω and $\Omega_{\mathbf{w},\mathbf{z}}$) centered at $-\mathbf{\Theta}^*$, and note that

$$\mathbf{\Theta}^*_{\mathcal{M}_1} = \mathbf{\Theta}^* \ , \quad \mathbf{\Theta}^*_{\mathcal{M}_2} = 0 \ .$$

Thus we obtain

$$\|\mathbf{\Theta}^*\|_{w,z} = \|\mathbf{\Theta}^*_{\mathcal{M}_1}\|_w = \sum_{i=1}^{r} \sigma_i^* \theta_i^* = \langle \boldsymbol{\sigma}^*, \boldsymbol{\theta}^* \rangle = \|\mathbf{\Theta}^*\| \ ,$$

which indicates that the two balls have the same radius. Hence we only need to show that $\|\cdot\|_{\mathbf{w},\mathbf{z}} \leq \|\cdot\|$. For any $\mathbf{\Delta} \in \mathbb{R}^{d\times p}$, assume that the SVD of $\mathbf{\Delta}_{\mathcal{M}_1}$ and $\mathbf{\Delta}_{\mathcal{M}_2}$ are given by $\mathbf{\Delta}_{\mathcal{M}_1} = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^T$ and $\mathbf{\Delta}_{\mathcal{M}_2} = \mathbf{U}_2\mathbf{\Sigma}_2\mathbf{V}_2^T$. The corresponding vectors of singular values are in the form of $\boldsymbol{\sigma}' = [\sigma_1', \sigma_2', \ldots, \sigma_r', 0, \ldots, 0]^T, \boldsymbol{\sigma}'' = [\sigma_1'', \sigma_2'', \ldots, \sigma_{d-r}'', 0, \ldots, 0]^T \in \mathbb{R}^d$, as $\operatorname{rank}(\boldsymbol{\Delta}_{\mathcal{M}_1}) \leq r$ and $\operatorname{rank}(\boldsymbol{\Delta}_{\mathcal{M}_2}) \leq d - r$. Then we have

$$\|\boldsymbol{\Delta}\|_{\mathbf{w},\mathbf{z}} = \|\boldsymbol{\Delta}_{\mathcal{M}_1}\|_{\mathbf{w}} + \|\boldsymbol{\Delta}_{\mathcal{M}_2}\|_{\mathbf{z}} = \langle \boldsymbol{\sigma}', \mathbf{w} \rangle + \langle \boldsymbol{\sigma}'', \mathbf{z} \rangle = \left\langle \boldsymbol{\theta}^*, \begin{bmatrix} \boldsymbol{\sigma}'_{1:r} \\ \boldsymbol{\sigma}''_{1:d-r} \end{bmatrix} \right\rangle = \langle\langle \boldsymbol{\Theta}, \boldsymbol{\Delta} \rangle\rangle \ ,$$

where $\boldsymbol{\Theta} = \mathbf{U}_1 \operatorname{Diag}(\boldsymbol{\theta}^*_{1:r})\mathbf{V}_1 + \mathbf{U}_2 \operatorname{Diag}(\boldsymbol{\theta}^*_{r+1:n})\mathbf{V}_2$. From this construction, we can see that $\boldsymbol{\theta}^*$ are the singular values of $\boldsymbol{\Theta}$, thus $\|\boldsymbol{\Theta}\|_* \leq 1$. It follows that

$$\langle\langle \boldsymbol{\Theta}, \boldsymbol{\Delta} \rangle\rangle \leq \max_{\|\mathbf{Z}\|_* \leq 1} \langle\langle \mathbf{Z}, \boldsymbol{\Delta} \rangle\rangle = \|\boldsymbol{\Delta}\| \ ,$$

which completes the proof. ■

5.A.2 Proof of Theorem 14

Proof: Under the setting of Lemma 11, as $\boldsymbol{\Theta}^* \in \mathcal{M}_1$, we have

$$\begin{aligned} \|\boldsymbol{\Delta} + \boldsymbol{\Theta}^*\|_{\mathbf{w},\mathbf{z}} \leq \|\boldsymbol{\Theta}^*\|_{\mathbf{w},\mathbf{z}} \quad &\Longrightarrow \quad \|\boldsymbol{\Delta}_{\mathcal{M}_1} + \boldsymbol{\Theta}^*\|_{\mathbf{w}} + \|\boldsymbol{\Delta}_{\mathcal{M}_2}\|_{\mathbf{z}} \leq \|\boldsymbol{\Theta}^*\|_{\mathbf{w}} \Longrightarrow \\ -\|\boldsymbol{\Delta}_{\mathcal{M}_1}\|_{\mathbf{w}} + \|\boldsymbol{\Theta}^*\|_{\mathbf{w}} + \|\boldsymbol{\Delta}_{\mathcal{M}_2}\|_{\mathbf{z}} \leq \|\boldsymbol{\Theta}^*\|_{\mathbf{w}} \quad &\Longrightarrow \quad \|\boldsymbol{\Delta}_{\mathcal{M}_2}\|_{\mathbf{z}} \leq \|\boldsymbol{\Delta}_{\mathcal{M}_1}\|_{\mathbf{w}} \ . \end{aligned}$$

As the set $\{\boldsymbol{\Delta} \mid \|\boldsymbol{\Delta}_{\mathcal{M}_2}\|_{\mathbf{z}} \leq \|\boldsymbol{\Delta}_{\mathcal{M}_1}\|_{\mathbf{w}}\}$ itself is a cone, we obtain

$$\mathcal{T}' \ \subseteq \ \{\boldsymbol{\Delta} \mid \|\boldsymbol{\Delta}_{\mathcal{M}_2}\|_{\mathbf{z}} \leq \|\boldsymbol{\Delta}_{\mathcal{M}_1}\|_{\mathbf{w}}\}$$

Define $\mathcal{M}^\perp$ as the orthogonal complement of $\mathcal{M}_1 \oplus \mathcal{M}_2$. By the definition and Lemma 11, we have

$$
\begin{aligned}
\Psi = \sup_{\boldsymbol{\Delta}\in\mathcal{T}} \frac{\|\boldsymbol{\Delta}\|}{\|\boldsymbol{\Delta}\|_F} \leq \sup_{\boldsymbol{\Delta}\in\mathcal{T}'} \frac{\|\boldsymbol{\Delta}\|}{\|\boldsymbol{\Delta}\|_F} &\leq \sup_{\|\boldsymbol{\Delta}_{\mathcal{M}_2}\|_{\mathbf{z}}\leq\|\boldsymbol{\Delta}_{\mathcal{M}_1}\|_{\mathbf{w}}} \frac{\|\boldsymbol{\Delta}\|}{\|\boldsymbol{\Delta}\|_F} \\
&\leq \sup_{\|\boldsymbol{\Delta}_{\mathcal{M}_2}\|_{\mathbf{z}}\leq\|\boldsymbol{\Delta}_{\mathcal{M}_1}\|_{\mathbf{w}}} \frac{\|\boldsymbol{\Delta}_{\mathcal{M}^\perp}\| + \|\boldsymbol{\Delta}_{\mathcal{M}_1} + \boldsymbol{\Delta}_{\mathcal{M}_2}\|}{\|\boldsymbol{\Delta}\|_F} \\
&\leq \sup_{\boldsymbol{\Delta}\in\mathcal{M}^\perp} \frac{\|\boldsymbol{\Delta}\|}{\|\boldsymbol{\Delta}\|_F} + \sup_{\frac{\|\boldsymbol{\Delta}_{\mathcal{M}_2}\|_{\mathrm{tr}}}{\|\boldsymbol{\Delta}_{\mathcal{M}_1}\|_{\mathrm{tr}}}\leq\rho} \frac{\|\boldsymbol{\Delta}_{\mathcal{M}_1} + \boldsymbol{\Delta}_{\mathcal{M}_2}\|}{\|\boldsymbol{\Delta}\|_F}
\end{aligned}
$$

It is not difficult to see that any $\boldsymbol{\Delta} \in \mathcal{M}^\perp$ has rank at most $2r$, thus

$$
\sup_{\boldsymbol{\Delta}\in\mathcal{M}^\perp} \frac{\|\boldsymbol{\Delta}\|}{\|\boldsymbol{\Delta}\|_F} = \sup_{\boldsymbol{\Delta}\in\mathcal{M}^\perp} \frac{f(\boldsymbol{\sigma}(\boldsymbol{\Delta}))}{\|\boldsymbol{\sigma}(\boldsymbol{\Delta})\|_2} \leq \sup_{\|\boldsymbol{\delta}\|_0\leq 2r} \frac{f(\boldsymbol{\delta})}{\|\boldsymbol{\delta}\|_2} \leq 2 \sup_{\|\boldsymbol{\delta}\|_0\leq r} \frac{f(\boldsymbol{\delta})}{\|\boldsymbol{\delta}\|_2} = 2\Phi_f(r) \ .
$$

Using (5.15) and $\|\boldsymbol{\Delta}_{\mathcal{M}_1} + \boldsymbol{\Delta}_{\mathcal{M}_2}\|_F \leq \|\boldsymbol{\Delta}\|_F$, we have

$$
\begin{aligned}
\sup_{\frac{\|\boldsymbol{\Delta}_{\mathcal{M}_2}\|_{\mathrm{tr}}}{\|\boldsymbol{\Delta}_{\mathcal{M}_1}\|_{\mathrm{tr}}}\leq\rho} \frac{\|\boldsymbol{\Delta}_{\mathcal{M}_1} + \boldsymbol{\Delta}_{\mathcal{M}_2}\|}{\|\boldsymbol{\Delta}\|_F} &\leq \sup_{\frac{\|\boldsymbol{\Delta}_{\mathcal{M}_2}\|_{\mathrm{tr}}}{\|\boldsymbol{\Delta}_{\mathcal{M}_1}\|_{\mathrm{tr}}}\leq\rho} \frac{\max\left\{\eta_2\|\boldsymbol{\Delta}\|_F,\ \eta_1\|\boldsymbol{\Delta}_{\mathcal{M}_1} + \boldsymbol{\Delta}_{\mathcal{M}_2}\|_{\mathrm{tr}}\right\}}{\|\boldsymbol{\Delta}\|_F} \\
&\leq \max\left\{\eta_2,\ \sup_{\boldsymbol{\Delta}\in\mathcal{M}_1} \frac{\eta_1(1+\rho)\|\boldsymbol{\Delta}\|_{\mathrm{tr}}}{\|\boldsymbol{\Delta}\|_F}\right\} \\
&\leq \max\left\{\eta_2, \eta_1(1+\rho)\sqrt{r}\right\} \ ,
\end{aligned}
$$

where the last inequality uses the fact that any $\boldsymbol{\Delta} \in \mathcal{M}_1$ is at most rank-r, and $\|\boldsymbol{\delta}\|_1 \leq \sqrt{r}\|\boldsymbol{\delta}\|_2$ for any r-sparse vector $\boldsymbol{\delta}$. Combining all the inequalities, we complete the proof.

■

Appendix 5.B Proof of Probabilistic Analysis

5.B.1 Proof of Theorem 15

Proof: Let (Ω, μ) be the probability space that $\mathbf{X}$ is defined on, and construct

$$\mathcal{H} = \{h(\cdot) = \langle\langle \cdot, \mathbf{\Delta} \rangle\rangle \mid \mathbf{\Delta} \in \mathcal{C}\} \ .$$

$|||\mathbf{X}|||_{\psi_2} \leq \kappa$ immediately implies that $\sup_{h \in \mathcal{H}} |||h|||_{\psi_2} \leq \kappa$. As $\mathbf{X}$ is isotropic, i.e., $\mathbb{E}[\langle\langle \mathbf{X}, \mathbf{\Delta} \rangle\rangle^2] = 1$ for any $\mathbf{\Delta} \in \mathcal{C} \subseteq \mathbb{S}^{dp-1}$, thus $\mathcal{H} \subseteq S_{L_2}$ and $\mathbb{E}[h^2] = 1$ for any $h \in \mathcal{H}$. Given $h_1 = \langle\langle \cdot, \mathbf{\Delta}_1 \rangle\rangle, h_2 = \langle\langle \cdot, \mathbf{\Delta}_2 \rangle\rangle \in \mathcal{H}$, where $\mathbf{\Delta}_1, \mathbf{\Delta}_2 \in \mathcal{C}$, the metric induced by ψ_2 norm satisfies

$$|||h_1 - h_2|||_{\psi_2} = |||\langle\langle \mathbf{X}, \mathbf{\Delta}_1 - \mathbf{\Delta}_2 \rangle\rangle|||_{\psi_2} \leq \kappa \left\| \mathbf{\Delta}_1 - \mathbf{\Delta}_2 \right\|_F .$$

Using the properties of γ_2-functional and Lemma 4, we have

$$\gamma_2(\mathcal{H}, |||\cdot|||_{\psi_2}) \ \leq \ \kappa \gamma_2(\mathcal{C}, \| \cdot \|_F) \ \leq \ \kappa c_4 w(\mathcal{C}) \ ,$$

where c_4 is an absolute constant. Hence, by choosing $\beta = \frac{c_1 c_4 \kappa^2 w(\mathcal{C})}{\sqrt{n}}$, we can guarantee that condition $c_1 \kappa \gamma_2(\mathcal{H}, |||\cdot|||_{\psi_2}) \leq \beta \sqrt{n}$ holds for $\mathcal{H}$. Applying Lemma 3 to this $\mathcal{H}$, with probability at least $1 - \exp(-c_2 c_1^2 c_4^2 w^2(\mathcal{C}))$, we have

$$\sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} h^2(\mathbf{X}_i) - 1 \right| \leq \beta \ ,$$

which implies

$$\inf_{\Delta \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^{n} \langle\langle X_i, \Delta \rangle\rangle^2 \geq 1 - \beta \ .$$

Letting $\zeta = c_2 c_1^2 c_4^2$, $\xi = c_1 c_4$, we complete the proof. ■

5.B.2 proof of Theorem 17

Proof: Let $\boldsymbol{\omega} = [\omega_1, \omega_2, \ldots, \omega_n]^T$. For each entry in $\boldsymbol{\omega}$, we have

$$\sqrt{\mathbb{E}[\omega_i^2]} \leq \sqrt{2} |||\omega_i|||_{\psi_2} = \sqrt{2}\tau \ ,$$
$$\left|\left|\left|\omega_i^2 - \mathbb{E}[\omega_i^2]\right|\right|\right|_{\psi_1} \leq 2\left|\left|\left|\omega_i^2\right|\right|\right|_{\psi_1} \leq 4|||\omega_i|||_{\psi_2}^2 \leq 4\tau^2 \ ,$$

where we use the definition of ψ_2 norm and its relation to ψ_1 norm. By Bernstein's inequality, we get

$$\mathbb{P}\left(\|\boldsymbol{\omega}\|_2^2 - 2\tau^2 \geq \epsilon\right) \leq \mathbb{P}\left(\|\boldsymbol{\omega}\|_2^2 - \mathbb{E}[\|\boldsymbol{\omega}\|_2^2] \geq \epsilon\right) \leq \exp\left(-c_1 \min\left(\frac{\epsilon^2}{16\tau^4 n}, \ \frac{\epsilon}{4\tau^2}\right)\right) \ .$$

Taking $\epsilon = 4\tau^2 n$, we have

$$\mathbb{P}\left(\|\boldsymbol{\omega}\|_2 \geq \tau\sqrt{6n}\right) \leq \exp\left(-c_1 n\right) \ .$$

Denote $\mathbf{Y}_{\mathbf{u}} = \sum_{i=1}^n u_i \mathbf{X}_i$ for $\mathbf{u} \in \mathbb{R}^n$. For any $\mathbf{u} \in \mathbb{S}^{n-1}$, we get $|||\mathbf{Y}_{\mathbf{u}}|||_{\psi_2} \leq c\kappa$ due to

$$|||\langle\langle \mathbf{Y}_{\mathbf{u}}, \boldsymbol{\Delta}\rangle\rangle|||_{\psi_2} = \left|\left|\left|\sum_{i=1}^n u_i \langle\langle \mathbf{X}_i, \boldsymbol{\Delta}\rangle\rangle\right|\right|\right|_{\psi_2} \leq c\sqrt{\sum_{i=1}^n u_i^2 |||\langle\langle \mathbf{X}_i, \boldsymbol{\Delta}\rangle\rangle|||_{\psi_2}^2} \leq \ c\kappa, \ \forall \ \boldsymbol{\Delta} \in \mathbb{S}^{dp-1}.$$

For the rest of the proof, we may drop the subscript of $\mathbf{Y}_{\mathbf{u}}$ for convenience. We construct the stochastic process $\{Z_{\boldsymbol{\Delta}} = \langle\langle \mathbf{Y}, \boldsymbol{\Delta}\rangle\rangle\}_{\boldsymbol{\Delta}\in\Omega}$, and note that any $Z_{\mathbf{U}}$ and $Z_{\mathbf{V}}$ from this process satisfy

$$\mathbb{P}\left(|Z_{\mathbf{U}} - Z_{\mathbf{V}}| \geq \epsilon\right) = \mathbb{P}\left(|\langle\langle \mathbf{Y}, \mathbf{U} - \mathbf{V}\rangle\rangle| \geq \epsilon\right) \leq e \cdot \exp\left(-\frac{C\epsilon^2}{\kappa^2\|\mathbf{U} - \mathbf{V}\|_F^2}\right) \ ,$$

for some universal constant C due to the sub-Gaussianity of $\mathbf{Y}$. As Ω is symmetric, it follows that

$$\sup_{\mathbf{U},\mathbf{V}\in\Omega} |Z_{\mathbf{U}} - Z_{\mathbf{V}}| = 2 \sup_{\boldsymbol{\Delta}\in\Omega} Z_{\boldsymbol{\Delta}} \ ,$$

$$\sup_{\mathbf{U},\mathbf{V}\in\Omega} \|\mathbf{U} - \mathbf{V}\|_F = 2 \sup_{\boldsymbol{\Delta}\in\Omega} \|\boldsymbol{\Delta}\|_F = 2\eta \ .$$

Using Lemma 2, we have

$$\mathbb{P}\left(2 \sup_{\boldsymbol{\Delta}\in\Omega} Z_{\boldsymbol{\Delta}} \geq c_4\kappa\left(\gamma_2(\Omega, \|\cdot\|_F) + \epsilon\cdot 2\eta\right)\right) \leq c_2 \exp\left(-\epsilon^2\right) \ ,$$

where c_2 and c_4 are absolute constant. By Lemma 4, there exist constants c_3 and c_5 such that

$$\mathbb{P}\left(2\|\mathbf{Y}\|_* \geq c_5\kappa\left(w(\Omega) + \epsilon\right)\right) = \mathbb{P}\left(2 \sup_{\boldsymbol{\Delta}\in\Omega} Z_{\boldsymbol{\Delta}} \geq c_5\kappa\left(w(\Omega)+\epsilon\right)\right) \leq c_2 \exp\left(-\frac{\epsilon^2}{c_3^2\eta^2}\right) .$$

Letting $\epsilon = w(\Omega)$, we have for any $\mathbf{u} \in \mathbb{S}^{n-1}$

$$\mathbb{P}\left(\|\mathbf{Y}_{\mathbf{u}}\| \geq c_5\kappa w(\Omega)\right) \leq c_2 \exp\left(-\left(\frac{w(\Omega)}{c_3\eta}\right)^2\right)$$

Combining this with the bound for $\|\boldsymbol{\omega}\|_2$ and letting $c_0 = \sqrt{6}c_5$, by union bound, we have

$$\begin{aligned}
\mathbb{P}\left(\left\|\sum_{i=1}^n \omega_i \mathbf{X}_i\right\|_* \geq c_0\kappa\tau\sqrt{n}w(\Omega)\right) &\leq \mathbb{P}\left(\frac{\|\mathbf{Y}_{\boldsymbol{\omega}}\|_*}{\|\boldsymbol{\omega}\|_2} \geq c_5\kappa w(\Omega)\right) + \mathbb{P}\left(\|\boldsymbol{\omega}\|_2 \geq \tau\sqrt{6n}\right) \\
&\leq \sup_{\mathbf{u}\in\mathbb{S}^{n-1}} \mathbb{P}\left(\|\mathbf{Y}_{\mathbf{u}}\|_* \geq c_5\kappa w(\Omega)\right) + \mathbb{P}\left(\|\boldsymbol{\omega}\|_2 \geq \tau\sqrt{6n}\right) \\
&\leq c_2 \exp\left(-\frac{w^2(\Omega)}{c_3^2\eta^2}\right) + \exp\left(-c_1 n\right) \ ,
\end{aligned}$$

which completes the proof. ■

5.B.3 Proof of Theorem 16

To facilitate the proof of Theorem 16, we will use some properties specific to the Gaussian random matrix $\mathbf{G} \in \mathbb{R}^{d\times p}$, which are summarized as follows. The symbol "$\sim$" means "has the same distribution as".

Property 1: Given an m-dimensional subspace $\mathcal{M} \subseteq \mathbb{R}^{d\times p}$ spanned by orthonormal basis $\mathbf{U}_1, \ldots, \mathbf{U}_m$,

$$\mathbf{G}_{\mathcal{M}} \sim \sum_{i=1}^{m} g_i \mathbf{U}_i,$$

where g_i's are i.i.d. standard Gaussian random variables. Moreover, $\mathbb{E}\left[\|\mathbf{G}_{\mathcal{M}}\|_F^2\right] = m$.

Proof: Given the orthonormal basis $\mathbf{U}_1, \ldots, \mathbf{U}_m$ of subspace $\mathcal{M}$, $\mathbf{G}_{\mathcal{M}}$ can be written as

$$\mathbf{G}_{\mathcal{M}} = \sum_{i=1}^{m} \langle\langle \mathbf{G}, \mathbf{U}_i \rangle\rangle \cdot \mathbf{U}_i$$

Since $\|\mathbf{U}_1\|_F = \ldots = \|\mathbf{U}_m\|_F = 1$, each $\langle\langle \mathbf{G}, \mathbf{U}_i \rangle\rangle$ is standard Gaussian. Moreover, as $\mathbf{U}_1, \ldots, \mathbf{U}_m$ are orthogonal, $\langle\langle \mathbf{G}, \mathbf{U}_i \rangle\rangle$ are independent of each other. ■

Property 2: $\mathbf{G}_{\mathcal{M}_1}$ and $\mathbf{G}_{\mathcal{M}_2}$ are independent if $\mathcal{M}_1, \mathcal{M}_2 \subseteq \mathbb{R}^{d\times p}$ are orthogonal subspaces.

Proof: Suppose that the orthonormal bases of $\mathcal{M}_1, \mathcal{M}_2$ are given by $\mathbf{U}_1, \ldots, \mathbf{U}_{m_1}$ and

$\mathbf{V}_1, \ldots, \mathbf{V}_{m_2}$ respectively. Using Property 1 above, $\mathbf{G}_{\mathcal{M}_1}$ and $\mathbf{G}_{\mathcal{M}_2}$ can be written as

$$\mathbf{G}_{\mathcal{M}_1} = \sum_{i=1}^{m_1} \langle\langle \mathbf{G}, \mathbf{U}_i \rangle\rangle \cdot \mathbf{U}_i \;\sim\; \sum_{i=1}^{m_1} g_i \mathbf{U}_i \ ,$$
$$\mathbf{G}_{\mathcal{M}_2} = \sum_{i=1}^{m_2} \langle\langle \mathbf{G}, \mathbf{V}_i \rangle\rangle \cdot \mathbf{V}_i \;\sim\; \sum_{i=1}^{m_2} h_i \mathbf{V}_i \ ,$$

where $g_1, \ldots, g_{m_1}$ and $h_1, \ldots, h_{m_2}$ are all standard Gaussian. As $\mathcal{M}_1, \mathcal{M}_2 \subseteq \mathbb{R}^{d \times p}$ are orthogonal, $\mathbf{U}_1, \ldots, \mathbf{U}_{m_1}$ and $\mathbf{V}_1, \ldots, \mathbf{V}_{m_2}$ are orthogonal to each other as well, which implies that $g_1, \ldots, g_{m_1}$ and $h_1, \ldots, h_{m_2}$ are all independent. Therefore $\mathbf{G}_{\mathcal{M}_1}$ and $\mathbf{G}_{\mathcal{M}_2}$ are independent. ■

Property 3: Given a subspace

$$\mathcal{M} = \{\boldsymbol{\Theta} \in \mathbb{R}^{d \times p} \mid \text{colsp}(\boldsymbol{\Theta}) \subseteq \mathcal{U},\ \text{rowsp}(\boldsymbol{\Theta}) \subseteq \mathcal{V}\} \ ,$$

where $\mathcal{U} \subseteq \mathbb{R}^d$, $\mathcal{V} \subseteq \mathbb{R}^p$ are two subspaces of dimension m_1 and m_2 respectively, then $\|\mathbf{G}_{\mathcal{M}}\|_{\text{op}}$ satisfies

$$\|\mathbf{G}_{\mathcal{M}}\|_{\text{op}} \sim \|\mathbf{G}'\|_{\text{op}} \ ,$$

where $\mathbf{G}'$ is an $m_1 \times m_2$ matrix with i.i.d. standard Gaussian entries.

Proof: Suppose that the orthonormal bases for $\mathcal{U}$ and $\mathcal{V}$ are $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_{m_1}]$ and $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_{m_2}]$ respectively, and $\mathbf{U}_\perp$ and $\mathbf{V}_\perp$ denote the orthonormal bases for their orthogonal complement. It is easy to see that the orthonormal basis for $\mathcal{M}$ can be given by $\{\mathbf{u}_i\mathbf{v}_j^T \mid 1 \leq i \leq m_1,\ 1 \leq j \leq m_2\}$. Using Property 1, we have

$$\begin{aligned}
\mathbf{G}_{\mathcal{M}} \ \sim\ \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} G'_{ij}\mathbf{u}_i\mathbf{v}_j^T &= \mathbf{U}\mathbf{G}'\mathbf{V} \\
&= [\mathbf{U},\mathbf{U}_\perp]\cdot \begin{bmatrix} \mathbf{G}' & \mathbf{0}_{m_1\times(p-m_2)} \\ \mathbf{0}_{(d-m_1)\times m_2} & \mathbf{0}_{(d-m_1)\times(p-m_2)} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{V}^T \\ \mathbf{V}_\perp^T \end{bmatrix}
\end{aligned}$$

where $\mathbf{G}'$ is a $m_1 \times m_2$ standard Gaussian random matrix. Note that both $[\mathbf{U},\mathbf{U}_\perp] \in \mathbb{R}^{d\times d}$ and $[\mathbf{V},\mathbf{V}_\perp] \in \mathbb{R}^{p\times p}$ are unitary matrices, because they form the orthonormal bases for $\mathbb{R}^d$ and $\mathbb{R}^p$ respectively. If we denote $\begin{bmatrix} \mathbf{G}' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$ by $\mathbf{W}$, then $\|\mathbf{G}_{\mathcal{M}}\|_{\mathrm{op}} = \|\mathbf{W}\|_{\mathrm{op}}$ as spectral norm is unitarily invariant. Further, if the SVD of $\mathbf{G}'$ is $\mathbf{G}' = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^T$, where $\mathbf{U}_1 \in \mathbb{R}^{m_1\times m_1}$, $\mathbf{\Sigma}_1 \in \mathbb{R}^{m_1\times m_2}$ and $\mathbf{V}_1 \in \mathbb{R}^{m_2\times m_2}$, then the SVD of $\mathbf{W}$ is given by

$$\begin{aligned}
\mathbf{W} = &\begin{bmatrix} \mathbf{U}_1 & \mathbf{0}_{m_1\times(d-m_1)} \\ \mathbf{0}_{(d-m_1)\times m_1} & \mathbf{U}_2 \end{bmatrix} \times \begin{bmatrix} \mathbf{\Sigma}_1 & \mathbf{0}_{m_1\times(p-m_2)} \\ \mathbf{0}_{(d-m_1)\times m_2} & \mathbf{0}_{(d-m_1)\times(p-m_2)} \end{bmatrix} \\
&\times \begin{bmatrix} \mathbf{V}_1^T & \mathbf{0}_{m_2\times(p-m_2)} \\ \mathbf{0}_{(p-m_2)\times m_2} & \mathbf{V}_2^T \end{bmatrix} ,
\end{aligned}$$

where $\mathbf{U}_2 \in \mathbb{R}^{(d-m_1)\times(d-m_1)}$ and $\mathbf{V}_2 \in \mathbb{R}^{(p-m_2)\times(p-m_2)}$ are arbitrary unitary matrices. From the equation above, we can see that $\mathbf{W}$ and $\mathbf{G}'$ share the same singular values, thus $\|\mathbf{G}_{\mathcal{M}}\|_{\mathrm{op}} = \|\mathbf{W}\|_{\mathrm{op}} = \|\mathbf{G}'\|_{\mathrm{op}}$. ■

Property 4: The operator norm $\|\mathbf{G}\|_{\mathrm{op}}$ satisfies

$$\mathbb{P}\left(\|\mathbf{G}\|_{\mathrm{op}} \geq \sqrt{d} + \sqrt{p} + \epsilon\right) \leq \exp\left(-\frac{\epsilon^2}{2}\right) \ , \tag{5.35}$$

$$\mathbb{E}\left[\|\mathbf{G}\|_{\mathrm{op}}\right] \leq \sqrt{d} + \sqrt{p} \ , \tag{5.36}$$

$$\mathbb{E}\left[\|\mathbf{G}\|_{\mathrm{op}}^2\right] \leq \left(\sqrt{d} + \sqrt{p}\right)^2 + 2 \ . \tag{5.37}$$

(5.35) and (5.36) are the classical results on the extreme singular value of Gaussian random matrix [146,172] (see Theorem 5.32 and Corollary 5.35 in [172]). (5.37) is used in [40] (see (82) - (87) in [40]).

Property 5: For a subset of unit sphere $\mathcal{A} \subseteq \mathbb{S}^{dp-1}$, A useful inequality [4,40] is given by the Gaussian width satisfies

$$w^2(\mathcal{A}) \leq \mathbb{E}_{\mathbf{G}}\left[\inf_{\mathbf{Z}\in\mathcal{N}} \|\mathbf{G}-\mathbf{Z}\|_F^2\right] \ , \tag{5.38}$$

in which $\mathcal{N} = \{\mathbf{Z} \mid \langle\langle \mathbf{Z}, \mathbf{\Delta}\rangle\rangle \leq 0 \text{ for all } \mathbf{\Delta} \in \mathcal{A}\}$ is the polar cone of $\text{cone}(\mathcal{A})$.

This property is essentially Proposition 10.2 in [4], and the right-hand side is often called *statistical dimension*. Now we are ready to present the proof of Theorem 16.

Proof of Theorem 16: Let $\boldsymbol{\theta}^*$ be any subgradient of $f(\cdot)$ at $\boldsymbol{\sigma}^*$, i.e., $\boldsymbol{\theta}^* \in \partial f(\boldsymbol{\sigma}^*)$, and $\mathbf{\Gamma} = \mathbf{U}\,\text{Diag}(\boldsymbol{\theta}^*_{1:r})\mathbf{V}$. We define

$$\mathcal{D} = \{\mathbf{W} \mid \mathbf{W} \in \mathcal{M}_2,\ \boldsymbol{\sigma}(\mathbf{W}) \preceq \mathbf{z}\} \ , \quad \mathcal{K} = \{\mathbf{\Gamma} + \mathbf{W} \mid \mathbf{W} \in \mathcal{D}\} \ ,$$

where the symbol "$\preceq$" means "elementwise less than or equal". It is not difficult to see that $\mathcal{K}$ is a subset of $\partial\|\mathbf{\Theta}^*\|$, as any $\mathbf{Z} \in \mathcal{K}$ satisfies $\|\mathbf{Z}\|_* = f_*(\boldsymbol{\sigma}(\mathbf{Z})) \leq f_*(\boldsymbol{\theta}^*) = 1$ and $\langle\langle \mathbf{Z}, \mathbf{\Theta}^*\rangle\rangle = \langle \boldsymbol{\sigma}(\mathbf{Z}), \boldsymbol{\sigma}^*\rangle = \langle \boldsymbol{\theta}^*_{1:r}, \boldsymbol{\sigma}^*_{1:r}\rangle = f(\boldsymbol{\sigma}^*) = \|\mathbf{\Theta}^*\|$. Hence we have

$$\text{cone}(\mathcal{K}) \subset \text{cone}\{\partial\|\mathbf{\Theta}^*\|\} = \mathcal{N} \ ,$$

where $\mathcal{N}$ is the polar cone of $\mathcal{T}$, and the equality follows from the Theorem 23.7 of [144]. We define the subspace $\mathcal{M}^\perp$ as the orthogonal complement of $\mathcal{M}_1 \oplus \mathcal{M}_2$. For the sake of convenience, we denote by $\mathbf{G}_1$ ($\mathbf{G}_2$, $\mathbf{G}_\perp$) the orthogonal projection of $\mathbf{G}$ onto $\mathcal{M}_1$

$(\mathcal{M}_2, \mathcal{M}_\perp)$, and denote cone($\mathcal{K}$) by $\mathcal{C}_\mathcal{K}$. Using (5.38), we obtain

$$\begin{aligned} w(\mathcal{C})^2 &\leq \mathbb{E}\left[\inf_{\mathbf{Z}\in\mathcal{N}} \|\mathbf{G}-\mathbf{Z}\|_F^2\right] \\ &\leq \mathbb{E}\left[\inf_{\mathbf{Z}\in\mathcal{C}_\mathcal{K}} \|\mathbf{G}_1-\mathbf{Z}_1\|_F^2 + \|\mathbf{G}_2-\mathbf{Z}_2\|_F^2 + \|\mathbf{G}_\perp-\mathbf{Z}_\perp\|_F^2\right] \\ &= \mathbb{E}\left[\inf_{t\geq 0,\ \mathbf{W}\in t\mathcal{D}} \|\mathbf{G}_1-t\mathbf{\Gamma}\|_F^2 + \|\mathbf{G}_2-\mathbf{W}\|_F^2\right] + \mathbb{E}\left[\|\mathbf{G}_\perp\|_F^2\right]\ . \end{aligned} \tag{5.39}$$

To further bound the expectations, we let $t_0 = \frac{\|\mathbf{G}_2\|_{\mathrm{op}}}{\theta^*_{\min}}$, which is a random quantity depending on $\mathbf{G}$. Therefore, we have

$$\begin{aligned} &\mathbb{E}\left[\inf_{t\geq 0,\ \mathbf{W}\in t\mathcal{D}} \|\mathbf{G}_1-t\mathbf{\Gamma}\|_F^2 + \|\mathbf{G}_2-\mathbf{W}\|_F^2\right] \\ &\leq \mathbb{E}\left[\|\mathbf{G}_1-t_0\mathbf{\Gamma}\|_F^2\right] + \mathbb{E}\left[\inf_{\mathbf{W}\in t_0\mathcal{D}} \|\mathbf{G}_2-\mathbf{W}\|_F^2\right] \\ &= \mathbb{E}\left[\|\mathbf{G}_1\|_F^2\right] + 2\mathbb{E}\left[\langle\langle \mathbf{G}_1, t_0\mathbf{\Gamma}\rangle\rangle\right] + \|\boldsymbol{\theta}^*_{1:r}\|_2^2\cdot\mathbb{E}\left[t_0^2\right] + 0 \\ &= r^2 + 0 + \mathbb{E}\left[\|\mathbf{G}_2\|_{\mathrm{op}}^2\right]\cdot\frac{\|\boldsymbol{\theta}^*_{1:r}\|_2^2}{\theta^{*2}_{\min}} \\ &\leq r^2 + ((\sqrt{d-r}+\sqrt{p-r})^2+2)\cdot\frac{\|\boldsymbol{\theta}^*_{1:r}\|_2^2}{\theta^{*2}_{\min}} \\ &\leq r^2 + 2\rho^2 r\,(d+p-2r)\ , \end{aligned} \tag{5.40}$$

where the second equality uses Property 1 and 2, and the second inequality follows from Property 3 and 4. Since $\mathcal{M}_\perp$ is a $r(d+p-2r)$-dimensional subspace, by Property 1 we have $\mathbb{E}\left[\|\mathbf{G}_\perp\|_F^2\right] = r(d+p-2r)$. Combining it with (5.39) and (5.40), we have

$$w(\mathcal{C}) \leq \sqrt{(2\rho^2+1)\,(d+p-r)\,r}\ . \tag{5.41}$$

On the other hand, as $\mathcal{C}\subseteq\mathbb{S}^{dp-1}$, we always have

$$w(\mathcal{C}) \leq \mathbb{E}\left[\|\mathbf{G}\|_F\right] \leq \sqrt{\mathbb{E}\left[\|\mathbf{G}\|_F^2\right]} = \sqrt{dp}\ . \tag{5.42}$$

We finish the proof by combining the two bounds for $w(\mathcal{C})$. ■

Chapter 6

Robust Structured Estimation for Single-Index Models

6.1 Introduction

In previous chapters, we focus on structured estimation for linear models. The simplicity of linear model leads to its great interpretability and computational efficiency, which are often favored in practical applications. Despite these attractive merits, one main drawback of linear models is the stringent assumption of linear relationship between $\mathbf{x}$ and y, which may fail to hold in complicated scenarios. To introduce more flexibility, one option is to consider the general single-index models (SIMs) [73, 77],

$$\mathbb{E}[y|\mathbf{x}] = f^*(\langle \boldsymbol{\theta}^*, \mathbf{x} \rangle) \ , \tag{6.1}$$

where $f^* : \mathbb{R} \mapsto \mathbb{R}$ is an *unknown* univariate transfer function (a.k.a. link function). This class of models enjoys rich modeling power in the sense that it encompasses several useful models as special cases, which are briefly described below:

- **One-bit Compressed Sensing**: In one-bit compressed sensing (1-bit CS) [23, 134], the response y is restricted to be binary, i.e., $y \in \{+1, -1\}$, and the range of transfer function f^* is $[-1, 1]$. Given the measurement vector $\mathbf{x}$, one can generate y from the Bernoulli model,

$$\frac{y+1}{2} \sim \text{Ber}\left(\frac{f^*(\langle \boldsymbol{\theta}^*, \mathbf{x} \rangle) + 1}{2}\right) . \tag{6.2}$$

  In the noiseless case, $f^*(z) = \text{sign}(z)$ and y always reflects the true sign of $\langle \boldsymbol{\theta}^*, \mathbf{x} \rangle$, while y can be incorrect for other f^* whose shape determines the noise level in some way.

- **Generalized Linear Models**: In generalized linear models (GLMs) [119], the transfer function is assumed to be *monotonically increasing* and conditional distribution of $y|\mathbf{x}$ belongs to exponential family. Different choices of f^* give rise to different members in GLMs. If f^* is identity function $f^*(z) = z$, one has the simple linear models, while the sigmoid function $f^*(z) = \frac{1}{1+e^{-z}}$ results in the logistic model for binary classification. In this work, however, we have *no access* to exact f^* other than knowing it is monotone.

- **Noise in Monotone Transfer**: Instead of having the general expectation form of y as GLMs, one could directly introduce the noise inside monotone transfer $\tilde{f}$ to model the randomness of y [135],

$$y = \tilde{f}\left(\langle \boldsymbol{\theta}^*, \mathbf{x} \rangle + \epsilon\right) . \tag{6.3}$$

  In this setting, the transfer function $\tilde{f}$ is slightly different from the f^* in (6.1), which are related by $f^*(z) = \mathbb{E}_\epsilon[\tilde{f}(z+\epsilon)|z]$.

A key advantage of SIM is its robustness. First, allowing unknown f^* prevents the mis-specification of transfer function, which could otherwise lead to a poor estimate of

$\boldsymbol{\theta}^*$. Second, the model in (6.1) makes minimal assumption on the distribution of y, thus being able to tolerate potentially heavy-tailed noise.

In the absence of exact f^*, though 1-bit CS and related variants were well-studied in recent years [23, 42, 62, 81, 104, 134, 151, 152, 181, 186, 191], the exploration of general SIMs or the cases with monotone transfers is relatively limited, especially in the high-dimensional regime. [93] and [92] investigated the low-dimensional SIMs with monotone transfers, and they proposed perceptron-type algorithms to estimate both f^* and $\boldsymbol{\theta}^*$, with provable guarantees on prediction error. In high dimension, general SIMs were studied by [3] and [136], in which only unstructured sparsity of $\boldsymbol{\theta}^*$ is considered. The algorithm developed in [3] relies on reversible jump MCMC, which could be slow. In [136], a path fitting algorithm is designed to recover f^* and $\boldsymbol{\theta}^*$, but only asymptotic guarantees are provided. [58] considered the high-dimensional setting with monotone transfer, and their iterative algorithm is based on non-convex optimization, for which it is hard to establish the convergence. Besides, the prediction error bound they derived is also weak (in the sense that it is even worse than the initialization of the algorithm). Recently [131] proposed a constrained least-squares method to estimate $\boldsymbol{\theta}^*$, with recovery error characterized by Gaussian width and related quantities. Though their analysis considered the general structure of $\boldsymbol{\theta}^*$, it only holds for noiseless setting where $y = f(\langle \boldsymbol{\theta}^*, \mathbf{x} \rangle)$. General structure of $\boldsymbol{\theta}^*$ was also explored in [173] and [135]. Other types of statistical guarantees for high-dimensional SIMs is also available, such as support recovery of $\boldsymbol{\theta}^*$ in [129]. It is worth noting that all the aforementioned statistical analyses rely on sub-Gaussian noise or the transfer function being bounded or Lipschitz, which indicates that none of the results can immediately hold for heavy-tailed noise (or without Lipschitzness and boundedness).

In this chapter, we focus on the parameter estimation of $\boldsymbol{\theta}^*$ instead of the prediction of y given new $\mathbf{x}$, given n measurements of $(\mathbf{x}, y) \in \mathbb{R}^p \times \mathbb{R}$, denoted by $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$.

In particular, we propose two families of generalized estimators, constrained and regularized, for model (6.1) under Gaussian measurement. The parameter $\boldsymbol{\theta}^*$ is assumed to possess certain low-complexity structure, which can be either captured by a constraint $\boldsymbol{\theta}^* \in \mathcal{K}$ or a norm regularization term $\|\boldsymbol{\theta}^*\|$. Our general approach is inspired by U-statistics [98] and the advances in 1-bit CS, and subsumes several existing 1-bit CS algorithms [42,186] as special cases. Similar to those algorithms, our estimator is simple and often admits closed-form solutions. Apart from 1-bit CS, we particularly investigate the model (6.3), for which the generalized estimator is specialized in a novel way. The resulting estimator better leverages the monotonicity of the transfer function, which is also demonstrated through experiments. Regarding the recovery analysis, there are two appealing aspects. First our results work for general structure, with error bound characterized by Gaussian width and some other easy-to-compute geometric measures. Instantiating our results with specific structure of $\boldsymbol{\theta}^*$ recovers previously established error bounds for 1-bit CS [42,186], which are sharper than those yielded by the general analysis in [134]. Second, our analysis works with limited assumptions on the condition distribution of y. In particular, our estimator is robust to heavy-tailed noise and permit unbounded transfer functions f^* as well as non-Lipschitz ones. For the ease of exposition, whenever we say "monotone", it means "monotonically increasing" by default. Throughout the chapter, we will use c, C, C', C_0, C_1 and so on to denote absolute constants, which may differ from context to context.

The rest of the chapter is organized as follows. In Section 6.2, we introduce our robust estimators for SIMs. Section 6.3 presents the statistical guarantees of the proposed estimators. Different structures of $\boldsymbol{\theta}^*$ are also discussed. In Section 6.4, we provide two few examples, 1-bit CS and monotone transfer, to illustrate the specialization of the general estimators. In Section 6.5, we complement our theoretical developments with some experiment results.

6.2 Robust Estimators

6.2.1 Assumptions

For the sake of identifiability, we assume w.l.o.g. that $\|\boldsymbol{\theta}^*\|_2 = 1$ throughout the chapter. At the first glimpse of model (6.1), we may realize that it is difficult to recover $\boldsymbol{\theta}^*$ due to unknown f^*. In contrast, when f^* is given, the recovery guarantees of $\boldsymbol{\theta}^*$ can be established under mild assumptions of $\mathbf{x}$ and y, such as boundedness or sub-Gaussianity. If we know certain properties of the transfer function like the monotonicity introduced in GLMs and (6.3), the structure of f^* is largely restricted, and it is tempting to expect that similar results will continue to hold. Unfortunately, we first have the following claim, which indicates that without other constraints on f^* beyond strict monotonicity, $\boldsymbol{\theta}^*$ cannot be consistently estimated under general sub-Gaussian (or bounded) measurement, even in the noiseless setting of (6.3).

Claim 1 *Suppose that each element x_i of $\mathbf{x}$ is sampled i.i.d. from Rademacher distribution, i.e., $\mathbb{P}(x_i = 1) = \mathbb{P}(x_i = -1) = 0.5$. Under model (6.3) with noise $\epsilon = 0$, there exists a $\bar{\boldsymbol{\theta}} \in \mathbb{S}^{p-1}$ together with a monotone $\bar{f}$, such that $\mathrm{supp}(\bar{\boldsymbol{\theta}}) = \mathrm{supp}(\boldsymbol{\theta}^*)$ and $y_i = \bar{f}(\langle \bar{\boldsymbol{\theta}}, \mathbf{x}_i \rangle)$ for data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with arbitrarily large sample size n, while $\|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 > \delta$ for some constant δ.*

Proof: In the noiseless setting with unknown f^*, provided that $\mathcal{S} \triangleq \mathrm{supp}(\boldsymbol{\theta}^*)$ is given and $|\mathcal{S}| = s$, the estimation of $\boldsymbol{\theta}^*$ is simplified as

$$\text{Find } \ \boldsymbol{\theta}_{\mathcal{S}} \in \mathbb{S}^{s-1} \qquad \text{s.t.} \qquad \mathrm{sign}\left(\langle \boldsymbol{\theta}_{\mathcal{S}}, \mathbf{x}_{i\mathcal{S}} - \mathbf{x}_{j\mathcal{S}} \rangle\right) = \mathrm{sign}(y_i - y_j), \quad \forall \ 1 \le i < j \le n \ , \tag{6.4}$$

any of whose solution $\boldsymbol{\theta}$ can be true $\boldsymbol{\theta}^*$ on the premise that no other information is available, since there always exists a monotone f satisfying $f(\langle \boldsymbol{\theta}, \mathbf{x}_i \rangle) = y_i$. Given the distribution of $\mathbf{x}$, $\mathbf{x}_{i\mathcal{S}} - \mathbf{x}_{j\mathcal{S}}$ only has 3^s possibilities even if $n \to +\infty$. We denote the feasible set of (6.4) by $\mathcal{C}$, which is basically an intersection of $\mathbb{S}^{s-1}$ and at most $\min\{n(n-1), 3^p\}$ halfspaces (or hyperplanes if $y_i = y_j$). Depending on the 3 different values of each $\text{sign}(y_i - y_j)$, this feasible set $\mathcal{C}$ has at most $3^{\min\{n(n-1),3^p\}}$ possibilities, which is finite, and the union of them should be $\mathbb{S}^{s-1}$. When $s \geq 2$ and the constant δ is small enough, we can always find a $\mathcal{C}$, in which there exist two different points away by δ. Specify them as $\boldsymbol{\theta}_{*\mathcal{S}}$ and $\bar{\boldsymbol{\theta}}_{\mathcal{S}}$ respectively, and we are unable to distinguish between them, as both can be solution to (6.4) for any samples. ■

Now that consistent estimation of $\boldsymbol{\theta}^*$ is not possible for general sub-Gaussian measurement, it might be reasonable to focus on certain special cases. For this work, we assume that $\mathbf{x}$ is standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$. For SIM (6.1), we additionally assume that the distribution of y depends on $\mathbf{x}$ *only* through the value of $\langle \boldsymbol{\theta}^*, \mathbf{x} \rangle$, i.e., the distribution of $y|\mathbf{x}$ is fixed if $\langle \boldsymbol{\theta}^*, \mathbf{x} \rangle$ is given (no matter what the exact $\mathbf{x}$ is). This assumption is quite minimal, and it turns out that the examples we provide in Section 6.1 all satisfy it (if noise ϵ is independent of $\mathbf{x}$ in (6.3)). The same assumption is used in [135] as well.

Under the assumptions above, given m i.i.d. observations $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$, we define

$$\mathbf{u}\left((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\right) = \sum_{i=1}^{m} q_i\left(y_1, \ldots, y_m\right) \cdot \mathbf{x}_i \ , \tag{6.5}$$

where all $q_i : \mathbb{R}^m \mapsto \mathbb{R}$ are bounded functions with $|q_i| \leq 1$, which are chosen along with m based on the properties of the transfer function. In Section 6.4, we will see examples for their choices. The vector $\mathbf{u} \in \mathbb{R}^p$ is critical due to the key observation below.

Lemma 12 *Suppose the distribution of y in model* (6.1) *depends on $\mathbf{x}$ through $\langle \boldsymbol{\theta}^*, \mathbf{x} \rangle$ and we define accordingly*

$$b_i\left(z_1, \ldots, z_m; \boldsymbol{\theta}^*\right) = \mathbb{E}\left[q_i\left(y_1, \ldots, y_m\right) | \langle \boldsymbol{\theta}^*, \mathbf{x}_1 \rangle = z_1, \ldots, \langle \boldsymbol{\theta}^*, \mathbf{x}_m \rangle = z_m\right] \ . \tag{6.6}$$

With $\mathbf{x}$ *being standard Gaussian* $\mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{u}$ *defined in* (6.5) *satisfies*

$$\mathbb{E}\left[\mathbf{u}\left((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\right)\right] = \beta\boldsymbol{\theta}^* \ , \tag{6.7}$$

where $\beta = \sum_{i=1}^m \mathbb{E}[b_i\left(g_1, \ldots, g_m; \boldsymbol{\theta}^*\right) \cdot g_i]$, *and* $g_1, \ldots, g_m$ *are i.i.d. standard Gaussian.*

Proof: Let $\boldsymbol{\theta}_\perp$ be any vector orthogonal to $\boldsymbol{\theta}^*$. For convenience, we use the shorthand notation $\mathbf{u}$ for $\mathbf{u}\left((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\right)$. Then we have

$$\begin{aligned}\langle \mathbb{E}\mathbf{u}, \boldsymbol{\theta}_\perp \rangle &= \mathbb{E}\left[\sum_{i=1}^m q_i\left(y_1, \ldots, y_m\right) \cdot \langle \mathbf{x}_i, \boldsymbol{\theta}_\perp \rangle\right] = \sum_{i=1}^m \mathbb{E}\left[q_i\left(y_1, \ldots, y_m\right) \cdot \langle \mathbf{x}_i, \boldsymbol{\theta}_\perp \rangle\right] \\ &= \sum_{i=1}^m \mathbb{E}\left[\langle \mathbf{x}_i, \boldsymbol{\theta}_\perp \rangle \cdot \mathbb{E}\left[q_i\left(y_1, \ldots, y_m\right) | \mathbf{x}_1, \ldots, \mathbf{x}_m\right]\right] \qquad (\boldsymbol{*})\end{aligned}$$

As $\mathbf{x}_i$ follows $\mathcal{N}(\mathbf{0}, \mathbf{I})$, $\langle \mathbf{x}_i, \boldsymbol{\theta}^* \rangle$ and $\langle \mathbf{x}_i, \boldsymbol{\theta}_\perp \rangle$ are two zero-mean independent Gaussian random variables. Since the distribution of y_i depends on $\mathbf{x}$ only via $\langle \boldsymbol{\theta}^*, \mathbf{x}_i \rangle$, we can split the expectation and obtain

$$\begin{aligned}(\boldsymbol{*}) &= \sum_{i=1}^m \mathbb{E}\left[\langle \mathbf{x}_i, \boldsymbol{\theta}_\perp \rangle \cdot b_i\left(\langle \boldsymbol{\theta}^*, \mathbf{x}_1 \rangle, \ldots, \langle \boldsymbol{\theta}^*, \mathbf{x}_m \rangle; \boldsymbol{\theta}^*\right)\right] \\ &= \sum_{i=1}^m \mathbb{E}\left[\langle \mathbf{x}_i, \boldsymbol{\theta}_\perp \rangle\right] \cdot \mathbb{E}\left[b_i\left(\langle \boldsymbol{\theta}^*, \mathbf{x}_1 \rangle, \ldots, \langle \boldsymbol{\theta}^*, \mathbf{x}_m \rangle; \boldsymbol{\theta}^*\right)\right] = 0 \ .\end{aligned}$$

Hence $\mathbf{u}$ has to point towards either $\boldsymbol{\theta}^*$ or $-\boldsymbol{\theta}^*$, and note that

$$\begin{aligned}\langle \mathbb{E}\mathbf{u}, \boldsymbol{\theta}^* \rangle &= \sum_{i=1}^m \mathbb{E}\left[q_i\left(y_1, \ldots, y_m\right) \cdot \langle \mathbf{x}_i, \boldsymbol{\theta}^* \rangle\right] \\ &= \sum_{i=1}^m \mathbb{E}\left[b_i\left(\langle \boldsymbol{\theta}^*, \mathbf{x}_1 \rangle, \ldots, \langle \boldsymbol{\theta}^*, \mathbf{x}_m \rangle; \boldsymbol{\theta}^*\right) \cdot \langle \mathbf{x}_i, \boldsymbol{\theta}^* \rangle\right] \\ &= \sum_{i=1}^m \mathbb{E}\left[b_i\left(g_1, \ldots, g_m; \boldsymbol{\theta}^*\right) \cdot g_i\right] = \beta\end{aligned}$$

We complete the proof by recalling that $\|\boldsymbol{\theta}^*\|_2 = 1$, thus $\mathbb{E}\mathbf{u} = \beta\boldsymbol{\theta}^*$. ∎

Note that Lemma 12 is true for all choices of q_i, and the proof is given in the appendix. This lemma presents an insight towards the design of our estimator, that is, the direction of $\boldsymbol{\theta}^*$ can be approximated if we have a good sense about $\mathbb{E}\mathbf{u}$. As we will see in the sequel, the scalar β plays a key role in the estimation error bound, which can give us clues to the choice of q_i. We can assume w.l.o.g. that $\beta \geq 0$ since we can flip the sign of each q_i.

6.2.2 Estimators

Inspired by Lemma 12, we define the vector $\hat{\mathbf{u}}$ for the observed data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$,

$$\hat{\mathbf{u}} = \frac{(n-m)!}{n!} \sum_{\substack{1 \leq i_1, \ldots, i_m \leq n \\ i_1 \neq \ldots \neq i_m}} \mathbf{u}\left((\mathbf{x}_{i_1}, y_{i_1}), \ldots, (\mathbf{x}_{i_m}, y_{i_m})\right) \ , \tag{6.8}$$

which is an unbiased estimator of $\mathbb{E}\mathbf{u}$, meaning that $\mathbb{E}\hat{\mathbf{u}} = \mathbb{E}\mathbf{u} = \beta\boldsymbol{\theta}^*$. For instance, when $m = 2$, we essentially have

$$\hat{\mathbf{u}} = \frac{1}{n(n-1)} \sum_{\substack{1 \leq i,j \leq n \\ i \neq j}} \mathbf{u}\left((\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j)\right) \tag{6.9}$$

In fact, $\hat{\mathbf{u}}$ can be treated as a vector version of U-statistics with order m. Given $\hat{\mathbf{u}}$, a naive way to estimate $\boldsymbol{\theta}^*$ is to simply normalize $\hat{\mathbf{u}}$, i.e., $\hat{\boldsymbol{\theta}} = \hat{\mathbf{u}}/\|\hat{\mathbf{u}}\|_2$, . which is the solution to

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\operatorname{argmin}} \ -\langle \hat{\mathbf{u}}, \boldsymbol{\theta} \rangle \quad \text{s.t.} \quad \boldsymbol{\theta} \in \mathbb{S}^{p-1} \ .$$

In high-dimensional setting, $\boldsymbol{\theta}^*$ is often structured, but the naive estimator fails to take such information into account, which would lead to large error. To incorporate the prior knowledge on $\boldsymbol{\theta}^*$, we design two types of estimator, the constrained one and the regularized one.

Constrained Estimator: If we assume that $\boldsymbol{\theta}^*$ belongs to some structured set $\mathcal{K} \subseteq \mathbb{S}^{p-1}$, then the estimation of $\boldsymbol{\theta}^*$ is carried out via the constrained optimization

$$\hat{\boldsymbol{\theta}}_{\text{cs}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\operatorname{argmin}} \; - \langle \hat{\mathbf{u}}, \boldsymbol{\theta} \rangle \quad \text{s.t.} \quad \boldsymbol{\theta} \in \mathcal{K} \ . \tag{6.10}$$

Similar estimator has been used in [135], but they only focused on specific $\hat{\mathbf{u}}$. Here the set $\mathcal{K}$ can be non-convex, as long as the optimization can be solved globally. Since the objective function is very simple (i.e., linear), we can often end up with a global minimizer. Moreover, the solution to (6.10) also minimizes its relaxed convex counterpart,

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \; - \langle \hat{\mathbf{u}}, \boldsymbol{\theta} \rangle \quad \text{s.t.} \quad \boldsymbol{\theta} \in \operatorname{conv}(\mathcal{K}) \ . \tag{6.11}$$

Although (6.11) may yield a solution outside $\mathcal{K}$, it could help recover a feasible solution for the original problem.

Regularized Estimator: If we assume that the structure of $\boldsymbol{\theta}^*$ can be captured by certain norm $\|\cdot\|$, we may alternatively use the regularized estimator to find $\boldsymbol{\theta}^*$,

$$\hat{\boldsymbol{\theta}}_{\text{rg}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\operatorname{argmin}} \; - \langle \hat{\mathbf{u}}, \boldsymbol{\theta} \rangle + \lambda \|\boldsymbol{\theta}\| \quad \text{s.t.} \quad \|\boldsymbol{\theta}\|_2 \leq 1 \ . \tag{6.12}$$

Previously this estimator was used in 1-bit CS scenario with L_1 norm [186]. The optimization associated with $\hat{\boldsymbol{\theta}}_{\text{rg}}$ is convex, thus the global minimum is always attained. In fact, the following theorem characterizes the solution to (6.12).

Theorem 19 *The regularized estimator* $\hat{\boldsymbol{\theta}}_{rg}$ *in* (6.12) *is given by*

$$\hat{\boldsymbol{\theta}}_{rg} = \begin{cases} \frac{\operatorname{prox}_{\lambda\|\cdot\|}(\hat{\mathbf{u}})}{\left\|\operatorname{prox}_{\lambda\|\cdot\|}(\hat{\mathbf{u}})\right\|_2}, & \textit{if } \lambda < \|\hat{\mathbf{u}}\|_* \\ \\ \mathbf{0}, & \textit{otherwise} \end{cases}, \tag{6.13}$$

where $\operatorname{prox}_{\lambda\|\cdot\|}(\cdot)$ *is the proximal operator for* $\lambda\|\cdot\|$*, and* $\|\cdot\|_*$ *is the dual norm of* $\|\cdot\|$*.*

Remark: When the regularization parameter λ is appropriately chosen, the regularized estimator is the solution to a proximal operator with normalization. The simplicity of the solution makes the computation highly efficient.

6.3 Statistical Analysis

Regarding the constrained estimator, the recovery of $\boldsymbol{\theta}^*$ relies on the geometry of $\hat{\boldsymbol{\theta}}_{\text{cs}}$, which is described by

$$\mathcal{C}_{\mathcal{K}} = \operatorname{cone}\left\{\mathbf{v} \;\middle|\; \mathbf{v} = \boldsymbol{\theta} - \boldsymbol{\theta}^*,\ \boldsymbol{\theta} \in \mathcal{K}\right\} \bigcap \mathbb{S}^{p-1} \tag{6.14}$$

The set $\mathcal{C}_{\mathcal{K}}$ essentially contains all possible directions that error $\boldsymbol{\delta} = \hat{\boldsymbol{\theta}}_{\text{cs}} - \boldsymbol{\theta}^*$ could lie in. The following theorem characterizes the error of $\hat{\boldsymbol{\theta}}$.

Theorem 20 *Suppose that the optimization* (6.10) *can be solved to global minimum. Then the following error bound holds for the minimizer* $\hat{\boldsymbol{\theta}}_{cs}$ *with probability at least* $1 - C'' \exp\left(-w^2\left(\mathcal{C}_{\mathcal{K}}\right)\right)$,

$$\left\|\hat{\boldsymbol{\theta}}_{cs} - \boldsymbol{\theta}^*\right\|_2 \leq \frac{C\kappa m^{\frac{3}{2}}}{\beta} \cdot \frac{w(\mathcal{C}_{\mathcal{K}}) + C'}{\sqrt{n}}, \tag{6.15}$$

where κ *is the sub-Gaussian norm of a standard Gaussian random variable, and* C*,* C'*,*

C'' are all absolute constant.

Remark: Note that estimator is consistent as long as $\beta \neq 0$. The error bound inversely depends on the scale of β, which implies that we should construct suitable q_i such that β is large according to its definition in Lemma 12. The choice of q_i further depends on the assumed property of f^*. Though dependency on m may prevent us from using higher-order $\mathbf{u}$, m is typically small in practice and can be treated as constant.

For regularized estimator, we can similarly establish the recovery guarantee in terms of Gaussian width.

Theorem 21 *Define the following set for any $\rho > 1$,*

$$\mathcal{C}_\rho = \text{cone}\left\{\mathbf{v} \;\middle|\; \|\mathbf{v} + \boldsymbol{\theta}^*\| \leq \|\boldsymbol{\theta}^*\| + \frac{\|\mathbf{v}\|}{\rho}\right\} \bigcap \mathbb{S}^{p-1} \ .$$

Let $\|\cdot\|_*$ be the dual norm of $\|\cdot\|$. If we set $\lambda = \rho \left\|\hat{\mathbf{u}} - \beta\boldsymbol{\theta}^*\right\|_* = O\left(\frac{\rho m^{3/2} w(\Omega)}{\sqrt{n}}\right)$ and it satisfies $\lambda < \|\hat{\mathbf{u}}\|_*$, then with probability at least $1 - C' \exp\left(-w^2(\Omega)\right)$, $\hat{\boldsymbol{\theta}}_{rg}$ in (6.12) satisfies

$$\left\|\hat{\boldsymbol{\theta}}_{rg} - \boldsymbol{\theta}^*\right\|_2 \leq \frac{C(1+\rho)\kappa m^{\frac{3}{2}}}{\beta} \cdot \frac{\Psi \cdot w(\Omega)}{\sqrt{n}} \ , \tag{6.16}$$

where $\Psi = \sup_{\mathbf{v} \in \mathcal{C}_\rho} \|\mathbf{v}\|$, and Ω is the unit ball of norm $\|\cdot\|$.

Remark: The geometry of the regularized estimator is slightly different from the constrained one. Instead of having $\mathcal{C}_\mathcal{K}$, here the set $\mathcal{C}_\rho$ depends on the choice of the regularization parameter λ through the coefficient ρ. There is a tradeoff regarding ρ. A larger ρ results in a larger coefficient in the error bound, while the spherical cap $\mathcal{C}_\rho$ will shrink, which potentially leads to a smaller Ψ. The same phenomenon also appears in the [14].

6.4 Applications

6.4.1 1-bit Compressed Sensing

For 1-bit CS problem (6.2), the $\mathbf{u}$ defined in (6.5) can be chosen with $m = 1$ and $q_i = y_i$, ending up with

$$\mathbf{u}((\mathbf{x}, y)) = y\mathbf{x} \quad \text{and} \quad \hat{\mathbf{u}} = \frac{1}{n}\sum_{i=1}^{n} y_i \mathbf{x}_i \ . \tag{6.17}$$

By such choice of $\mathbf{u}$, the β defined in Lemma 12 is simply $\beta = \mathbb{E}[f^*(g)g]$ with g being standard Gaussian random vector. Under reasonably mild noise, y is likely to take the sign of the linear measurement, which means that $f^*(g)$ should be close to 1 (or -1) if g is positive (or negative). Thus we expect $f^*(g)g$ to be positive most of time and β to be large. Given the choice of $\mathbf{u}$, we can specialize our generalized constrained/regularized estimator to obtain previous results. If $\boldsymbol{\theta}^*$ is assumed to be s-sparse, for constrained estimator, we can choose a straightforward $\mathcal{K} = \{\boldsymbol{\theta} \mid \|\boldsymbol{\theta}\|_0 \leq s\} \cap \mathbb{S}^{p-1}$, which results in the k-support norm estimator [42],

$$\hat{\boldsymbol{\theta}}^{\text{ks}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\text{argmin}} \ -\langle \hat{\mathbf{u}}, \boldsymbol{\theta} \rangle \ \text{ s.t. } \ \|\boldsymbol{\theta}\|_0 \leq s, \|\boldsymbol{\theta}\|_2 = 1 \tag{6.18}$$

Though $\mathcal{K}$ is non-convex, the global minimizer can actually be obtained in closed form,

$$\hat{\theta}_j^{\text{ks}} = \begin{cases} \frac{\hat{u}_j}{\left\| |\hat{\mathbf{u}}|_{1:s}^{\downarrow} \right\|_2} \ , & \text{if } |\hat{u}_j| \text{ is in } |\hat{\mathbf{u}}|_{1:s}^{\downarrow} \\ 0 \ , & \text{otherwise} \end{cases} \tag{6.19}$$

where $|\hat{\mathbf{u}}|^{\downarrow}$ is the absolute-value counterpart of $\hat{\mathbf{u}}$ with entries sorted in descending order, and the subscript takes the top s entries. Similarly if the regularized estimator is instantiated with L_1 norm $\|\cdot\|_1$, we obtain the so-called passive algorithm introduced in [186],

$$\hat{\boldsymbol{\theta}}^{\mathrm{ps}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\operatorname{argmin}} \; -\langle \hat{\mathbf{u}}, \boldsymbol{\theta} \rangle + \lambda \|\boldsymbol{\theta}\|_1 \;\; \text{s.t.} \;\; \|\boldsymbol{\theta}\|_2 \leq 1 \; , \tag{6.20}$$

whose solution is given by the elementwise soft-thresholding operator

$$\hat{\boldsymbol{\theta}}^{\mathrm{ps}} = \frac{S\left(\hat{\mathbf{u}}, \lambda\right)}{\left\|S\left(\hat{\mathbf{u}}, \lambda\right)\right\|_2} \; , \quad \text{where} \;\; S_i(\hat{\mathbf{u}}, \lambda) = \max\{\operatorname{sign}(\hat{u}_i)(|\hat{u}_i| - \lambda), 0\} \; . \tag{6.21}$$

Based on Theorem 20 and 21, we can easily obtain the error bound for both k-support norm estimator and passive algorithm.

Corollary 4 *Assume that* $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ *follow 1-bit CS model in* (6.2) *and* $\hat{\mathbf{u}}$ *is given as* (6.17)*. For any s-sparse $\boldsymbol{\theta}^*$, with high probability, $\hat{\boldsymbol{\theta}}$ produced by both* (6.18) *and* (6.20) *(i.e., $\hat{\boldsymbol{\theta}}^{ks}$ and $\hat{\boldsymbol{\theta}}^{ps}$) satisfy*

$$\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_2 \leq O\left(\sqrt{\frac{s \log p}{n}}\right) \tag{6.22}$$

Proof: For the k-support norm estimator, the cone $\mathcal{C}_{\mathcal{K}}$ is given by

$$\begin{aligned} \mathcal{C}_{\mathcal{K}} &= \operatorname{cone}\left\{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \;\middle|\; \|\hat{\boldsymbol{\theta}}\|_0 \leq s, \|\hat{\boldsymbol{\theta}}\|_2 \leq 1\right\} \bigcap \mathbb{S}^{p-1} \\ \implies \quad & \mathcal{C}_{\mathcal{K}} \subseteq \mathcal{S} = \{\mathbf{v} \mid \|\mathbf{v}\|_0 \leq 2s\} \cap \mathbb{S}^{p-1} \end{aligned}$$

Using (19) from [43], we have

$$w(\mathcal{C}_{\mathcal{K}}) \leq w(\mathcal{S}) \leq O\left(\sqrt{s \log p}\right) \; .$$

By Theorem 20, the error of k-support norm estimator satisfies

$$\left\|\hat{\boldsymbol{\theta}}^{\mathrm{ks}} - \boldsymbol{\theta}^*\right\|_2 \leq O\left(\sqrt{\frac{s\log p}{n}}\right)$$

For the passive algorithm, if we choose $\rho = 2$, the restricted norm compatibility Ψ for L_1 norm satisfies

$$\Psi_{L_1} \leq 4\sqrt{s} \tag{6.23}$$

according to the results in [14, 127]. [43] also show that the Gaussian width of the L_1-norm ball is bounded by

$$w(\Omega_{L_1}) \leq O\left(\sqrt{\log p}\right) \ . \tag{6.24}$$

Now combining (6.23), (6.24) and Theorem 21, we can conclude that

$$\left\|\hat{\boldsymbol{\theta}}^{\mathrm{ps}} - \boldsymbol{\theta}^*\right\|_2 \leq O\left(\sqrt{\frac{s\log p}{n}}\right) \ ,$$

which completes the proof. ■

The above result was shown by [151] and [186], but their analyses do not consider the general structure. Compared with $O\left(\sqrt[4]{\frac{s\log p}{n}}\right)$ yielded by the general result in [134], our bound is much sharper.

6.4.2 A New Estimator for Monotone Transfer

In this subsection, we specifically study model (6.3). Here we further assume that $\tilde{f}$ is *strictly* increasing. What is worth mentioning is that the estimator we develop here can be applied to GLMs as well. To avoid the confusion with $\mathbf{u}$ and $\hat{\mathbf{u}}$ used for 1-bit CS, we instead use new notations $\mathbf{h}$ and $\hat{\mathbf{h}}$ respectively in this monotone transfer setting. To motivate the design of $\mathbf{h}$, it is helpful to rewrite model (6.3) by applying the inverse of $\tilde{f}$ on both sides,

$$\tilde{f}^{-1}(y) = \langle \boldsymbol{\theta}^*, \mathbf{x} \rangle + \epsilon \ . \tag{6.25}$$

Note that the new formulation resembles the linear model except that we have no access to the value of $\tilde{f}^{-1}(y)$. Instead, all we know about $\mathbf{r} = [\tilde{f}^{-1}(y_1), \ldots, \tilde{f}^{-1}(y_n)]^T \in \mathbb{R}^n$ is that it preserves the ordering of $\mathbf{y} = [y_1, \ldots, y_n]^T$. Put in another way, $\mathbf{r}$ needs to satisfy the constraint that $r_i > r_j$ iff. $y_i > y_j$ and $r_i < r_j$ iff. $y_i < y_j$. To move one step further, it is equivalent to $\text{sign}(y_i - y_j) = \text{sign}(r_i - r_j) = \text{sign}(\langle \boldsymbol{\theta}^*, \mathbf{x}_i - \mathbf{x}_j \rangle + \epsilon_i - \epsilon_j)$ based on model assumption. Hence the information contained in sample $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ can be interpreted from the perspective of 1-bit CS, where $\text{sign}(y_i - y_j)$ reflects the perturbed sign of linear measurement $\langle \boldsymbol{\theta}^*, \mathbf{x}_i - \mathbf{x}_j \rangle$. Inspired by the $\mathbf{u}$ for 1-bit CS, we may choose $m = 2$ and define $\mathbf{h}$, $\hat{\mathbf{h}}$ as

$$\mathbf{h}\left((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)\right) = \text{sign}(y_1 - y_2) \cdot (\mathbf{x}_1 - \mathbf{x}_2) \ , \tag{6.26}$$

$$\hat{\mathbf{h}} = \frac{1}{n(n-1)} \sum_{\substack{1 \le i,j \le n \\ i \ne j}} \mathbf{h}\left((\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j)\right) \ , \tag{6.27}$$

Given the definition of $\hat{\mathbf{h}}$, Lemma 12 directly implies the following corollary.

Corollary 5 *Suppose that* $(\mathbf{x}_1, y_2)$ *and* $(\mathbf{x}_2, y_2)$ *are generated by model* (6.3)*, where* $\mathbf{x}_1, \mathbf{x}_2$ *follow Gaussian distribution* $\mathcal{N}(\mathbf{0}, \mathbf{I})$*, and the noise* ϵ_1, ϵ_2 *are independent of* $\mathbf{x}_1, \mathbf{x}_2$ *and identically (but arbitrarily) distributed. Then the expectation of* $\mathbf{h}\left((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)\right)$ *satisfies*

$$\mathbb{E}\left[\mathbf{h}\left((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)\right)\right] = \sqrt{2}\beta' \boldsymbol{\theta}^* \ , \tag{6.28}$$

where $\beta' = \mathbb{E}_{g \sim \mathcal{N}(0,1)}\left[\text{sign}\left(g + (\epsilon_1 - \epsilon_2)/\sqrt{2}\right) \cdot g\right]$.

Remark: The scalar $\sqrt{2}\beta'$ serves as the role of β in Lemma 12, and β' is always guaranteed to be strictly positive regardless how the noise is distributed, which keeps $\boldsymbol{\theta}^*$ distinguishable all the time. To see this, let $\xi = (\epsilon_1 - \epsilon_2)/\sqrt{2}$. Note that ξ is symmetric, thus $\varepsilon\xi$ has the same distribution as ξ, where ε is a Rademacher random variable. Therefore

$$\begin{aligned}\beta' &= \mathbb{E}\left[\text{sign}\left(g+\xi\right)\cdot g\right] = \mathbb{E}_{g,\xi}\mathbb{E}_{\varepsilon}\left[\text{sign}\left(g+\varepsilon\xi\right)\cdot g\right] \\ &= \mathbb{E}_{\xi}\mathbb{E}_{g}\left[\frac{\text{sign}\left(g-\xi\right)+\text{sign}\left(g+\xi\right)}{2}\cdot g\right]\end{aligned}$$

Since $g(g-\xi)+g(g+\xi) = 2g^2 \geq 0$, it follows that $\text{sign}(g(g-\xi)) + \text{sign}(g(g+\xi)) = (\text{sign}(g-\xi)+\text{sign}(g+\xi))\cdot\text{sign}(g) \geq 0$, thus $(\text{sign}(g-\xi)+\text{sign}(g+\xi))\cdot g$ is always nonnegative. Find a large enough $M > 0$ such that $\mathbb{P}(|\xi| \leq M) = 0.5 > 0$, and we have

$$\begin{aligned}\beta' &= \mathbb{E}\left[\text{sign}\left(g+\xi\right)\cdot g\right] \geq \mathbb{E}_{\xi}\mathbb{E}_{g}\left[|g|\cdot\mathbb{I}\{|g|>|\xi|\}\right] \\ &\geq 0.5\mathbb{E}_{g}\left[|g|\cdot\mathbb{I}\{|g|>M\}\right] = \frac{M}{2}\cdot\mathbb{P}(|g|>M) > 0\ .\end{aligned}$$

In the ideal noiseless case, β' achieve its maximum, $\beta'_{\max} = \mathbb{E}[\text{sign}(g)g] = \mathbb{E}[|g|] = \sqrt{2/\pi}$. In the worst case, if ϵ_1 and ϵ_2 are heavy-tailed and dominate g, then $\beta' \approx \mathbb{E}\left[\text{sign}\left((\epsilon_1-\epsilon_2)/\sqrt{2}\right)\cdot g\right] \approx 0$.

Now we can instantiate the generalized estimator based on $\hat{\mathbf{h}}$. For example, if $\boldsymbol{\theta}^*$ is s-sparse, we estimate it by

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\text{argmin}}\ -\langle\hat{\mathbf{h}},\boldsymbol{\theta}\rangle \ \text{ s.t. } \ \|\boldsymbol{\theta}\|_0 \leq s, \|\boldsymbol{\theta}\|_2 = 1 \tag{6.29}$$

which enjoys $O\left(\sqrt{\frac{s\log p}{n}}\right)$ error rate as shown in Corollary 4. The regularized estimator can also be obtained with the same $\hat{\mathbf{h}}$ according to (6.20). The bottleneck of computing $\hat{\boldsymbol{\theta}}$ lies in the calculation of $\hat{\mathbf{h}}$. A simple proposition below enables us to get $\hat{\mathbf{h}}$ in a fast manner.

Proposition 15 *Given* $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$*, let* $\pi^{\downarrow}$ *be the permutation of* $\{1, \ldots, n\}$ *such that* $y_{\pi_1^{\downarrow}} > y_{\pi_2^{\downarrow}} > \ldots > y_{\pi_n^{\downarrow}}$*. Then we have*

$$\hat{\mathbf{h}} = \frac{2}{n(n-1)} \sum_{i=1}^{n} (n+1-2i) \cdot \mathbf{x}_{\pi_i^{\downarrow}} \tag{6.30}$$

Proof: We rearrange the terms inside the summation of (6.27) based on $\pi^{\downarrow}$,

$$\begin{aligned} \hat{\mathbf{h}} &= \frac{1}{n(n-1)} \sum_{\substack{1 \le i,j \le n \\ i \ne j}} \operatorname{sign}(y_i - y_j) \cdot (\mathbf{x}_i - \mathbf{x}_j) = \frac{2}{n(n-1)} \sum_{\substack{1 \le i,j \le n \\ i \ne j}} \operatorname{sign}(y_i - y_j) \cdot \mathbf{x}_i \\ &= \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j \ne \pi_i^{\downarrow}} \operatorname{sign}\left(y_{\pi_i^{\downarrow}} - y_j\right) \cdot \mathbf{x}_{\pi_i^{\downarrow}} = \frac{2}{n(n-1)} \sum_{i=1}^{n} (n+1-2i) \cdot \mathbf{x}_{\pi_i^{\downarrow}} \ , \end{aligned}$$

where the last inequality uses the fact that there are $(i-1)$ y_j larger than and $(n-i)$ smaller than $y_{\pi_i^{\downarrow}}$, thus $\sum_{j \ne \pi_i^{\downarrow}} \operatorname{sign}\left(y_{\pi_i^{\downarrow}} - y_j\right) = (n-i) - (i-1) = n+1-2i$. ■

Remark: Based on the proposition above, $\hat{\mathbf{h}}$ can be efficiently computed in $O(np + n\log n)$ time, i.e., $O(n \log n)$ time for sorting $\mathbf{y}$ and $O(np)$ time for the weighted sum of all $\mathbf{x}_i$. This is a significant improvement compared with the the naive calculation using (6.27), which takes $O(n^2 p)$ time.

6.4.3 Other Parameter Structures

So far we have illustrated the Gaussian width based error bounds, viz (6.15) and (6.16), only through unstructured sparsity of $\boldsymbol{\theta}^*$. Here we provide two more examples, non-overlapping group sparsity and fused sparsity.

Non-Overlapping Group Sparsity: Suppose the coordinates of $\boldsymbol{\theta}^*$ has been partitioned into K predefined disjoint groups $\mathcal{I}_1, \ldots, \mathcal{I}_K \subseteq \{1, 2, \ldots, p\}$ with $|\mathcal{I}_j| \le G$

$(1 \leq j \leq K)$, out of which only k groups are non-zero. If we use the regularized estimator with $L_{2,1}$ norm $\|\boldsymbol{\theta}\|_{2,1} = \sum_{j=1}^{K} \|\boldsymbol{\theta}_{\mathcal{I}_j}\|_2$, the optimal solution can be similarly obtained as (6.20), with elementwise soft-thresholding replaced by the groupwise one,

$$\hat{\boldsymbol{\theta}}_{\text{rg}} = \frac{GS\left(\hat{\mathbf{u}}, \lambda\right)}{\left\|GS\left(\hat{\mathbf{u}}, \lambda\right)\right\|_2} , \quad \text{where } \; GS_{\mathcal{I}}(\hat{\mathbf{u}}, \lambda) = \max\left\{1 - \frac{\lambda}{\left\|\hat{\mathbf{u}}_{\mathcal{I}}\right\|_2}, 0\right\} \cdot \hat{\mathbf{u}}_{\mathcal{I}} \; . \tag{6.31}$$

The related geometric measures that appears in (6.16) can be found in [14], which are given by

$$\Psi_{L_{2,1}} \leq O(\sqrt{k}) \tag{6.32}$$

$$w\left(\Omega_{L_{2,1}}\right) \leq O(\sqrt{\log K} + \sqrt{G}) \; , \tag{6.33}$$

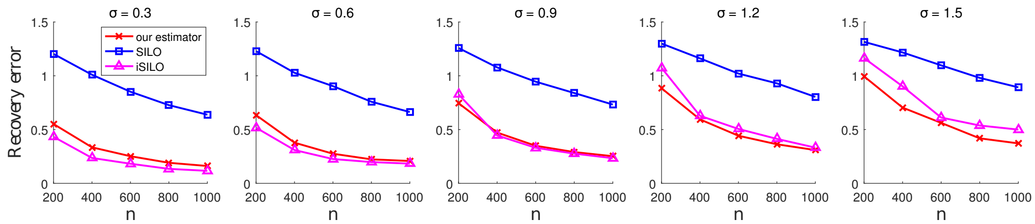
Fused Sparsity: $\boldsymbol{\theta}^*$ is said to be *s-fused-sparse* if the cardinality of the set $\mathcal{F}(\boldsymbol{\theta}^*) = \{1 \leq i < p \mid \theta^*_i \neq \theta^*_{i+1}\}$ is smaller than s. If we resort to the constrained estimator (6.10) with $\mathcal{K} = \{\boldsymbol{\theta} \mid |\mathcal{F}(\boldsymbol{\theta})| \leq s, \; \|\boldsymbol{\theta}\|_2 = 1\}$, the associated optimization can be solved by dynamic programming [18]. The proposition below upper bounds the corresponding Gaussian width $w(\mathcal{C}_{\mathcal{K}})$ in (6.15).

Proposition 16 *For s-fused-sparse $\boldsymbol{\theta}^*$, the Gaussian width of set $\mathcal{C}_{\mathcal{K}}$ with $\mathcal{K} = \{\boldsymbol{\theta} \in \mathbb{R}^p \mid |\mathcal{F}(\boldsymbol{\theta})| \leq s, \; \|\boldsymbol{\theta}\|_2 = 1\}$ satisfies*

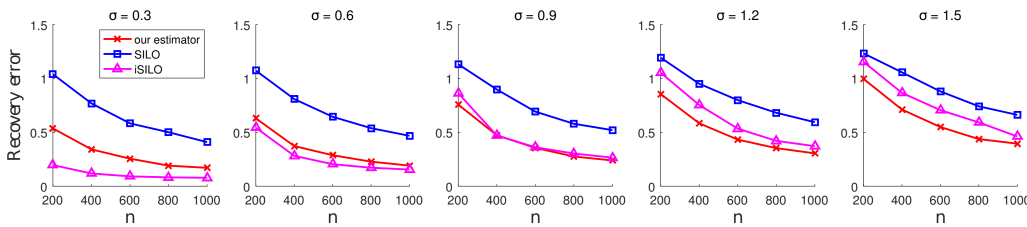
$$w(\mathcal{C}_{\mathcal{K}}) \leq O(\sqrt{s \log p}) \tag{6.34}$$

6.5 Experimental Results

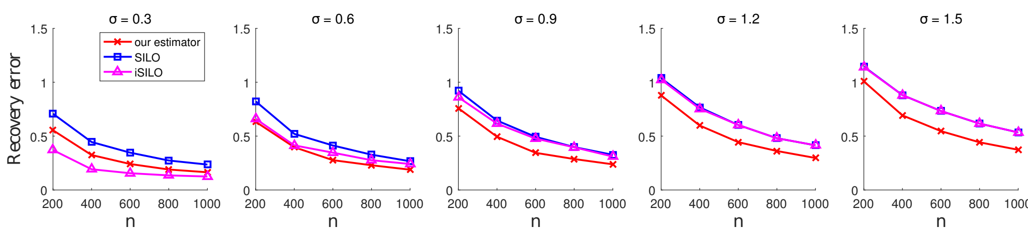
In the experiment, we focus on model (6.3) with sparse $\boldsymbol{\theta}^*$. The problem dimension is fixed as $p = 1000$, and the sparsity of $\boldsymbol{\theta}^*$ is set to 10. Essentially we generate our data



(a) Error for $\tilde{f}(z) = 1/(1 + \exp(-z))$



(b) Error for $\tilde{f}(z) = \log(1 + \exp(z))$



(c) Error for $\tilde{f}(z) = z^3$

Figure 6.1: Recovery error vs. sample size. (a) Our estimator has similar performance compared with iSILO, both of which outperform SILO by a large margin. (b) iSILO has smaller error when σ is small, while our estimator works better in high-noise regime (c) The error of SILO is reduced compared with other $\tilde{f}$, but iSILO fails to give further improvement over SILO when σ is large. Our estimator still outperforms them when $\sigma \geq 0.6$.

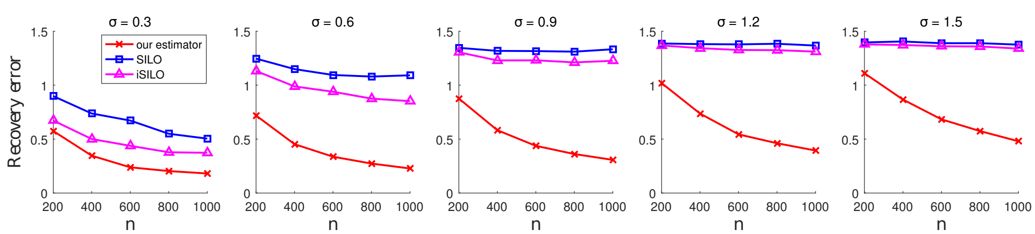


Figure 6.2: Recovery error vs. sample size, with $\tilde{f}(z) = z^3$ under heavy-tailed noise

$(\mathbf{x}, y)$ from

$$y = \tilde{f}\left(\langle \boldsymbol{\theta}^*, \mathbf{x} \rangle + \epsilon\right) \ ,$$

where $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. σ ranges from 0.3 to 1.5. We choose three monotonically increasing $\tilde{f}$, $\tilde{f}(z) = 1/(1 + \exp(-z))$ (which is bounded and Lipschitz), $\tilde{f}(z) = z^3$ (which is unbounded and non-Lipschitz), and $\tilde{f}(z) = \log(1 + \exp(z))$ (which is unbounded but Lipschitz). The sample size n varies from 200 to 1000. We use the estimator (6.29) in Section 6.4.2. The baselines we compare with is the SILO and iSILO algorithm introduced in [58]. SILO does not quite take the monotonicity in account. In fact, it is the special case of our generalized constrained estimator which uses the same choice of $\mathbf{u}$ as 1-bit CS. The original SILO use the constraint set $\{\boldsymbol{\theta} \mid \|\boldsymbol{\theta}\|_1 \leq \sqrt{s},\ \|\boldsymbol{\theta}\|_2 \leq 1\}$, which is computationally less efficient and statistically no better than $\mathcal{K} = \{\boldsymbol{\theta} \mid \|\boldsymbol{\theta}\|_0 \leq s\} \cap \mathbb{S}^{p-1}$ [42, 186]. Hence we also use $\mathcal{K}$ in SILO for a fair comparison. iSILO relies on a specific implementation of isotonic regression which explicitly restricts the Lipschitz constant of $\tilde{f}$ to be one. To fit iSILO into our setting, we remove the Lipschitzness constraint and perform the standard isotonic regression. Since the convergence is not guaranteed for the iterative procedure of iSILO, the number of its iterations is fixed to 100. The best tuning parameter of iSILO is obtained by grid search.

The experiment results are shown in Figure 6.1. Overall the iSILO algorithm works well under small noise, while our estimator has better performance when the variance of noise increases. To better demonstrate the robustness of our estimator to heavy-tailed noise, instead of Gaussian noise, we sample ϵ from the Student's t distribution with degrees of freedom equal to 3. We repeat the experiments for $\tilde{f}(z) = z^3$, and obtain the plots in Figure 6.2. We can see that the error of our estimator consistently decreases for all choice of σ as n increases. For SILO and iSILO, the errors are relatively large, and unable to shrink for large σ even when more data are provided.

Appendix

Appendix 6.A Supplementary Proofs

6.A.1 Proof of Theorem 19

Proof: For $\lambda \geq \|\hat{\mathbf{u}}\|_*$, by Hölder's inequality, the objective function of (6.12) satisfies

$$-\langle \hat{\mathbf{u}}, \boldsymbol{\theta} \rangle + \lambda \|\boldsymbol{\theta}\| \geq -\langle \hat{\mathbf{u}}, \boldsymbol{\theta} \rangle + \|\hat{\mathbf{u}}\|_* \|\boldsymbol{\theta}\| \geq 0 \ ,$$

and we can easily verify that $\mathbf{0}$ is a solution. When $\lambda < \|\hat{\mathbf{u}}\|_*$, the minimum of (6.12) is negative, thus the optimal solution is always obtained at the boundary of the constraint, i.e., $\|\hat{\boldsymbol{\theta}}_{\mathrm{rg}}\|_2 = 1$. For this case, we construct the Langrangian of (6.12) and swap the minimization and maximization step,

$$\max_{\beta \geq 0} \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \ -\langle \hat{\mathbf{u}}, \boldsymbol{\theta} \rangle + \lambda \|\boldsymbol{\theta}\| + \beta(\|\boldsymbol{\theta}\|_2^2 - 1) \ .$$

The minimization step over $\boldsymbol{\theta}$ can be equivalently written as

$$\begin{aligned}
& \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \ -\langle \hat{\mathbf{u}}, \boldsymbol{\theta} \rangle + \lambda \|\boldsymbol{\theta}\| + \beta(\|\boldsymbol{\theta}\|_2^2 - 1) + \frac{1}{4\beta} \|\hat{\mathbf{u}}\|_2^2 \\
\Longleftrightarrow \quad & \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \ \beta \|\boldsymbol{\theta}\|_2^2 - \langle \hat{\mathbf{u}}, \boldsymbol{\theta} \rangle + \frac{1}{4\beta} \|\hat{\mathbf{u}}\|_2^2 + \lambda \|\boldsymbol{\theta}\| \\
\Longleftrightarrow \quad & \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \ \beta \left\| \boldsymbol{\theta} - \frac{\hat{\mathbf{u}}}{2\beta} \right\|_2^2 + \lambda \|\boldsymbol{\theta}\| \\
\Longleftrightarrow \quad & \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \ \frac{1}{2} \|2\beta\boldsymbol{\theta} - \hat{\mathbf{u}}\|_2^2 + \lambda \|2\beta\boldsymbol{\theta}\| \quad \Longrightarrow \quad 2\beta\hat{\boldsymbol{\theta}}_{\mathrm{rg}} = \mathrm{prox}_{\lambda\|\cdot\|}(\hat{\mathbf{u}})
\end{aligned}$$

As we have shown that $\|\hat{\boldsymbol{\theta}}_{\text{rg}}\|_2 = 1$ for $\lambda < \|\hat{\mathbf{u}}\|_*$, $\hat{\boldsymbol{\theta}}_{\text{rg}}$ must be the normalized version of the proximal operator of $\hat{\mathbf{u}}$ for $\lambda\|\cdot\|$, which completes the proof. ■

6.A.2 Proof of L_2-Error Bound

We prove Theorem 20 and Theorem 21 here. To show them, we need a Hoeffding-type inequality for sub-Gaussian U-statistics. In the literature, most of the studies are centered around *bounded* U-statistics, for which the celebrated concentration is established by [72]. Yet it is not easy to locate the counterpart for sub-Gaussian case. Therefore we provide the following result and attach a proof.

Lemma 13 (concentration for sub-Gaussian U-statistics) *Define the U-statistic*

$$U_{n,m}(h) = \frac{(n-m)!}{n!} \sum_{\substack{1 \le i_1, \ldots, i_m \le n \\ i_1 \neq i_2 \neq \ldots \neq i_m}} h\left(\mathbf{z}_{i_1}, \ldots, \mathbf{z}_{i_m}\right) \tag{6.35}$$

with order m and kernel $h : \mathbb{R}^{d \times m} \mapsto \mathbb{R}$ based on n independent copies of random vector $\mathbf{z} \in \mathbb{R}^d$, denoted by $\mathbf{z}_1, \cdots, \mathbf{z}_n$. If $h(\cdot, \ldots, \cdot)$ is sub-Gaussian with $\|h\|_{\psi_2} \le \kappa$, then the following inequality holds for $U_{n,m}(h)$ with any $\delta > 0$,

$$\mathbb{P}\left(|U_{n,m}(h) - \mathbb{E}U_{n,m}(h)| > \delta\right) \le 2\exp\left(-C\left\lfloor\frac{n}{m}\right\rfloor \cdot \frac{\delta^2}{\kappa^2}\right) , \tag{6.36}$$

in which C is an absolute constant.

Proof: Our proof is based on Hoeffding's decomposition for U-statistics. For simplicity, we use U as shorthand for $U_{n,m}(h)$. Given a permutation π of $\{1, \ldots, n\}$, define

$$W_\pi = \frac{1}{\left\lfloor\frac{n}{m}\right\rfloor} \sum_{k=0}^{\left\lfloor\frac{n}{m}\right\rfloor - 1} h\left(\mathbf{z}_{\pi_{mk+1}}, \ldots, \mathbf{z}_{\pi_{m(k+1)}}\right) ,$$

The U-statistic can be rewritten as $U = \frac{1}{n!}\sum_\pi W_\pi$, and the summation is over all possible permutations of $\{1, \ldots, n\}$. As no copy of $\mathbf{z}$ appears more than twice in a single W_π, W_π is an average of $\lfloor \frac{n}{m} \rfloor$ independent sub-Gaussian random variables. Hence the ψ_2-norm of its centered version satisfies $\|W_\pi - \mathbb{E}W_\pi\|_{\psi_2} \leq c\kappa/\sqrt{\lfloor \frac{n}{m} \rfloor}$. Using Chernoff technique, we have for any $t > 0$,

$$
\begin{aligned}
\mathbb{P}\left(U - \mathbb{E}U > \delta\right) &\leq e^{-t\delta} \cdot \mathbb{E}\left[\exp(t(U - \mathbb{E}U))\right] \\
&= e^{-t\delta} \cdot \mathbb{E}\left[\exp\left(\frac{t}{n!}\sum_\pi (W_\pi - \mathbb{E}U)\right)\right] \\
&\leq e^{-t\delta} \cdot \mathbb{E}\left[\frac{1}{n!}\sum_\pi \exp\left(t(W_\pi - \mathbb{E}U)\right)\right] \qquad (6.37) \\
&= e^{-t\delta} \cdot \mathbb{E}\left[\exp\left(t(W_\pi - \mathbb{E}W_\pi)\right)\right] \\
&\leq \exp\left(-t\delta + ct^2 \cdot \frac{\kappa^2}{\lfloor \frac{n}{m} \rfloor}\right) ,
\end{aligned}
$$

where the second inequality is obtained via Jensen's inequality and the last one follows the moment generating function bound for centered sub-Gaussian random variable. Choosing $t = \lfloor \frac{n}{m} \rfloor \delta / 2c\kappa^2$ to minimize right-hand side of (6.37), we obtain

$$
\mathbb{P}\left(U - \mathbb{E}U > \delta\right) \leq \exp\left(-C\left\lfloor \frac{n}{m} \right\rfloor \cdot \frac{\delta^2}{\kappa^2}\right) ,
$$

where $C = 1/2c$. To complete the proof, we just need to repeat the argument above for $\mathbb{P}\left(U - \mathbb{E}U < -\delta\right)$. ■

Proof of Theorem 20: As $\hat{\boldsymbol{\theta}}$ attains the global minimum of (6.10), we have

$$
\langle \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*, \hat{\mathbf{u}} \rangle \geq 0 \iff \left\langle \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*, \frac{\hat{\mathbf{u}}}{\beta} - \boldsymbol{\theta}^* + \boldsymbol{\theta}^* \right\rangle \geq 0
$$

$$
\Longrightarrow \langle \hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^* \rangle \geq 1 - \left\langle \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*, \frac{\hat{\mathbf{u}}}{\beta} - \boldsymbol{\theta}^* \right\rangle \geq 1 - \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \cdot \sup_{\mathbf{v} \in \mathcal{C}_\mathcal{K} \cup \{\mathbf{0}\}} \left\langle \mathbf{v}, \frac{\hat{\mathbf{u}}}{\beta} - \boldsymbol{\theta}^* \right\rangle
$$

In order to bound the supremum above, we use the result from generic chaining. We define the stochastic process $\{Z_{\mathbf{v}} = \langle \mathbf{v}, \frac{\hat{\mathbf{u}}}{\beta} - \boldsymbol{\theta}^* \rangle\}_{\mathbf{v} \in \mathcal{C}_{\mathcal{K}} \cup \{\mathbf{0}\}}$. First, we need to check the process has sub-Gaussian incremental. For simplicity, we denote $\mathbf{u}\left((\mathbf{x}_{i_1}, y_{i_1}), \ldots, (\mathbf{x}_{i_m}, y_{i_m})\right)$ by $\mathbf{u}_{i_1,\ldots,i_m}$. By the definitions and properties of sub-Gaussian norm, the sub-Gaussian norm of $\mathbf{u}_{i_1,\ldots,i_m}$ satisfies

$$
\begin{aligned}
\|\mathbf{u}_{i_1,\ldots,i_m}\|_{\psi_2} &= \sup_{\mathbf{v} \in \mathbb{S}^{p-1}} \left\| \sum_{j=1}^{m} q_j \left(y_{1_1}, \ldots, y_{i_m}\right) \cdot \langle \mathbf{x}_j, \mathbf{v} \rangle \right\|_{\psi_2} \leq \sup_{\mathbf{v} \in \mathbb{S}^{p-1}} \left\| \sum_{j=1}^{m} |\langle \mathbf{x}_j, \mathbf{v} \rangle| \right\|_{\psi_2} \\
&\leq m \cdot \sup_{\mathbf{v} \in \mathbb{S}^{p-1}} \||\langle \mathbf{x}_j, \mathbf{v} \rangle|\|_{\psi_2} \leq \kappa m \ ,
\end{aligned}
$$

thus we know $\|\langle \mathbf{u}_{i_1,\ldots,i_m}, \mathbf{v} - \mathbf{w} \rangle\|_{\psi_2} \leq \kappa m \cdot \|\mathbf{v} - \mathbf{w}\|_2$. By Lemma 13, we have

$$
\begin{aligned}
&\mathbb{P}\left(|Z_{\mathbf{v}} - Z_{\mathbf{w}}| > \delta\right) \\
&= \mathbb{P}\left(\left| \left\langle \mathbf{v} - \mathbf{w}, \frac{\hat{\mathbf{u}}}{\beta} - \boldsymbol{\theta}^* \right\rangle \right| > \delta \right) \\
&= \mathbb{P}\left(\left| \frac{(n-m)!}{n!} \sum_{\substack{1 \leq i_1, \ldots, i_m \leq n \\ i_1 \neq \ldots \neq i_m}} \frac{1}{\beta} \cdot \langle \mathbf{u}_{i_1,\ldots,i_m}, \mathbf{v} - \mathbf{w} \rangle - \langle \mathbf{v} - \mathbf{w}, \boldsymbol{\theta}^* \rangle \right| > \delta \right) \\
&\leq 2 \exp\left(-C \left\lfloor \frac{n}{m} \right\rfloor \cdot \frac{\beta^2 \delta^2}{m^2 \kappa^2 \cdot \|\mathbf{v} - \mathbf{w}\|_2^2} \right) \\
&\leq 2 \exp\left(-C' \cdot \frac{n \beta^2 \delta^2}{m^3 \kappa^2 \cdot \|\mathbf{v} - \mathbf{w}\|_2^2} \right) \ ,
\end{aligned}
$$

where we set $C' = C/2$. Therefore we can conclude that $\{Z_{\mathbf{v}}\}$ has sub-Gaussian incremental w.r.t. the metric $s(\mathbf{v}, \mathbf{w}) \triangleq \frac{\kappa m^{\frac{3}{2}} \cdot \|\mathbf{v} - \mathbf{w}\|_2}{\beta \sqrt{n}}$. Now applying Lemma 2 to $\{Z_{\mathbf{v}}\}$, we obtain

$$\mathbb{P}\left(\sup_{\mathbf{v},\mathbf{w}\in\mathcal{C}_{\mathcal{K}}\cup\{\mathbf{0}\}}|Z_{\mathbf{v}}-Z_{\mathbf{w}}|\geq C_1\Big(\gamma_2\left(\mathcal{C}_{\mathcal{K}}\cup\{\mathbf{0}\},s\right)+\delta\cdot\text{diam}\left(\mathcal{C}_{\mathcal{K}}\cup\{\mathbf{0}\},s\right)\Big)\right)$$

$$\leq C_2\exp\left(-\delta^2\right)$$

$$\Longrightarrow\qquad \mathbb{P}\left(\sup_{\mathbf{v}\in\mathcal{C}_{\mathcal{K}}\cup\{\mathbf{0}\}}|Z_{\mathbf{v}}|\geq\frac{C_1\kappa m^{\frac{3}{2}}}{\beta\sqrt{n}}\cdot\left(\gamma_2\left(\mathcal{C}_{\mathcal{K}}\cup\{\mathbf{0}\},\|\cdot\|_2\right)+2\delta\right)\right)\leq C_2\exp\left(-\delta^2\right)$$

Using Lemma 4 $\gamma_2\left(\mathcal{C}_{\mathcal{K}}\cup\{\mathbf{0}\},\|\cdot\|_2\right)\leq C_0\cdot w\left(\mathcal{C}_{\mathcal{K}}\cup\{\mathbf{0}\}\right)$ and taking $\delta=w\left(\mathcal{C}_{\mathcal{K}}\cup\{\mathbf{0}\}\right)$, we get

$$\sup_{\mathbf{v}\in\mathcal{C}_{\mathcal{K}}\cup\{\mathbf{0}\}}\left\langle\mathbf{v},\frac{\hat{\mathbf{u}}}{\beta}-\boldsymbol{\theta}^*\right\rangle\leq\sup_{\mathbf{v}\in\mathcal{C}_{\mathcal{K}}\cup\{\mathbf{0}\}}|Z_{\mathbf{v}}|\leq\frac{C_3\kappa m^{\frac{3}{2}}}{\beta\sqrt{n}}\cdot w\left(\mathcal{C}_{\mathcal{K}}\cup\{\mathbf{0}\}\right)$$

$$\leq\frac{C_3\kappa m^{\frac{3}{2}}}{\beta}\cdot\frac{w\left(\mathcal{C}_{\mathcal{K}}\right)+C_4}{\sqrt{n}}$$

with probability at least $1-C_2\exp\left(-w^2\left(\mathcal{C}_{\mathcal{K}}\right)\right)$. The last inequality follows from Lemma 1. Now we turn to the quantity $\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2$,

$$\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2^2\leq 2-2\langle\hat{\boldsymbol{\theta}},\boldsymbol{\theta}^*\rangle\leq 2-2\left(1-\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2\cdot\frac{C_3\kappa m^{\frac{3}{2}}}{\beta}\cdot\frac{w\left(\mathcal{C}_{\mathcal{K}}\right)+C_4}{\sqrt{n}}\right)$$

$$\leq\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2\cdot\frac{2C_3\kappa m^{\frac{3}{2}}}{\beta}\cdot\frac{w\left(\mathcal{C}_{\mathcal{K}}\right)+C_4}{\sqrt{n}}\ .$$

We finish the proof by letting $C=2C_3$, $C'=C_4$ and $C''=C_2$. ∎

Proof of Theorem 21: Based on the optimality of $\hat{\boldsymbol{\theta}}$, we have

$$-\langle\hat{\mathbf{u}},\hat{\boldsymbol{\theta}}\rangle+\lambda\|\hat{\boldsymbol{\theta}}\|\leq-\langle\hat{\mathbf{u}},\boldsymbol{\theta}^*\rangle+\lambda\|\boldsymbol{\theta}^*\|$$

$$\Longrightarrow\qquad\langle\beta\boldsymbol{\theta}^*-\hat{\mathbf{u}}-\beta\boldsymbol{\theta}^*,\hat{\boldsymbol{\theta}}\rangle+\lambda\|\hat{\boldsymbol{\theta}}\|\leq\langle\beta\boldsymbol{\theta}^*-\hat{\mathbf{u}}-\beta\boldsymbol{\theta}^*,\boldsymbol{\theta}^*\rangle+\lambda\|\boldsymbol{\theta}^*\|$$

$$\Longrightarrow\qquad\beta(1-\langle\boldsymbol{\theta}^*,\hat{\boldsymbol{\theta}}\rangle)\leq\langle\hat{\mathbf{u}}-\beta\boldsymbol{\theta}^*,\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\rangle+\lambda(\|\boldsymbol{\theta}^*\|-\|\hat{\boldsymbol{\theta}}\|)$$

Since $\langle \boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}} \rangle \leq 1$, we have

$$\begin{aligned}
\langle \hat{\mathbf{u}} - \beta \boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \rangle + \lambda \left(\|\boldsymbol{\theta}^*\| - \|\hat{\boldsymbol{\theta}}\| \right) \geq 0 \qquad \Longrightarrow \\
\|\hat{\boldsymbol{\theta}}\| \leq \|\boldsymbol{\theta}^*\| + \frac{1}{\lambda} \cdot \langle \hat{\mathbf{u}} - \beta \boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \rangle \\
\leq \|\boldsymbol{\theta}^*\| + \frac{1}{\lambda} \cdot \|\hat{\mathbf{u}} - \beta \boldsymbol{\theta}^*\|_* \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| \\
= \|\boldsymbol{\theta}^*\| + \frac{1}{\rho} \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| \qquad \Longrightarrow \qquad \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \in \mathcal{C}_\rho
\end{aligned}$$

Therefore it follows that

$$\begin{aligned}
1 - \langle \boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}} \rangle &\leq \langle \frac{\hat{\mathbf{u}}}{\beta} - \boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \rangle + \frac{\lambda}{\beta} \left(\|\boldsymbol{\theta}^*\| - \|\hat{\boldsymbol{\theta}}\| \right) \\
&\leq \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \left(\left\| \frac{\hat{\mathbf{u}}}{\beta} - \boldsymbol{\theta}^* \right\|_* \cdot \frac{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|}{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2} + \frac{\lambda}{\beta} \cdot \frac{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|}{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2} \right) \\
&\leq (1+\rho) \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \cdot \left\| \frac{\hat{\mathbf{u}}}{\beta} - \boldsymbol{\theta}^* \right\|_* \cdot \sup_{\mathbf{v} \in \mathcal{C}_\rho} \|\mathbf{v}\| \\
&= (1+\rho) \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \cdot \left\| \frac{\hat{\mathbf{u}}}{\beta} - \boldsymbol{\theta}^* \right\|_* \cdot \Psi
\end{aligned} \tag{6.38}$$

Now we try to bound $\left\| \frac{\hat{\mathbf{u}}}{\beta} - \boldsymbol{\theta}^* \right\|_*$. We first rewrite it as $\left\| \frac{\hat{\mathbf{u}}}{\beta} - \boldsymbol{\theta}^* \right\|_* = \sup_{\mathbf{v} \in \Omega} \left\langle \frac{\hat{\mathbf{u}}}{\beta} - \boldsymbol{\theta}^*, \mathbf{v} \right\rangle$. Construct the stochastic process $\{Z_{\mathbf{v}} = \langle \mathbf{v}, \hat{\mathbf{u}}/\beta - \boldsymbol{\theta}^* \rangle\}_{\mathbf{v} \in \Omega}$, and it is not difficult to verify that $\{Z_{\mathbf{v}}\}$ has sub-Gaussian incremental using the proof in Theorem 20. Now applying Lemma 2 and 4, we have

$$\sup_{\mathbf{v} \in \Omega} \left\langle \frac{\hat{\mathbf{u}}}{\beta} - \boldsymbol{\theta}^*, \mathbf{v} \right\rangle = \frac{1}{2} \cdot \sup_{\mathbf{v}, \mathbf{w} \in \Omega} |Z_{\mathbf{v}} - Z_{\mathbf{w}}| \leq \frac{C_1 \kappa m^{\frac{3}{2}}}{\beta} \cdot \frac{w(\Omega)}{\sqrt{n}} \ , \tag{6.39}$$

with probability at least $1 - C' \exp\left(-w^2(\Omega)\right)$. Therefore we know that λ satisfies

$$\lambda = O\left(\frac{\rho m^{3/2} w(\Omega)}{\sqrt{n}} \right)$$

As indicated by Theorem 19, if $\lambda < \|\hat{\mathbf{u}}\|_*$, we can assert that $\|\hat{\boldsymbol{\theta}}\|_2 = 1$. Combining (6.38) and (6.39), we finally get

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| = \frac{2 - 2\langle\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*\rangle}{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|} \leq \frac{Cm\kappa(1+\rho)}{\beta} \cdot \frac{\Psi \cdot w\,(\Omega)}{\sqrt{n}}\ ,$$

where the equality uses the fact that $\|\hat{\boldsymbol{\theta}}\|_2 = 1$. ■

6.A.3 Proof of Proposition 16

Proof: Define the following sets

$$\mathcal{T} = \bigcup_{i\leq j} \mathcal{T}_{i,j}\ , \quad \text{where} \tag{6.40}$$

$$\mathcal{T}_{i,j} = \left\{ \alpha\mathbf{u} \in \mathbb{R}^p \;\middle|\; |\alpha| \leq \sqrt{2s+1},\ u_i = \ldots = u_j = \frac{1}{\sqrt{j-i+1}}, \right. \\ \left. u_k = 0\,(k < i \text{ or } \ k < j) \right\} \tag{6.41}$$

For each $\mathcal{T}_{i,j}$, its Gaussian width can be calculated as

$$w(\mathcal{T}_{i,j}) = \mathbb{E}\left[\sup_{\mathbf{v}\in\mathcal{T}_{i,j}} \langle\mathbf{v}, \mathbf{g}\rangle\right] = \sqrt{2s+1} \cdot \mathbb{E}\,[|\langle\mathbf{u}, \mathbf{g}\rangle|] = \sqrt{2s+1} \cdot \mathbb{E}\,|g| = O(\sqrt{2s+1})\ ,$$

where $\mathbf{u}$ is defined in (6.41) and g is a standard Gaussian random variable. We apply Lemma 1 to $\mathcal{T}$, and obtain

$$\begin{aligned} w(\mathcal{T}) &\leq \max_{i\leq j} w(\mathcal{T}_{i,j}) + 2\sup_{\mathbf{z}\in\mathcal{T}} \|\mathbf{z}\|_2 \sqrt{\log\left(\binom{p}{2} + p\right)} \\ &\leq O(\sqrt{2s+1}) + O(\sqrt{2s+1} \cdot \sqrt{\log p}) \\ &= O(\sqrt{s\log p}) \end{aligned}$$

Next we show that $\mathcal{C}_{\mathcal{K}} \subseteq \text{conv}(\mathcal{T})$. Since $\mathcal{K} = \{\boldsymbol{\theta} \mid |\mathcal{F}(\boldsymbol{\theta})| \leq s, \ \|\boldsymbol{\theta}\|_2 = 1\}$ and $\mathcal{C}_{\mathcal{K}} = \text{cone}\left\{\mathbf{v} \;\middle|\; \mathbf{v} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*, \ \hat{\boldsymbol{\theta}} \in \mathcal{K}\right\} \bigcap \mathbb{S}^{p-1}$ by definition, we have $|\mathcal{F}(\mathbf{v})| \leq 2s$ for any $\mathbf{v} \in \mathcal{C}_{\mathcal{K}}$. Suppose $|\mathcal{F}(\mathbf{v})| = t \leq 2s$ and $\mathcal{F}(\mathbf{v}) = \{i_1, i_2, \ldots, i_t\}$. For simplicity, we also let $i_0 = 0$ and $i_{t+1} = p$. Then any $\mathbf{v} \in \mathcal{C}_{\mathcal{K}}$ can be written as a convex combination of $t+2$ points in $\mathcal{T}$. To see this, we rewrite $\mathbf{v}$ as

$$\mathbf{v} = \sum_{r=0}^{t} \mathbf{v}_{i_r+1:i_{r+1}} = \sum_{r=0}^{t} \frac{\|\mathbf{v}_{i_r+1:i_{r+1}}\|_2}{\sqrt{t+1}} \cdot \frac{\sqrt{t+1}\mathbf{v}_{i_r+1:i_{r+1}}}{\|\mathbf{v}_{i_r+1:i_{r+1}}\|_2} + \left(1 - \sum_{r=0}^{t} \frac{\|\mathbf{v}_{i_r+1:i_{r+1}}\|_2}{\sqrt{t+1}}\right) \cdot \mathbf{0} \ , \tag{6.42}$$

where $\mathbf{v}_{i_r+1:i_{r+1}}$ is obtained from $\mathbf{v}$ by keeping the entries from index i_r+1 to i_{r+1} while zeroing out the rest. Let $\mathbf{u}_{i_r+1:i_{r+1}} = \frac{\sqrt{t+1}\mathbf{v}_{i_r+1:i_{r+1}}}{\|\mathbf{v}_{i_r+1:i_{r+1}}\|_2}$, and we have

$$\|\mathbf{u}_{i_r+1:i_{r+1}}\|_2 = \sqrt{t+1} \leq \sqrt{2s+1} \quad \Longrightarrow \quad \mathbf{u}_{i_r+1:i_{r+1}} \in \mathcal{T}_{i_r+1:i_{r+1}} \subseteq \mathcal{T} \ .$$

It follows from $\|\mathbf{v}\|_2 = 1$ that

$$\sum_{r=0}^{t} \frac{\|\mathbf{v}_{i_r+1:i_{r+1}}\|_2}{\sqrt{t+1}} \leq \frac{\sqrt{(t+1)\sum_{r=0}^{t} \|\mathbf{v}_{i_r+1:i_{r+1}}\|_2^2}}{\sqrt{t+1}} = 1$$

$$\Longrightarrow \quad 1 - \sum_{r=0}^{t} \frac{\|\mathbf{v}_{i_r+1:i_{r+1}}\|_2}{\sqrt{t+1}} \geq 0$$

Hence (6.42) is indeed a convex combination of $t+2$ points in $\mathcal{T}$, which implies $\mathcal{C}_{\mathcal{K}} \subseteq \text{conv}(\mathcal{T})$. Finally, by the properties of Gaussian width, we conclude that

$$w(\mathcal{C}_{\mathcal{K}}) \leq w(\text{conv}(\mathcal{T})) = w(\mathcal{T}) \leq O(\sqrt{s \log p})$$

■

Chapter 7

Sparse Linear Isotonic Models

7.1 Introduction

As discussed in Chapter 6, despite the prevalent success of linear models, modern data often arise from complex environments in which the linear correlation could break down, leading to poor performance of linear models. Similar to the single-index model that captures the nonlinearity in response, progress has been made to relax the stringent assumption of linear models by allowing nonlinearity in features. In particular, [9] consider the following *additive isotonic models* (AIMs),

$$y = \sum_{j=1}^{p} f_j(x_j) + \epsilon \ , \tag{7.1}$$

where $\{f_j\}_{j=1}^{p} \triangleq \mathcal{F}$ is a set of *monotone* univariate functions. To estimate $\mathcal{F}$, a commonly-used procedure is *cyclic pooled adjacent violators* (CPAV). At each iteration of CPAV, *isotonic regression* is called to estimate one f_j and its solution can be efficiently found by the *pooled adjacent violators algorithm* (PAVA) [16]. Though the nonlinearity can be captured by $\mathcal{F}$, one need to specify the monotonicity for each f_j

(either increasing or decreasing) in advance, which could be unknown in real-world applications, and enumerating all possible combinations can be computationally prohibitive. In high dimension, the estimation of $\mathcal{F}$ becomes even more challenging, because the number of monotone functions is very large.

To address the challenges in AIMs, we propose the *sparse linear isotonic models* (SLIMs) for the high-dimensional setting, which assume

$$\mathbb{E}\left[y|\mathbf{x}\right] = \sum_{j=1}^{p} \tilde{\theta}_j f_j(x_j) = \left\langle \tilde{\boldsymbol{\theta}}, \mathbf{f}(\mathbf{x}) \right\rangle \ , \tag{7.2}$$

where $\mathbf{f}(\mathbf{x}) \triangleq \tilde{\mathbf{x}} = [f_1(x_1), \ldots, f_p(x_p)]^T$. SLIMs combine the parametric form from the sparse linear models with the monotone transformations from AIMs, and generalize the assumption of additive noise ϵ to the conditional expectation form $\mathbb{E}[y|\mathbf{x}]$. Throughout the chapter, the parameter $\tilde{\boldsymbol{\theta}}$ is assumed to be *s-sparse*. For identifiability, we also assume w.l.o.g. that each f_j is monotonically increasing (as the monotonicity can be flipped by changing signs of $\tilde{\theta}_j$), and properly normalized such that every $\tilde{x}_j = f_j(x_j)$ is zero-mean and unit-variance. Note that without losing any representational power of AIM, the assumption of increasing f_j avoids the pre-specification of monotonicity for each f_j as required in (7.1). For such hybrid model, given n i.i.d. samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, our goal is to estimate both $\tilde{\boldsymbol{\theta}}$ and $\mathcal{F}$. Since the hidden predictor $\tilde{\mathbf{x}}$ is inaccessible, brutally fitting data into a linear model could result in a poor estimate of $\tilde{\boldsymbol{\theta}}$. In this work, we design a two-step algorithm to accomplish this goal, which estimates $\tilde{\boldsymbol{\theta}}$ followed by $\mathcal{F}$. The estimation of $\tilde{\boldsymbol{\theta}}$ is inspired by the *rank-based* approaches for structure learning of graphical models. At the high level, those approaches do not rely on the exact values of samples generated from the graphical model, in order to learn its structure. Instead they resort to rank correlations (e.g., *Kendall's tau correlation* [94]) that are invariant under monotonically increasing transformation, so that observing $\mathbf{x}$ and $\tilde{\mathbf{x}}$ makes no difference to the method. By leveraging a similar idea, we propose the Kendall's tau Dantzig selector (KDS) to estimate $\tilde{\boldsymbol{\theta}}$, with certain Kendall's tau correlation coefficients appropriately plugged in. Under some distributional assumptions, we show that this estimator is guaranteed to recover a normalized version of $\tilde{\boldsymbol{\theta}}$ with small error. After $\tilde{\boldsymbol{\theta}}$ is estimated, we have a CPAV-type algorithm tailored for estimating transformations $\mathcal{F}$, which efficiently extends CPAV at little cost.

To sum up, we highlight a few merits of SLIM as follows. First, as aforementioned, SLIM need not specify the monotonicity of f_j whereas AIM requires. Second, the two-step estimation for SLIM is particularly useful in high-dimensional settings. The estimation of $\tilde{\boldsymbol{\theta}}$ may identify many "don't-care" f_j's as their corresponding $\tilde{\theta}_j$'s are zero, thus reducing the problem size of estimating $\mathcal{F}$. Besides, estimating $\tilde{\boldsymbol{\theta}}$ will suffice if one only focuses on variable selection.

For the ease of exposition, we introduce a few notations which will be used in the rest of the chapter. We let $\mathbf{y} = [y_1, y_2, \ldots, y_n]^T \in \mathbb{R}^n$ be the response vector, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times p}$ be the observed design matrix , and denote its columns by $\mathbf{x}^j \in \mathbb{R}^n$. Similarly $\tilde{\mathbf{X}}$, $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{x}}^j$ will denote the hidden counterpart of $\mathbf{X}$, $\mathbf{x}_i$ and $\mathbf{x}^j$. Matrix is bold capital, and the corresponding bold lowercase is reserved for its rows (columns) with suitable subscripts (superscripts), and its entries are plain lowercase with subscripts indexing both row and column. In general, vectors are bold lowercase while scalars are plain lowercase. For a matrix, $\|\cdot\|_2$ denotes its spectral norm (i.e. the largest eigenvalue) and $\|\cdot\|_{\max}$ denotes the value of the largest entry in magnitude. The rest of the chapter is organized as follows: we first review the related work in Section 7.2, and provide an overview of estimation for SLIM in Section 7.3. Next we analyze the recovery of $\tilde{\boldsymbol{\theta}}$ and present the algorithmic details for estimating $\mathcal{F}$ in Section 7.4. In Section 7.5, we demonstrate the effectiveness of SLIM through experiments.

7.2 Related Work

AIM was initially proposed in [9]. [117] established the asymptotic properties of the CPAV procedure. The high-dimensional counterpart of AIMs (i.e., assuming most of f_j's are zero), Lasso ISO (LISO), was studied by [53], where a modified CPAV is used to achieved the sparsity in $\mathcal{F}$. [49] considered a semiparametric additive isotonic model by introducing an additional parametric model into (7.1). On the other hand, [71] considered an additive model of the same form as (7.1) for general $\mathcal{F}$. With suitable smoothing operator on f_j's, a coordinate descent procedure called *backfitting* can be applied to estimating $\mathcal{F}$. In high-dimensional regime , [140] correspondingly investigated the sparse additive models (SpAMs), which is solved a backfitting algorithm with extra soft-thresholding steps. Many other efforts have been spent by relying on the smoothness of f_j's, including [106], [122], [74], and etc.

The method we use to estimate $\tilde{\boldsymbol{\theta}}$ is closely related to the high-dimensional structure learning of graphical models. For sparse Gaussian graphical model, [123] proposed a neighborhood selection procedure for estimating the graph structure, which iteratively regresses each variable against the rest via Lasso. The neighborhood Dantzig selector [182] shares the similar spirit with this approach, which switches Lasso to Dantzig selector. Recent progress has shown that these approaches continue to work for some non-Gaussian distributions, such as *nonparanormal distribution* [110], by using rank correlations to approximate the latent correlation matrix [108, 178]. Similar results have been further generalized to *transelliptical distribution* [69, 70, 109].

7.3 Overview of Two-Step Algorithm

In this section, we present an overview of the two-step algorithm for the estimation of SLIM, which first estimates $\tilde{\boldsymbol{\theta}}$ and then $\mathcal{F}$.

For the estimation of $\tilde{\boldsymbol{\theta}}$, if the hidden design matrix $\tilde{\mathbf{X}}$ could be observed, Dantzig selector [32] can be used to estimate $\tilde{\boldsymbol{\theta}}$ as normal linear models,

$$\hat{\boldsymbol{\theta}}_{\text{orc}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\operatorname{argmin}} \|\boldsymbol{\theta}\|_1 \quad \text{s.t.} \quad \left\| \frac{1}{n}\tilde{\mathbf{X}}^T \left(\tilde{\mathbf{X}}\boldsymbol{\theta} - \mathbf{y}\right) \right\|_\infty \leq \gamma_n \ , \tag{7.3}$$

where γ_n is a tuning parameter. A key observation from (7.3) is that instead of exactly knowing $\tilde{\mathbf{X}}$ and $\mathbf{y}$, it is sufficient to be given the (approximate) value of $\frac{\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}}{n}$ and $\frac{\tilde{\mathbf{X}}^T\mathbf{y}}{n}$ in order for (7.3) to work. Note that the quantity $\frac{\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}}{n}$ and its expectation $\tilde{\boldsymbol{\Sigma}} = \mathbb{E}[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^T]$ also arise in the structure learning of nonparanormal graphical models. Specifically if $\tilde{\mathbf{x}}$ follows a multivariate Gaussian $\mathcal{N}(\mathbf{0}, \tilde{\boldsymbol{\Sigma}})$, then the observed predictor $\mathbf{x}$, represented as $\mathbf{f}^{-1}(\tilde{\mathbf{x}}) \triangleq [f_1^{-1}(\tilde{x}_1), \ldots, f_1^{-1}(\tilde{x}_p)]^T$, is by definition a nonparanormal distribution $NPN(\tilde{\boldsymbol{\Sigma}}, \mathbf{f}^{-1})$, in which $\tilde{\boldsymbol{\Sigma}}$ is often called *latent correlation matrix*. Simply speaking, the nonparanormal distribution models the random vector whose coordinates are element-wise monotone transformations of a Gaussian random vector. To estimate $\tilde{\boldsymbol{\Sigma}}$ without knowing $\mathbf{f}$ or $\mathbf{f}^{-1}$, Kendall's tau correlation coefficient [94] plays a key role in rank-based methods. Given data $\mathbf{X} = [x_{ij}] \in \mathbb{R}^{n\times p}$, we define the *sample Kendall's tau correlation matrix* $\hat{\mathbf{T}} = [\hat{t}_{ij}] \in \mathbb{R}^{p\times p}$ as

$$\hat{t}_{ij} = \sum_{1\leq k,k'\leq n} \frac{\operatorname{sign}((x_{ki} - x_{k'i})(x_{kj} - x_{k'j}))}{n(n-1)} \ , \tag{7.4}$$

and its transformed version $\hat{\boldsymbol{\Sigma}} = [\hat{\sigma}_{ij}] \in \mathbb{R}^{p\times p}$,

$$\hat{\sigma}_{ij} = \sin\left(\frac{\pi}{2}\hat{t}_{ij}\right) , \tag{7.5}$$

One straightforward yet critical property of $\hat{\mathbf{T}}$ and $\hat{\boldsymbol{\Sigma}}$ is the invariance to monotone increasing transformations on columns of $\mathbf{X}$, indicating that the two quantities remain unchanged if $\mathbf{X}$ is replaced by $\tilde{\mathbf{X}}$ in the definitions. More importantly, later analysis will reveal for the class of transelliptical distributions (a generalization of nonparanormal distribution) the closeness between the transformed sample Kendall's tau correlation matrix $\hat{\mathbf{\Sigma}}$ and the latent correlation matrix $\tilde{\mathbf{\Sigma}}$, thus $\hat{\mathbf{\Sigma}}$ can serve as an approximation to $\frac{\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}}{n}$ as $\frac{\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}}{n} \approx \tilde{\mathbf{\Sigma}}$ in expectation. For $\frac{\tilde{\mathbf{x}}^T\mathbf{y}}{n}$ and its expectation $\tilde{\boldsymbol{\beta}} = \mathbb{E}[y\tilde{\mathbf{x}}] = \tilde{\mathbf{\Sigma}}\tilde{\boldsymbol{\theta}}$, we similarly define the *sample Kendall's tau correlation vector* $\hat{\mathbf{b}} \in \mathbb{R}^p$ and its transformation $\hat{\boldsymbol{\beta}}$

$$\hat{b}_j = \sum_{1 \le k,k' \le n} \frac{\text{sign}((x_{kj} - x_{k'j})(y_k - y_{k'}))}{n(n-1)} \ , \tag{7.6}$$

$$\hat{\beta}_j = \sin\left(\frac{\pi}{2}\hat{b}_j\right) \ , \tag{7.7}$$

and use $\hat{\boldsymbol{\beta}}$ as a replacement for $\frac{\tilde{\mathbf{x}}^T\mathbf{y}}{n}$. Therefore the estimation of $\tilde{\boldsymbol{\theta}}$ can proceed with (7.3) by replacing $\frac{\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}}{n}$ and $\frac{\tilde{\mathbf{X}}^T\mathbf{y}}{n}$ with $\hat{\mathbf{\Sigma}}$ and $\hat{\boldsymbol{\beta}}$ respectively, which leads to the following estimator which we call *Kendall's tau Dantzig selector* (KDS),

$$\check{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\text{argmin}} \ \ \|\boldsymbol{\theta}\|_1 \quad \text{s.t.} \quad \left\| \hat{\mathbf{\Sigma}}\boldsymbol{\theta} - \hat{\boldsymbol{\beta}} \right\|_\infty \le \gamma_n \ . \tag{7.8}$$

Later it will be shown in the analysis that the $\check{\boldsymbol{\theta}}$ only approximates the direction of $\tilde{\boldsymbol{\theta}}$, and the scale should be attached on the final estimate $\hat{\boldsymbol{\theta}}$ by calculating the sample variance of $\mathbf{y}$.

To estimate the transformations $\mathcal{F}$, one needs to first find out an $\hat{\mathbf{X}} = [\hat{x}_{ij}]$ that approximates the hidden design $\tilde{\mathbf{X}} = [f_j(x_{ij})]$ for the observed $\mathbf{X} = [x_{ij}]$, which essentially gives us the estimated values of each f_j at n points $x_{1j}, \ldots, x_{nj}$. To be specific, we fit $\hat{\mathbf{X}}$ into $\mathbf{y}$ and the estimated $\hat{\boldsymbol{\theta}}$ through the following convex program,

$$\begin{aligned} \hat{\mathbf{X}} \ &= \ \underset{\mathbf{Z} \in \mathbb{R}^{n \times p}}{\text{argmin}} \ \ \frac{1}{2}\|\mathbf{Z}\hat{\boldsymbol{\theta}} - \mathbf{y}\|_2^2 \\ &\text{s.t.} \ \ \mathbf{z}^j \in \mathcal{M}(\mathbf{x}^j), \ \ \mathbf{1}^T\mathbf{z}^j = 0, \ \ \|\mathbf{z}^j\|_2 \le \sqrt{n}, \ \ \forall \ 1 \le j \le p, \end{aligned} \tag{7.9}$$

Algorithm 7 Estimating $\tilde{\boldsymbol{\theta}}$ and $\mathcal{F}$ for SLIM

Input: $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{y} \in \mathbb{R}^n$, tuning parameter γ_n
Output: Estimated $\hat{\boldsymbol{\theta}}$ for $\tilde{\boldsymbol{\theta}}$ and $\hat{\mathcal{F}}$ for $\mathcal{F}$
1: Compute the transformed sample Kendall's tau correlation matrix $\hat{\boldsymbol{\Sigma}}$ and vector $\hat{\boldsymbol{\beta}}$ using (7.4) - (7.7)
2: Estimate $\check{\boldsymbol{\theta}}$ via Kendall's tau Dantzig selector (7.8)
3: $\hat{\boldsymbol{\theta}} := \hat{\sigma}_y \check{\boldsymbol{\theta}}$, where $\hat{\sigma}_y$ is the sample variance of $\mathbf{y}$
4: Estimate the hidden design $\hat{\mathbf{X}}$ via (7.9)
5: $\hat{\mathcal{F}} := \{\hat{f}_j\}_{j=1}^p$, where $\hat{f}_j$ is given by (7.10)
6: **Return** $\hat{\boldsymbol{\theta}}$ and $\hat{\mathcal{F}}$

where $\mathcal{M}(\mathbf{x}) = \{\mathbf{v} \mid v_i \geq v_j \text{ iff } x_i \geq x_j, \ \forall\ 1 \leq i, j \leq p\}$. In order to get the f_j defined everywhere, we need to interpolate the n estimated points $\hat{x}_{1j}, \ldots, \hat{x}_{nj}$. In the algorithm, we simply use nearest-neighbor interpolation as follows,

$$\hat{f}_j(x) = \sum_{i=1}^{n} \hat{x}_{ij} \cdot \mathbb{I}\Big\{ i = \underset{1 \leq k \leq n}{\operatorname{argmin}} \ |x_{kj} - x| \Big\} \ , \tag{7.10}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function that outputs one if the predicate is true and zero otherwise. Other interpolation technique, e.g., linear/spline interpolation, can be applied in the need of certain desired properties of f_j. The full estimation algorithm is given in Algorithm 7.

7.4 Statistical and Algorithmic Analysis

In this section, we detail the Algorithm 7 in several aspects. We analyze the recovery guarantee of the KDS for estimating $\tilde{\boldsymbol{\theta}}$. Under the assumption of transelliptically distributed $(\mathbf{x}, y)$ and the so-called *sign sub-Gaussian condition*, we show that the sample complexity of KDS can be sharpened compared with [108, 178]. To estimate $\mathcal{F}$, we present a CPAV-type algorithm for solving (7.31), where each step can be solved almost at no more cost than isotonic regression.

7.4.1 Recovery Guarantee of $\tilde{\boldsymbol{\theta}}$

In this subsection, we consider the estimation of $\tilde{\boldsymbol{\theta}}$. The KDS (7.8) can be casted as a linear program, which can be solved efficiently by many optimization algorithms [41,99]. Hence we focus on the statistical aspect of KDS. From Section 7.3, we know that the success of KDS relies on $\hat{\mathbf{\Sigma}}$ and $\hat{\boldsymbol{\beta}}$, which replace $\frac{\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}}{n}$ and $\frac{\tilde{\mathbf{X}}^T\mathbf{y}}{n}$ in the Dantzig selector (7.3). Hence we first investigate the property of $\hat{\mathbf{\Sigma}}$ and $\hat{\boldsymbol{\beta}}$. The definition (7.4) - (7.7) are sample versions of (transformed) Kendall's tau correlation matrix and vector. Here we define their population counterparts.

Definition 17 Given $(\mathbf{x}, y)$ and its independent copy $(\mathbf{x}', y')$, the population Kendall's tau correlation matrix $\mathbf{T} = [t_{ij}] \in \mathbb{R}^{p\times p}$ and vector $\mathbf{b} \in \mathbb{R}^p$ are defined as

$$t_{ij} = \mathbb{P}\left((x_i - x_i')(x_j - x_j') > 0\right) - \mathbb{P}\left((x_i - x_i')(x_j - x_j') < 0\right) \ , \tag{7.11}$$

$$b_j = \mathbb{P}\left((x_j - x_j')(y - y') > 0\right) - \mathbb{P}\left((x_j - x_j')(y - y') < 0\right) \ , \tag{7.12}$$

and their transformed versions $\mathbf{\Sigma} = [\sigma_{ij}] \in \mathbb{R}^{p\times p}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$ are given by

$$\sigma_{ij} = \sin\left(\frac{\pi}{2}t_{ij}\right) \ , \quad \beta_j = \sin\left(\frac{\pi}{2}b_j\right) \ . \tag{7.13}$$

To specify the statistical assumptions, we first introduce two family of distributions, *elliptical* and *transelliptical*. The transelliptical distribution is defined based on the elliptical distribution given as follows.

Definition 18 (elliptical distribution) A random vector $\mathbf{z} \in \mathbb{R}^p$ follows an elliptical distribution $EC(\boldsymbol{\mu}, \tilde{\mathbf{\Sigma}}, \xi)$ iff $\mathbf{z}$ has a stochastic representation:

$$\mathbf{z} \sim \boldsymbol{\mu} + \xi \mathbf{A}\mathbf{u} \ . \tag{7.14}$$

Here $\boldsymbol{\mu} \in \mathbb{R}^p$, $q \triangleq \text{rank}(\mathbf{A})$, $\mathbf{A} \in \mathbb{R}^{p \times q}$, $\xi \geq 0$ is a random variable independent of $\mathbf{u}$, $\mathbf{u} \in \mathbb{S}^{q-1}$ is uniformly distributed on the unit sphere in $\mathbb{R}^q$, and $\mathbf{A}\mathbf{A}^T = \tilde{\mathbf{\Sigma}}$. Note that

$$\mathbb{E}[\mathbf{z}] = \boldsymbol{\mu} \ , \qquad \text{Cov}[\mathbf{z}] = \frac{\mathbb{E}[\xi^2]}{q} \tilde{\mathbf{\Sigma}} \ . \tag{7.15}$$

This family of distribution contains the Gaussian distribution as a special case, and more details can be found in [52]. The extension from elliptical to transelliptical distribution parallels that from normal to nonparanormal distribution.

Definition 19 (transelliptical distribution) A random vector $\mathbf{x} \in \mathbb{R}^p$ is said to follow the transelliptical distribution $TE(\tilde{\mathbf{\Sigma}}, \xi, \mathbf{f})$ if $\mathbf{f}(\mathbf{x}) = [f_1(x_1), f_2(x_2), \ldots, f_p(x_p)]^T \sim EC(\boldsymbol{\mu}, \tilde{\mathbf{\Sigma}}, \xi)$, where $f_1, f_2, \ldots f_p$ are all strictly increasing functions, $\boldsymbol{\mu} = \mathbf{0}$, $\text{diag}(\tilde{\mathbf{\Sigma}}) = \mathbf{I}$, and $\mathbb{P}(\xi = 0) = 0$.

The conditions on $\boldsymbol{\mu}$ and $\text{diag}(\tilde{\mathbf{\Sigma}})$ are imposed for identifiability. If the underlying elliptical distribution is multivariate Gaussian, then the transelliptical family is reduced to the nonparanormal. We refer the readers to [109] for more discussions on transelliptical distribution. Based on the elliptical and transelliptical family, we introduce our assumptions on distribution of $(\mathbf{x}, y)$:

- $\mathbf{x} \in \mathbb{R}^p$ follows transelliptical distribution $TE(\tilde{\mathbf{\Sigma}}, \xi, \mathbf{f})$, or equivalently $\tilde{\mathbf{x}}$ follows elliptical distribution $EC(\mathbf{0}, \tilde{\mathbf{\Sigma}}, \xi)$, where $\mathbb{E}[\xi^2] = p$.
- The smallest eigenvalue $\lambda_{\min}$ of $\tilde{\mathbf{\Sigma}}$ is strictly positive.
- $\tilde{\mathbf{x}}$ and y are jointly elliptically distributed.

The assumption $\mathbb{E}[\xi^2] = p$ is also out of the consideration of identifiability. The last assumption on the joint distribution of $(\tilde{\mathbf{x}}, y)$ may seem obscure. But it can be satisfied, for example, when $\mathbf{x}$ is nonparanormal and y is a noisy observation of $\langle \tilde{\boldsymbol{\theta}}, \mathbf{f}(\mathbf{x}) \rangle$

perturbed by an additive zero-mean Gaussian noise. Under these assumptions, we have the following recovery guarantee for the KDS $\check{\boldsymbol{\theta}}$.

Theorem 22 *For any s-sparse $\tilde{\boldsymbol{\theta}}$, if we set $\gamma_n = \frac{5\pi}{\sqrt{\lambda_{\min}}}\sqrt{\frac{s\log p}{n}}$ and $n \geq \left(\frac{24\pi}{\lambda_{\min}}\right)^2 s^2 \log p$, with probability at least $1 - \frac{2}{p} - \frac{1}{p^{2.5}}$, $\hat{\boldsymbol{\theta}}$ given by (7.8) satisfies*

$$\left\|\check{\boldsymbol{\theta}} - \frac{\tilde{\boldsymbol{\theta}}}{\sigma_y}\right\|_2 \leq \frac{40\pi}{\lambda_{\min}^{3/2}}\sqrt{\frac{s^2\log p}{n}} \ , \tag{7.16}$$

In the theorem above, though KDS only approximates a normalized version of $\tilde{\boldsymbol{\theta}}$, the scale σ_y can be estimated by computing the sample variance $\hat{\sigma}_y^2$ of $\mathbf{y}$, and the final estimate of $\tilde{\boldsymbol{\theta}}$ is $\hat{\boldsymbol{\theta}} = \hat{\sigma}_y\check{\boldsymbol{\theta}}$ as given in Algorithm 7.

To prove the theorem above, we need to characterize certain properties of $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\beta}}$. One notable result that has been shown for $\boldsymbol{\Sigma}$, $\hat{\boldsymbol{\Sigma}}$ and $\tilde{\boldsymbol{\Sigma}}$ is given in the following lemma.

Lemma 14 *For $\mathbf{x} \sim TE(\tilde{\boldsymbol{\Sigma}}, \xi, \mathbf{f})$, the transformed population Kendall's tau correlation matrix $\boldsymbol{\Sigma}$ satisfies*

$$\boldsymbol{\Sigma} = \tilde{\boldsymbol{\Sigma}} \ , \tag{7.17}$$

and the sample version $\hat{\boldsymbol{\Sigma}}$ for $\boldsymbol{\Sigma}$ defined in (7.5), with probability at least $1 - p^{-2.5}$, satisfies

$$\|\hat{\boldsymbol{\Sigma}} - \tilde{\boldsymbol{\Sigma}}\|_{\max} \leq 3\pi\sqrt{\frac{\log p}{n}} \tag{7.18}$$

The lemma is essentially Theorem 3.2 and 4.1 in [70]. Similarly we have the following lemma for $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$.

Lemma 15 *The transformed population Kendall's tau correlation vector $\boldsymbol{\beta}$ satisfies*

$$\boldsymbol{\beta} = \frac{\tilde{\boldsymbol{\beta}}}{\sigma_y} = \frac{\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{\theta}}}{\sigma_y} \ , \tag{7.19}$$

where σ_y^2 is the variance of y. The transformed sample Kendall's tau correlation vector $\hat{\boldsymbol{\beta}}$, with probability at least $1-\frac{2}{p}$, satisfies

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_\infty \leq 2\pi\sqrt{\frac{\log p}{n}} \tag{7.20}$$

Proof: By definition, $\tilde{\boldsymbol{\beta}} = \mathbb{E}[y\tilde{\mathbf{x}}] = \mathbb{E}_{\tilde{\mathbf{x}}}[\tilde{\mathbf{x}}\cdot\mathbb{E}_y[y|\tilde{\mathbf{x}}]] = \mathbb{E}[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^T\tilde{\boldsymbol{\theta}}] = \tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{\theta}}$. Given that $\lambda_{\min} > 0$ and the properties of elliptical distribution (7.15), we have $\mathbb{E}[\tilde{\mathbf{x}}] = \mathbf{0}$, $\text{rank}(\mathbf{A}) = \text{rank}(\tilde{\boldsymbol{\Sigma}}) = p$ and $\text{Cov}[\tilde{\mathbf{x}}] = \tilde{\boldsymbol{\Sigma}}$. Since $\tilde{\mathbf{x}}$, y are jointly elliptical and $\boldsymbol{\beta}$ is invariant to $\mathbf{f}$, using Theorem 2 in [107], we have for each β_j,

$$\beta_j = \frac{\mathbb{E}[y\tilde{x}_j]-\mathbb{E}[y]\mathbb{E}[\tilde{x}_j]}{\sqrt{\text{Var}[y]}\sqrt{\text{Var}[\tilde{x}_j]}} = \frac{\mathbb{E}\left[\langle\tilde{\boldsymbol{\theta}},\tilde{\mathbf{x}}\rangle\cdot\tilde{x}_j\right]}{\sqrt{\text{Var}[y]}} = \frac{\langle\tilde{\boldsymbol{\theta}},\tilde{\boldsymbol{\sigma}}_j\rangle}{\sigma_y} ,$$

which implies (7.19). Using Hoeffding's inequality for U-statistics [72], we have for each β_j and $\hat{\beta}_j$

$$\mathbb{P}\left(\left|\beta_j-\hat{\beta}_j\right|\geq\epsilon\right) \leq \mathbb{P}\left(\left|b_j-\hat{b}_j\right|\geq\frac{2}{\pi}\epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{2\pi^2}\right) .$$

Letting $\epsilon = 2\pi\sqrt{\frac{\log p}{n}}$ and taking union bound, we obtain

$$\mathbb{P}\left(\left\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right\|_\infty \geq 2\pi\sqrt{\frac{\log p}{n}}\right) \leq \frac{2}{p} ,$$

which completes the proof. ■

In the light of Lemma 14 and 15, it becomes clear that $\frac{\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}}{n}$ and $\frac{\tilde{\mathbf{X}}^T\mathbf{y}}{n}$ in (7.3) are replaced by $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\beta}}$ in (7.8). The population counterpart of $\hat{\boldsymbol{\Sigma}}$ is $\boldsymbol{\Sigma} = \tilde{\boldsymbol{\Sigma}} = \mathbb{E}[\frac{\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}}{n}]$. Unfortunately, the population version $\boldsymbol{\beta}$ of $\hat{\boldsymbol{\beta}}$ is not equal to $\tilde{\boldsymbol{\beta}} = \mathbb{E}[\frac{\tilde{\mathbf{X}}^T\mathbf{y}}{n}]$, which is additionally normalized by σ_y. Therefore we will see later that KDS recovers a scaled $\tilde{\boldsymbol{\theta}}$.

In order to bound the estimation error, we additionally need to show that the transformed sample Kendall's tau correlation matrix $\hat{\mathbf{\Sigma}}$ satisfies the following *restricted eigenvalue* (RE) condition [14, 21, 127, 139, 190], which is critical in the recovery analysis.

Lemma 16 *Define the error spherical cap for any s-sparse vector $\boldsymbol{\theta}^* \in \mathbb{R}^p$,*

$$\mathcal{C} = \{\mathbf{v} \in \mathbb{R}^p \mid \ \|\boldsymbol{\theta}^* + \mathbf{v}\|_1 \leq \|\boldsymbol{\theta}^*\|_1\} \cap \mathbb{S}^{p-1} \ . \tag{7.21}$$

If $\mathbf{x} \sim TE(\tilde{\mathbf{\Sigma}}, \xi, \mathbf{f})$ and $n \geq \left(\frac{24\pi}{\lambda_{\min}}\right)^2 s^2 \log p = O(s^2 \log p)$, with probability at least $1 - p^{-2.5}$, the following RE condition holds for $\hat{\mathbf{\Sigma}}$ and $\mathcal{C}$,

$$\inf_{\mathbf{v} \in \mathcal{C}} \mathbf{v}^T \hat{\mathbf{\Sigma}} \mathbf{v} \geq \frac{\lambda_{\min}}{2} \ , \tag{7.22}$$

where $\lambda_{\min}$ is the smallest eigenvalue of $\tilde{\mathbf{\Sigma}}$.

Remark: The proof is given in the appendix. Similar proof steps appear in [178] amid the analysis of rank-based neighborhood Dantzig selector, in which the concept of RE condition is not formulated. Here we single out this lemma in order for the later comparison in Section 7.4.2.

Now we are ready to present the proof of Theorem 22.

Proof of Theorem 22: For the sake of convenience, we denote $\boldsymbol{\theta}^* = \frac{\tilde{\boldsymbol{\theta}}}{\sigma_y}$, and it is easy to see that $\tilde{\mathbf{\Sigma}}\boldsymbol{\theta}^* = \boldsymbol{\beta}$. We first show that $\boldsymbol{\theta}^*$ is feasible when $\gamma_n = \frac{5\pi}{\sqrt{\lambda_{\min}}}\sqrt{\frac{s \log p}{n}}$, by bounding the left-hand side of the constraint for $\boldsymbol{\theta}^*$.

$$\begin{aligned}
\left\|\hat{\mathbf{\Sigma}}\boldsymbol{\theta}^* - \hat{\boldsymbol{\beta}}\right\|_\infty &= \left\|\left(\hat{\mathbf{\Sigma}} - \tilde{\mathbf{\Sigma}}\right)\boldsymbol{\theta}^* - (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right\|_\infty \\
&\leq \left\|\left(\hat{\mathbf{\Sigma}} - \tilde{\mathbf{\Sigma}}\right)\boldsymbol{\theta}^*\right\|_\infty + \left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_\infty \\
&\leq \|\boldsymbol{\theta}^*\|_1 \left\|\hat{\mathbf{\Sigma}} - \tilde{\mathbf{\Sigma}}\right\|_{\max} + 2\pi\sqrt{\frac{\log p}{n}}
\end{aligned}$$

$$\leq \sqrt{s} \cdot \|\boldsymbol{\theta}^*\|_2 \left\|\hat{\boldsymbol{\Sigma}} - \tilde{\boldsymbol{\Sigma}}\right\|_{\max} + 2\pi\sqrt{\frac{\log p}{n}}$$
$$\leq \frac{3\pi}{\sqrt{\lambda_{\min}}}\sqrt{\frac{s\log p}{n}} + 2\pi\sqrt{\frac{\log p}{n}} \leq \frac{5\pi}{\sqrt{\lambda_{\min}}}\sqrt{\frac{s\log p}{n}} \ ,$$

where we use Lemma 14 and 15, and thus $\boldsymbol{\theta}^*$ is feasible with probability $1 - \frac{2}{p} - \frac{1}{p^{2.5}}$ by union bound. On the other hand, since $\check{\boldsymbol{\theta}}$ is optimal solution to (7.8), it satisfies

$$\|\check{\boldsymbol{\theta}}\|_1 \leq \|\boldsymbol{\theta}^*\|_1 \qquad \text{and} \qquad \left\|\hat{\boldsymbol{\Sigma}}\check{\boldsymbol{\theta}} - \hat{\boldsymbol{\beta}}\right\|_\infty \leq \gamma_n \ .$$

Letting $\mathbf{z} = \check{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$, we thus have

$$\left\|\hat{\boldsymbol{\Sigma}}\mathbf{z}\right\|_\infty \leq \left\|\hat{\boldsymbol{\Sigma}}\check{\boldsymbol{\theta}} - \hat{\boldsymbol{\beta}}\right\|_\infty + \left\|\hat{\boldsymbol{\Sigma}}\boldsymbol{\theta}^* - \hat{\boldsymbol{\beta}}\right\|_\infty \leq 2\gamma_n \quad \Longrightarrow$$
$$\mathbf{z}^T\hat{\boldsymbol{\Sigma}}\mathbf{z} = \left\langle \mathbf{z}, \hat{\boldsymbol{\Sigma}}\mathbf{z} \right\rangle \leq \|\mathbf{z}\|_1 \left\|\hat{\boldsymbol{\Sigma}}\mathbf{z}\right\|_\infty \leq 2\gamma_n\|\mathbf{z}\|_1$$

Using Lemma 16 combined with the inequality above, with probability at least $1 - \frac{2}{p} - \frac{1}{p^{2.5}}$, we get

$$\frac{\lambda_{\min}}{2}\|\mathbf{z}\|_2^2 \leq \mathbf{z}^T\hat{\boldsymbol{\Sigma}}\mathbf{z} \leq 2\gamma_n\|\mathbf{z}\|_1 \qquad \Longrightarrow \qquad \|\mathbf{z}\|_2 \leq \frac{4\gamma_n}{\lambda_{\min}}\frac{\|\mathbf{z}\|_1}{\|\mathbf{z}\|_2} \leq \frac{40\pi}{\lambda_{\min}^{3/2}}\sqrt{\frac{s^2\log p}{n}} \ ,$$

where we use the fact that $\sup_{\mathbf{z}\in\mathcal{C}} \frac{\|\mathbf{z}\|_1}{\|\mathbf{z}\|_2} \leq 2\sqrt{s}$. ∎

7.4.2 Improved RE Condition

From the result stated in Lemma 16, we see that the $O(s^2 \log p)$ sample complexity for RE condition of $\hat{\boldsymbol{\Sigma}}$ is worse than that of $\frac{\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}}{n}$, which is $O(s\log p)$ [21,127]. Next we show that this sharper bound (see Theorem 23) can be obtained for $\hat{\boldsymbol{\Sigma}}$ if the distribution of $\mathbf{x}$ further satisfies the *sign sub-Gaussian condition* [69]. This result may be of independent interest.

Definition 20 (sign sub-Gaussian condition) For a random variable x, the operator $\psi : \mathbb{R} \mapsto \mathbb{R}$ is defined as

$$\psi(x;\alpha,t_0) \triangleq \inf\left\{c > 0 : \mathbb{E}\exp\{t(x^\alpha - \mathbb{E}x^\alpha)\} \leq \exp(ct^2),\ \textit{for}\ |t| < t_0\right\} . \quad (7.23)$$

The random vector $\mathbf{x} \in \mathbb{R}^p$ satisfies the sign sub-Gaussian condition iff

$$\sup_{\mathbf{v}\in\mathbb{S}^{p-1}} \psi\left(\left\langle \text{sign}(\mathbf{x}-\mathbf{x}'), \mathbf{v}\right\rangle; 2, t_0\right) \leq \kappa\|\mathbf{T}\|_2^2 , \quad (7.24)$$

for a fixed constant κ and some $t_0 > 0$ such that $t_0\kappa\|\mathbf{T}\|_2^2$ is lower bounded by a fixed constant, where $\mathbf{x}'$ is an independent copy of $\mathbf{x}$ and $\mathbf{T}$ is the population Kendall's tau correlation matrix defined in (7.11).

Detailed discussions on the sign sub-Gaussian condition can be found in [69], which is out of the scope of this work. In particular, [69] show that if sign sub-Gaussian condition for transelliptical $\mathbf{x}$, the $\hat{\mathbf{\Sigma}}$ will converge with high probability to $\tilde{\mathbf{\Sigma}}$ at rate $O(\sqrt{\frac{s\log p}{n}})$ in terms of *restricted spectral norm*,

$$\|\hat{\mathbf{\Sigma}} - \tilde{\mathbf{\Sigma}}\|_{2,s} \triangleq \sup_{\substack{\mathbf{v}\in\mathbb{S}^{p-1}\\ \|\mathbf{v}\|_0\leq s}} \left|\mathbf{v}^T(\hat{\mathbf{\Sigma}} - \tilde{\mathbf{\Sigma}})\mathbf{v}\right| = O\left(\sqrt{\frac{s\log p}{n}}\right) . \quad (7.25)$$

Starting from this result, we show that with high probability the RE condition will hold for $\hat{\mathbf{\Sigma}}$ with $O(s\log p)$ samples.

Theorem 23 *Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^T$ be i.i.d. samples of $\mathbf{x} \sim TE(\tilde{\mathbf{\Sigma}}, \xi, \mathbf{f})$ for which the sign sub-Gaussian condition holds with constant κ. Define the constant*

$$c_0 = \max\left\{\frac{320\kappa\pi^4\|\tilde{\mathbf{\Sigma}}\|_2^2}{\lambda_{\min}^2}, \frac{\pi^2}{\lambda_{\min}}\right\} ,$$

in which $\lambda_{\min}$ is the smallest eigenvalue $\tilde{\mathbf{\Sigma}}$. If $n \geq \frac{128c_0}{\lambda_{\min}} s \log p = O(s \log p)$, with probability at least $1 - \frac{2}{p} - \frac{1}{p^2}$, $\hat{\mathbf{\Sigma}}$ satisfies the following RE condition,

$$\inf_{\mathbf{v} \in \mathcal{C}} \mathbf{v}^T \hat{\mathbf{\Sigma}} \mathbf{v} \geq \frac{\lambda_{\min}}{2} \ , \tag{7.26}$$

where $\mathcal{C}$ is defined in (7.21).

Remark: The proof of Theorem 23 is deferred to the appendix. Note that Theorem 22 relies on the RE condition described in Lemma 16, but we emphasize that if sign sub-Gaussian condition holds we can obtain similar result as long as n attains the bound in Theorem 23, which is smaller than the one required in Lemma 16.

7.4.3 Computation of $\mathcal{F}$

After $\hat{\boldsymbol{\theta}}$ is obtained, we can turn to the estimation of transformations $\mathcal{F}$. As we only have access to a finite number of samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, it is impossible to know the exact function. Hence we use the simple nearest-neighbor interpolation to approximate the f_j as mentioned in (7.10). By leveraging the monotonicity of f_j, we can estimate $\tilde{\mathbf{X}}$ via solving the constrained least squares problem below,

$$\hat{\mathbf{X}} = \underset{\mathbf{Z} \in \mathbb{R}^{n \times p}}{\operatorname{argmin}} \ \ \ell(\mathbf{Z}) = \frac{1}{2} \|\mathbf{Z}\hat{\boldsymbol{\theta}} - \mathbf{y}\|_2^2 \quad \text{s.t.} \quad \mathbf{z}^j \in \mathcal{M}(\mathbf{x}^j), \ \forall \ 1 \leq j \leq p \ , \tag{7.27}$$

where the set $\mathcal{M}(\mathbf{x})$ denotes the *monotone cone* induced by vector $\mathbf{x}$, i.e.,

$$\mathcal{M}(\mathbf{x}) = \{\mathbf{v} \mid v_i \geq v_j \text{ iff } x_i \geq x_j, \ \forall \ 1 \leq i, j \leq p\} \ . \tag{7.28}$$

The problem (7.27) is convex w.r.t. $\mathbf{Z}$. Note that if $\hat{\boldsymbol{\theta}} = \mathbf{1}$, the problem (7.27) is reduced to the estimation of $\mathcal{F}$ in AIM, which can be solved by the CPAV algorithm. Hence similar CPAV-type algorithm applies here, which is essentially a procedure of cyclic block coordinate descent (BCD) with exact minimization (i.e., minimizing $\ell(\mathbf{Z})$ w.r.t. each $\mathbf{z}^j$ cyclically while keeping other blocks fixed). In this scheme, each subproblem turns out to be an isotonic regression [16]. To be specific, we let $\hat{\mathbf{X}}_{(k)}$ be the iterate of the k-th round update, and define the residue for the j-th block as

$$\mathbf{r}^j_{(k)} = \mathbf{y} - \sum_{i<j} \hat{\theta}_i \hat{\mathbf{x}}^i_{(k)} - \sum_{i>j} \hat{\theta}_i \hat{\mathbf{x}}^i_{(k-1)}. \tag{7.29}$$

Then each $\hat{\mathbf{x}}^j_{(k)}$ is obtained by solving

$$\hat{\mathbf{x}}^j_{(k)} = \operatorname*{argmin}_{\mathbf{z}^j \in \mathcal{M}(\mathbf{x}^j)} \ \frac{1}{2} \left\| \mathbf{z}^j - \frac{\mathbf{r}^j_{(k)}}{\hat{\theta}_j} \right\|_2^2 , \tag{7.30}$$

which can be efficiently computed in $O(n)$ time using a skillful implementation of PAVA [65]. If we define for a set $\mathcal{A}$ the projection operator as $P_{\mathcal{A}}(\mathbf{z}) = \operatorname{argmin}_{\mathbf{x}\in\mathcal{A}} \frac{1}{2}\|\mathbf{x} - \mathbf{z}\|_2^2$, the isotonic regression (7.30) is simply the projection of $\mathbf{r}^j_{(k)}/\hat{\theta}_j$ onto the monotone cone $\mathcal{M}(\mathbf{x}^j)$. Note that $\ell(\cdot)$ is a function of the design $\mathbf{Z}$ instead of the coefficient vector $\hat{\boldsymbol{\theta}}$. Though being convex, the problem (7.27) can have infinitely many solutions, some of which can be far from the original $\tilde{\mathbf{X}}$. For example, given any $\hat{\mathbf{X}}$, we can construct another optimum via shifting two columns $\hat{\mathbf{x}}^i$ and $\hat{\mathbf{x}}^j$ by μ_i and μ_j respectively, such that $\hat{\theta}_i\mu_i + \hat{\theta}_j\mu_j = 0$. To avoid these "bad" solutions, we further impose on each $\hat{\mathbf{x}}^j$ the constraints $\mathbf{1}^T\hat{\mathbf{x}}^j = 0$ and $\|\hat{\mathbf{x}}^j\|_2 \leq \sqrt{n}$, as the marginal distribution of $\tilde{x}_{ij}$ is zero-mean and unit-variance. With additional constraints, the new problem is given by

$$\begin{aligned} \hat{\mathbf{X}} \ &= \ \operatorname*{argmin}_{\mathbf{Z}\in\mathbb{R}^{n\times p}} \ell(\mathbf{Z}) \\ \text{s.t. } \ \mathbf{z}^j \in \mathcal{M}(\mathbf{x}^j), &\quad \mathbf{1}^T\mathbf{z}^j = 0, \quad \|\mathbf{z}^j\|_2 \leq \sqrt{n}, \quad \forall\ 1 \leq j \leq p, \end{aligned} \tag{7.31}$$

and the subproblem for each block boils down to

$$\hat{\mathbf{x}}^j_{(k)} = \underset{\mathbf{z}^j \in \mathcal{M}(\mathbf{x}^j)}{\operatorname{argmin}} \ \frac{1}{2}\left\|\mathbf{z}^j - \frac{\mathbf{r}^j_{(k)}}{\hat{\theta}_j}\right\|_2^2 \quad \text{s.t.} \quad \mathbf{1}^T\mathbf{z}^j = 0, \ \ \|\mathbf{z}^j\|_2 \leq \sqrt{n}, \tag{7.32}$$

which we name *standardized isotonic regression.* The solution to (7.32) can be viewed as the projection onto the intersection of monotone cone $\mathcal{M}(\mathbf{x}^i)$, hyperplane $\mathcal{L} = \{\mathbf{z} \mid \mathbf{1}^T\mathbf{z} = 0\}$, and scaled L_2-norm ball $\mathcal{B} = \{\mathbf{z} \mid \|\mathbf{z}\|_2 \leq \sqrt{n}\}$. The next theorem show that the standardized isotonic regression is equivalent to the ordinary isotonic regression followed by successive projection on $\mathcal{L}$ and $\mathcal{B}$.

Theorem 24 *Given any monotone cone $\mathcal{M}$, the following equality holds*

$$P_{\mathcal{M}\cap\mathcal{L}\cap\mathcal{B}}(\cdot) = P_{\mathcal{B}}(P_{\mathcal{L}}(P_{\mathcal{M}}(\cdot))) \ , \tag{7.33}$$

where $P_{\mathcal{L}}(\mathbf{z}) = \mathbf{z} - \frac{\mathbf{1}^T\mathbf{z}}{n} \cdot \mathbf{1}$ and $P_{\mathcal{B}}(\mathbf{z}) = \min\{\frac{\sqrt{n}}{\|\mathbf{z}\|_2}, 1\} \cdot \mathbf{z}$.

The proof of Theorem 24 is given in the appendix. Theorem 24 indicates that the extra cost for each subproblem of our CPAV algorithm is very minimal, since the projection onto $\mathcal{L}$ and $\mathcal{B}$ can be done in linear time. Note that the CPAV for AIM needs to work with p blocks of variables, and pre-specifying the monotonicity for each f_j could lead to as many as 2^p different combinations, which is computationally prohibitive. In contrast, our algorithm only deals with roughly $O(s)$ blocks and need not specify the monotonicity. The details of our CPAV is given Algorithm 8. For $\hat{\theta}_j = 0$, the corresponding f_j will have no contribution to the estimated SLIM, which is thus skipped in our CPAV. The convergence of Algorithm 8 basically follows from the extensive studies on cyclic BCD type algorithms [17, 115, 168]. Recently [158] show that the convergence rate of BCD with exact minimization achieves $O(1/t)$ for a family of quadratic nonsmooth problem without linear dependency on the number of blocks, which applies to Algorithm 8 for solving (7.31).

Algorithm 8 Estimating $\tilde{\mathbf{X}}$

Input: Data $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p}$, estimated $\hat{\boldsymbol{\theta}}$, number of round t
Output: Estimated hidden design $\hat{\mathbf{X}}$

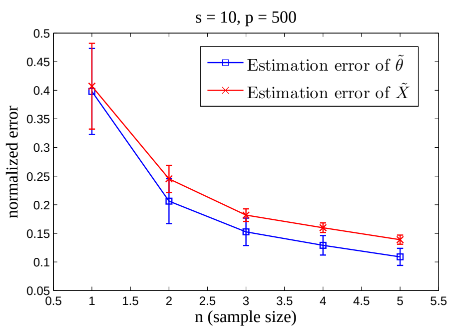
Initialize $\hat{\mathbf{X}}_{(0)} = \mathbf{0}_{n \times p}$
for $k := 1, 2, \ldots, t$ **do**
  for $j := 1, 2, \ldots, p$ **do**
    if $\hat{\theta}_j \neq 0$ **then**
      Compute $\mathbf{r}^j_{(k)}$ using (7.29)
      Compute $\mathbf{z}^j_{(k)} = P_{\mathcal{M}(\mathbf{x}^j)}\left(\frac{\mathbf{r}^j_{(k)}}{\hat{\theta}_j}\right)$ using PAVA
      $\hat{\mathbf{x}}^j_{(k)} := P_{\mathcal{B}}(P_{\mathcal{L}}(\mathbf{z}^j_{(k)}))$
    end if
  end for
end for
Return $\hat{\mathbf{X}} = \hat{\mathbf{X}}_{(t)}$

7.5 Experimental Results

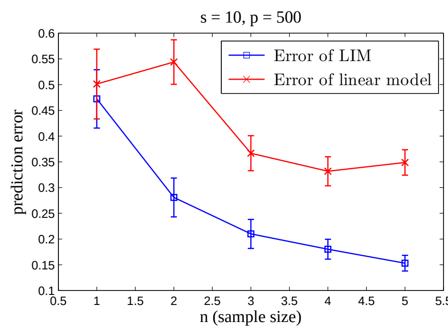
In this section, we show some experimental evidence for the effectiveness of SLIM. We test our estimation algorithm on the synthetic data. Specifically we fix the problem dimension $p = 500$, the sparsity level of $\tilde{\boldsymbol{\theta}}$, $s = 10$. The distribution of $\mathbf{x}$ is chosen as $NPN(\tilde{\boldsymbol{\Sigma}}, \mathbf{f})$, and $y \sim \langle \tilde{\boldsymbol{\theta}}, \tilde{\mathbf{x}} \rangle + \mathcal{N}(0, 0.25)$. The covariance matrix is given by $\tilde{\boldsymbol{\Sigma}} = \mathbf{A}\mathbf{A}^T$, where $\mathbf{A}$ is a Gaussian random matrix with normalized rows. In data preparation, we first generate $\tilde{\mathbf{x}}$ from $\mathcal{N}(\mathbf{0}, \tilde{\boldsymbol{\Sigma}})$. For the ten $\tilde{x}_j$'s whose corresponding $\tilde{\theta}_j$'s are nonzero, we then apply ten different monotonically increasing functions to obtain x_j's, which are basically the inverse of f_j's. The ten inverse functions are summarized in the table below. The $\Phi(\cdot)$ in f_4^{-1} is the CDF of standard norm distribution. For the rest of $\tilde{x}_j$, we randomly apply one of the functions above. All the results are obtained based on the average over 100 trials.

$f_1^{-1}(x) = x^3$	$f_6^{-1}(x) = x\log(\lvert x\rvert + 1)$
$f_2^{-1}(x) = \mathrm{sign}(x)\sqrt{\lvert x\rvert}$	$f_7^{-1}(x) = 1/(1+\exp(-x))$
$f_3^{-1}(x) = \exp(x)$	$f_8^{-1}(x) = x - 1$
$f_4^{-1}(x) = \Phi(x)$	$f_9^{-1}(x) = \mathrm{sign}(x)\log(\lvert x\rvert + 1)$
$f_5^{-1}(x) = x\exp(\sqrt{\lvert x\rvert})$	$f_{10}^{-1}(x) = \log(\exp(x) + 1)$

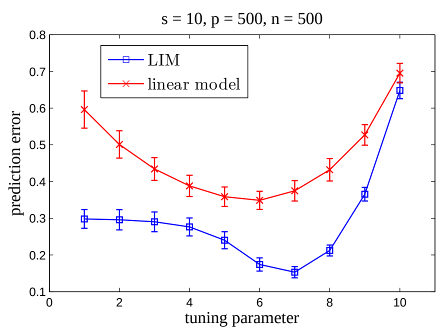
Table 7.1: Inverse of function f_j for nonzero $\tilde{\theta}_j$



(a) Estimation error vs. sample size



(b) Prediction error vs. sample size



(c) Prediction error vs. tuning parameter

Figure 7.1: Error for SLIM

We plot in Figure 7.1(a) the normalized estimation error of $\tilde{\boldsymbol{\theta}}$ and $\tilde{\mathbf{X}}$, $\frac{\|\tilde{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}\|_2}{\|\tilde{\boldsymbol{\theta}}\|_2}$ and $\frac{\|\tilde{\mathbf{X}}-\hat{\mathbf{X}}\|_2}{\|\tilde{\mathbf{X}}\|_2}$. As sample size n increases from 100 to 500, we can see the clear decreasing trend of error. We also compare the prediction error of SLIM with the simple linear model on 200 new data points, which is shown in Figure 7.1(b). The best tuning parameters for both methods are picked up via grid search. The simple linear model fails to capture the nonlinear correlation between $\mathbf{x}$ and y, thus incurring large prediction errors. In contrast, SLIM better fits the data and has substantially smaller errors. In Figure

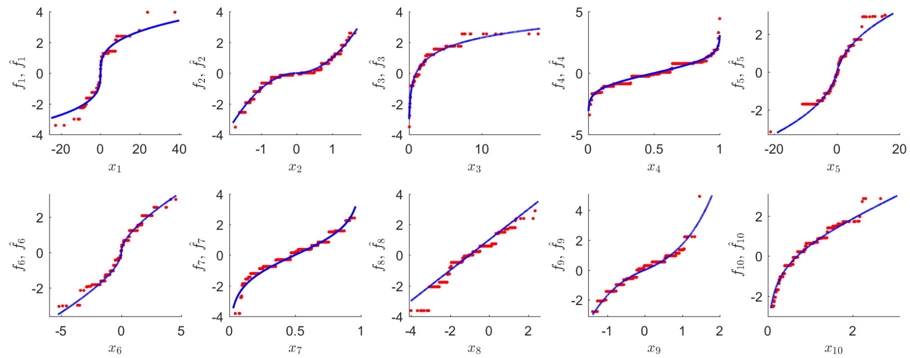


Figure 7.2: Function f_j (blue curves) and the corresponding estimated $\hat{f}_j$ at observed x_j (red dots) ($n = 500$)

7.1(c), we specifically plot the prediction errors along the parameter-tuning paths when $n = 500$, and see that SLIM always outperforms the linear model (The actual parameters are different for both methods, but we keep the largest as 2^9 times the smallest). In Figure 7.2, we also provide the plots for $f_1, f_2, \cdot, f_{10}$ and the corresponding estimated ones at the observed $x_1, x_2, \ldots, x_{10}$. It is not difficult to see that the red dots are closely distributed around the function curves except for some tails.

Appendix

Appendix 7.A Proof of Lemma 16

Proof: Let $\mathcal{S}$ be the support of $\boldsymbol{\theta}^*$, then we have

$$\mathbf{v} \in \mathcal{C} \implies \|\boldsymbol{\theta}^*_{\mathcal{S}} + \mathbf{v}_{\mathcal{S}}\|_1 + \|\mathbf{v}_{\mathcal{S}^c}\|_1 \leq \|\boldsymbol{\theta}^*\|_1$$

$$\implies \|\boldsymbol{\theta}^*_{\mathcal{S}}\|_1 - \|\mathbf{v}_{\mathcal{S}}\|_1 + \|\mathbf{v}_{\mathcal{S}^c}\|_1 \leq \|\boldsymbol{\theta}^*\|_1 \implies$$

$$\|\mathbf{v}_{\mathcal{S}^c}\|_1 \leq \|\mathbf{v}_{\mathcal{S}}\|_1 \implies \|\mathbf{v}\|_1 \leq 2\|\mathbf{v}_{\mathcal{S}}\|_1 \leq 2\sqrt{s}\|\mathbf{v}_{\mathcal{S}}\|_2 \leq 2\sqrt{s}$$

With probability at least $1-p^{-2.5}$, we have for any $\mathbf{v} \in \mathcal{C}$

$$\begin{aligned}
\mathbf{v}^T \hat{\boldsymbol{\Sigma}} \mathbf{v} &\geq \mathbf{v}^T \tilde{\boldsymbol{\Sigma}} \mathbf{v} - \left| \mathbf{v}^T \left(\hat{\boldsymbol{\Sigma}} - \tilde{\boldsymbol{\Sigma}} \right) \mathbf{v} \right| \\
&\geq \lambda_{\min} - \left| \sum_{1 \leq i,j \leq p} v_i v_j \left(\hat{\sigma}_{ij} - \tilde{\sigma}_{ij} \right) \right| \\
&\geq \lambda_{\min} - \|\mathbf{v}\|_1^2 \left\| \hat{\boldsymbol{\Sigma}} - \tilde{\boldsymbol{\Sigma}} \right\|_{\max} \\
&\geq \lambda_{\min} - 12\pi \sqrt{\frac{s^2 \log p}{n}} ,
\end{aligned}$$

where we use Lemma 14 and the fact $\|\mathbf{v}\|_1 \leq 2\sqrt{s}$. As $n \geq \left(\frac{24\pi}{\lambda_{\min}} \right)^2 s^2 \log p$, we have

$$\mathbf{v}^T \hat{\boldsymbol{\Sigma}} \mathbf{v} \geq \lambda_{\min} - 12\pi \sqrt{\frac{s^2 \log p}{n}} \geq \lambda_{\min} - \frac{\lambda_{\min}}{2} = \frac{\lambda_{\min}}{2} ,$$

which completes the proof. ■

Appendix 7.B Proof of Theorem 23

To prove Theorem 23, we first formally state below the convergence result for $\hat{\boldsymbol{\Sigma}}$ and $\tilde{\boldsymbol{\Sigma}}$ in [69].

Lemma 17 (Theorem 4.10 in [69]) *Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^T$ be i.i.d. samples of $\mathbf{x} \sim TE(\tilde{\boldsymbol{\Sigma}}, \xi, \mathbf{f})$ for which the sign sub-Gaussian condition holds with constant κ. With probability at least $1 - 2\alpha - \alpha^2$, $\hat{\boldsymbol{\Sigma}}$ constructed from $\mathbf{X}$ satisfies*

$$\|\hat{\boldsymbol{\Sigma}} - \tilde{\boldsymbol{\Sigma}}\|_{2,s_0} \ \leq \ \pi^2 \left(\frac{s_0 \log p}{n} + 2\sqrt{2\kappa} \|\tilde{\boldsymbol{\Sigma}}\|_2 \sqrt{\frac{s_0 \left(3 + \log(p/s_0)\right) + \log(1/\alpha)}{n}} \right) , \quad (7.34)$$

where $\|\mathbf{A}\|_{2,s_0} \triangleq \sup_{\mathbf{v} \in \mathbb{S}^{p-1}, \|\mathbf{v}\|_0 \leq s_0} \mathbf{v}^T \mathbf{A} \mathbf{v}$.

The next step for showing Theorem 23 is to extend the RE condition on all s_0-sparse unit vectors (s_0 needs to be appropriately specified) to all unit vectors inside the targeted error spherical cap $\mathcal{C}$. Lemma 18 accomplishes this goal.

Lemma 18 *Given $\hat{\mathbf{\Sigma}}$ constructed from $\mathbf{X}$ whose rows are generated from $TE(\tilde{\mathbf{\Sigma}}, \xi, \mathbf{f})$, we assume that for every s_0-sparse unit vector $\mathbf{v}$, the condition $\mathbf{v}^T\hat{\mathbf{\Sigma}}\mathbf{v} \geq \mu$ is satisfied. Then we have for any $\mathbf{u} \in \mathcal{C}$,*

$$\mathbf{u}^T\hat{\mathbf{\Sigma}}\mathbf{u} \geq \mu - \frac{4s}{s_0 - 1}(1-\mu) \ . \tag{7.35}$$

Proof: For any $\mathbf{u} \in \mathcal{C}$, let $\mathbf{z} \in \mathbb{R}^p$ be a random vector defined by

$$\mathbb{P}\left(\mathbf{z} = \|\mathbf{u}\|_1 \text{sign}(u_i) \cdot \mathbf{e}_i\right) = \frac{|u_i|}{\|\mathbf{u}\|_1} \ , \tag{7.36}$$

where $\{\mathbf{e}_i\}_{i=1}^p$ is the canonical basis of $\mathbb{R}^p$. Therefore, $\mathbb{E}[\mathbf{z}] = \mathbf{u}$. Let $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{s_0}$ be independent copies of $\mathbf{z}$ and set $\bar{\mathbf{z}} = \frac{1}{s_0}\sum_{i=1}^{s_0}\mathbf{z}_i$. Therefore $\bar{\mathbf{z}}$ is an s_0-sparse vector, and by our assumption on quadratic forms on s_0-sparse vectors

$$\bar{\mathbf{z}}^T\hat{\mathbf{\Sigma}}\bar{\mathbf{z}} \geq \mu\|\bar{\mathbf{z}}\|_2^2 \implies \mathbb{E}\left[\bar{\mathbf{z}}^T\hat{\mathbf{\Sigma}}\bar{\mathbf{z}}\right] \geq \mu\mathbb{E}\left[\|\bar{\mathbf{z}}\|_2^2\right] \ , \tag{7.37}$$

where the expectation is taken w.r.t $\bar{\mathbf{z}}$. Since $\bar{\mathbf{z}} = \frac{1}{s_0}\sum_{i=1}^{s_0}\mathbf{z}_i$, we have

$$\begin{aligned}
\mathbb{E}\left[\bar{\mathbf{z}}^T\hat{\mathbf{\Sigma}}\bar{\mathbf{z}}\right] &= \frac{1}{s_0^2}\sum_{1\leq i,j\leq s_0}\mathbb{E}\left[\mathbf{z}_i^T\hat{\mathbf{\Sigma}}\mathbf{z}_j\right] \\
&= \frac{1}{s_0^2}\sum_{\substack{1\leq i,j\leq s_0 \\ i\neq j}}\mathbb{E}\left[\mathbf{z}_i^T\hat{\mathbf{\Sigma}}\mathbf{z}_j\right] + \frac{1}{s_0^2}\sum_{1\leq i\leq s_0}\mathbb{E}\left[\mathbf{z}_i^T\hat{\mathbf{\Sigma}}\mathbf{z}_i\right] \\
&= \frac{s_0(s_0-1)}{s_0^2}\mathbf{u}^T\hat{\mathbf{\Sigma}}\mathbf{u} + \frac{s_0}{s_0^2}\sum_{i=1}^p\frac{|u_i|}{\|\mathbf{u}\|_1}\|\mathbf{u}\|_1^2\hat{\sigma}_{ii}
\end{aligned}$$

$$= \frac{s_0 - 1}{s_0} \mathbf{u}^T \hat{\mathbf{\Sigma}} \mathbf{u} + \frac{\|\mathbf{u}\|_1^2}{s_0} \ ,$$

since $\hat{\sigma}_{ii} = 1$, and $\sum_{i=1}^{p} \frac{|u_i|}{\|\mathbf{u}\|_1} = 1$. Replacing $\hat{\mathbf{\Sigma}}$ in the above expression by the identity matrix $\mathbf{I} \in \mathbb{R}^{p \times p}$, we have

$$\mathbb{E}\|\bar{\mathbf{z}}\|_2^2 = \frac{s_0 - 1}{s_0} \|\mathbf{u}\|_2^2 + \frac{\|\mathbf{u}\|_1^2}{s_0} \ .$$

Plugging both these expressions back in (7.37), we have

$$\begin{aligned} \frac{s_0 - 1}{s_0} \mathbf{u}^T \hat{\mathbf{\Sigma}} \mathbf{u} + \frac{\|\mathbf{u}\|_1^2}{s_0} \geq \mu \frac{s_0 - 1}{s_0} \|\mathbf{u}\|_2^2 + \mu \frac{\|\mathbf{u}\|_1^2}{s_0} \quad &\Longrightarrow \\ \mathbf{u}^T \hat{\mathbf{\Sigma}} \mathbf{u} \geq \mu \|\mathbf{u}\|_2^2 - \frac{\|\mathbf{u}\|_1^2}{s_0 - 1}(1 - \mu) \geq \mu - \frac{4s}{s_0 - 1}(1 - \mu) \ , \end{aligned}$$

where we use the facts that $\|\mathbf{u}\|_2 = 1$ and $\|\mathbf{u}\|_1 \leq 2\sqrt{s}$. That completes the proof. ∎

Equipped with Lemma 17 and 18, we present the proof of Theorem 23.

Proof of Theorem 23: For Lemma 17, we set $\alpha = \frac{1}{p}$, $s_0 = \frac{16s}{\lambda_{\min}}$, and let $c_0 = \max\{\frac{320\kappa\pi^4\|\tilde{\mathbf{\Sigma}}\|_2^2}{\lambda_{\min}^2}, \frac{\pi^2}{\lambda_{\min}}\}$. When $n \geq \frac{128c_0}{\lambda_{\min}} s \log p = 8c_0 s_0 \log p$, by Lemma 17, we have

$$\begin{aligned} \|\hat{\mathbf{\Sigma}} - \tilde{\mathbf{\Sigma}}\|_{2,s_0} &\leq \pi^2 \left(\frac{s_0 \log p}{n} + 2\sqrt{2\kappa} \|\tilde{\mathbf{\Sigma}}\|_2 \sqrt{\frac{s_0(3 + \log(p/s_0)) + \log p}{n}} \right) \\ &\leq \pi^2 \left(\frac{s_0 \log p}{\frac{\pi^2}{\lambda_{\min}} \cdot 8 s_0 \log p} + 2\sqrt{2\kappa} \|\tilde{\mathbf{\Sigma}}\|_2 \sqrt{\frac{s_0(3 + \log(p/s_0)) + \log p}{\frac{320\kappa\pi^4\|\tilde{\mathbf{\Sigma}}\|_2^2}{\lambda_{\min}^2} \cdot 8 s_0 \log p}} \right) \\ &\leq \pi^2 \left(\frac{\lambda_{\min}}{\pi^2} \sqrt{\frac{5 s_0 \log p}{320 s_0 \log p}} + \frac{\lambda_{\min}}{\pi^2} \frac{s_0 \log p}{8 s_0 \log p} \right) \\ &\leq \frac{\lambda_{\min}}{8} + \frac{\lambda_{\min}}{8} = \frac{\lambda_{\min}}{4} \ , \end{aligned}$$

with probability at least $1-\frac{2}{p}-\frac{1}{p^2}$. It follows that for any s_0-sparse unit vector $\mathbf{v}$,

$$\mathbf{v}^T\hat{\boldsymbol{\Sigma}}\mathbf{v} \geq \mathbf{v}^T\tilde{\boldsymbol{\Sigma}}\mathbf{v} - \left|\mathbf{v}^T\left(\hat{\boldsymbol{\Sigma}}-\tilde{\boldsymbol{\Sigma}}\right)\mathbf{v}\right| s \geq \lambda_{\min} - \|\hat{\boldsymbol{\Sigma}}-\tilde{\boldsymbol{\Sigma}}\|_{2,s_0} \geq \frac{3}{4}\lambda_{\min} \ ,$$

which satisfies the assumption in Lemma 18 with $\mu = \frac{3}{4}\lambda_{\min}$. With the same $s_0 = \frac{16s}{\lambda_{\min}}$, by Lemma 18, we have for any $\mathbf{v}\in\mathcal{C}$,

$$\begin{aligned}
\mathbf{v}^T\hat{\boldsymbol{\Sigma}}\mathbf{v} &\geq \frac{3}{4}\lambda_{\min} - \frac{4s}{\frac{16s}{\lambda_{\min}}-1}\left(1-\frac{3}{4}\lambda_{\min}\right) \\
&\geq \frac{3}{4}\lambda_{\min} - \frac{4s}{\frac{16s}{\lambda_{\min}}-12s}\left(1-\frac{3}{4}\lambda_{\min}\right) \\
&= \frac{3}{4}\lambda_{\min} - \frac{4s}{\frac{16s}{\lambda_{\min}}(1-\frac{3}{4}\lambda_{\min})}\left(1-\frac{3}{4}\lambda_{\min}\right) \\
&= \frac{3}{4}\lambda_{\min} - \frac{\lambda_{\min}}{4} = \frac{\lambda_{\min}}{2} \ ,
\end{aligned}$$

which completes the proof. ∎

Appendix 7.C Proof of Theorem 24

Proof: It is easy to verify the the analytic expression for $P_{\mathcal{L}}(\cdot)$ and $P_{\mathcal{B}}(\cdot)$. To show (7.33), we let $\mathbf{x}^* = P_{\mathcal{M}}(\mathbf{z})$ and $\tilde{\mathbf{x}}^* = P_{\mathcal{M}\cap\mathcal{L}\cap\mathcal{B}}(\mathbf{z})$. We assume w.l.o.g. that the monotone cone is $\mathcal{M} = \{\mathbf{x} \mid x_1 \geq x_2 \geq \ldots \geq x_n\}$. By introducing the Lagrange multipliers $\boldsymbol{\lambda} = [\lambda_1,\ldots,\lambda_{n-1}]^T$, the isotonic regression $P_{\mathcal{M}}(\mathbf{z})$ can be casted as

$$\max_{\boldsymbol{\lambda}\preceq\mathbf{0}}\min_{\mathbf{x}} \quad g(\mathbf{x},\boldsymbol{\lambda}) = \frac{1}{2}\|\mathbf{x}-\mathbf{z}\|_2^2 + \sum_{i=1}^{n-1}\lambda_i(x_i - x_{i+1}) \ ,$$

where we use the strong duality. The optimum $\mathbf{x}^*$ has to satisfy the stationarity $\nabla_{\mathbf{x}}\, g(\mathbf{x}, \boldsymbol{\lambda}) = 0$, i.e.,

$$\begin{aligned} x_1^* - z_1 + \lambda_1 = 0 \ , \\ x_2^* - z_2 - \lambda_1 + \lambda_2 = 0 \ , \\ \vdots \\ x_{n-1}^* - z_{n-1} - \lambda_{n-2} + \lambda_{n-1} = 0 \ , \\ x_n^* - z_n - \lambda_{n-1} = 0 \ . \end{aligned} \tag{7.38}$$

Using (7.38) to express $\mathbf{x}^*$ in terms of $\boldsymbol{\lambda}$, we denote $\min_{\mathbf{x}} g(\mathbf{x}, \boldsymbol{\lambda})$ by another function $h(\boldsymbol{\lambda})$, and the optimal dual variables $\boldsymbol{\lambda}^*$ satisfies

$$\boldsymbol{\lambda}^* = \underset{\boldsymbol{\lambda} \preceq \mathbf{0}}{\operatorname{argmax}} \ \ h(\boldsymbol{\lambda}) \ .$$

For the standardized isotonic regression $P_{\mathcal{M}\cap\mathcal{L}\cap\mathcal{B}}(\mathbf{z})$, we can also introduce the Lagrange multipliers $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_{n-1}]^T$, β and γ, and obtain the following optimization problem

$$\max_{\lambda \preceq \mathbf{0}, \gamma \leq 0, \beta} \min_{\mathbf{x}} \ \ \tilde{g}(\mathbf{x}, \boldsymbol{\lambda}, \beta, \gamma) = \frac{1}{2}\|\mathbf{x} - \mathbf{z}\|_2^2 + \sum_{i=1}^{n-1} \lambda_i (x_i - x_{i+1}) + \beta \sum_{i=1}^{n} x_i + \gamma(n - \|\mathbf{x}\|_2^2) \ .$$

Again the optimum $\tilde{\mathbf{x}}^*$ has to satisfy $\nabla_{\mathbf{x}}\, \tilde{g}(\tilde{\mathbf{x}}^*, \boldsymbol{\lambda}, \beta, \gamma)$,

$$\begin{aligned} (1-2\gamma)\tilde{x}_1^* - z_1 + \beta + \lambda_1 = 0 \ , \\ (1-2\gamma)\tilde{x}_2^* - z_2 + \beta - \lambda_1 + \lambda_2 = 0 \ , \\ \vdots \\ (1-2\gamma)\tilde{x}_{n-1}^* - z_{n-1} + \beta - \lambda_{n-2} + \lambda_{n-1} = 0 \ , \\ (1-2\gamma)\tilde{x}_n^* - z_n + \beta - \lambda_{n-1} = 0 \ . \end{aligned} \tag{7.39}$$

By substituting $\tilde{\mathbf{x}}^*$ for $\boldsymbol{\lambda}$, β and γ, we have

$$
\begin{aligned}
&\min_{\mathbf{x}} \tilde{g}(\mathbf{x}, \boldsymbol{\lambda}, \beta, \gamma) \\
&= \frac{1-2\gamma}{2} \sum_{i=1}^{n} \left(\tilde{x}_i^* - \frac{z_i - \beta}{1-2\gamma} \right)^2 + \sum_{i=1}^{n-1} \lambda_i (\tilde{x}_i^* - \tilde{x}_{i+1}^*) + \frac{\|\mathbf{z}\|_2^2}{2} - \frac{\sum_{i=1}^{n} (z_i - \beta)^2}{2(1-2\gamma)} + \gamma n \\
&= \frac{h(\boldsymbol{\lambda})}{1-2\gamma} + \frac{\|\mathbf{z}\|_2^2}{2} - \frac{\sum_{i=1}^{n} (z_i - \beta)^2}{2(1-2\gamma)} + \gamma n \ ,
\end{aligned}
$$

in which we note that the last three terms are free of $\boldsymbol{\lambda}$. Hence the optimal $\boldsymbol{\lambda}$ for standardized isotonic regression,

$$
\tilde{\boldsymbol{\lambda}}^* = \underset{\boldsymbol{\lambda} \preceq \mathbf{0}}{\operatorname{argmax}} \ \frac{h(\boldsymbol{\lambda})}{1-2\gamma} + \frac{\|\mathbf{z}\|_2^2}{2} - \frac{\sum_{i=1}^{n} (z_i - \beta)^2}{2(1-2\gamma)} + \gamma n = \underset{\boldsymbol{\lambda} \preceq \mathbf{0}}{\operatorname{argmax}} \ \ h(\boldsymbol{\lambda})
$$

is the same as the one for isotonic regression. Thus, combining (7.38) and (7.39), we have

$$
\tilde{\mathbf{x}}^* = \frac{\mathbf{x}^* - \beta \cdot \mathbf{1}}{1-2\gamma} \ . \tag{7.40}
$$

On the other hand, by summing up the equations respectively in (7.38) and (7.39) and using the primal feasibility $\sum_{i=1}^{n} \tilde{x}_i^* = 0$, we have

$$
\sum_{i=1}^{n} x_i^* = \sum_{i=1}^{n} z_i, \quad \sum_{i=1}^{n} z_i = n\beta \quad \Longrightarrow \quad \beta = \frac{\mathbf{1}^T \mathbf{x}^*}{n} \ ,
$$

which implies that

$$
\mathbf{x}^* - \beta \cdot \mathbf{1} = P_{\mathcal{L}}(\mathbf{x}^*) = P_{\mathcal{L}}(P_{\mathcal{M}}(\mathbf{z})) \ . \tag{7.41}
$$

Denoting $\mathbf{x}^* - \beta \cdot \mathbf{1}$ by $\hat{\mathbf{x}}^*$, we now show that scaling $\hat{\mathbf{x}}^*$ by $\frac{1}{1-2\gamma}$ is exactly the projection onto $\mathcal{B}$. If $\|\hat{\mathbf{x}}^*\|_2 > \sqrt{n}$, then $\gamma < 0$ due to (7.40) and primal feasibility $\|\tilde{\mathbf{x}}^*\|_2 \leq \sqrt{n}$. By complementary slackness $\gamma(n - \|\tilde{\mathbf{x}}^*\|_2^2) = 0$, we have $\|\tilde{\mathbf{x}}^*\|_2 = \sqrt{n}$. If $\|\hat{\mathbf{x}}^*\|_2 < \sqrt{n}$, then $\|\tilde{\mathbf{x}}^*\| < \sqrt{n}$ due to (7.40) and dual feasibility $\gamma \leq 0$. It follows from complementary slackness that $\gamma = 0$, which result in $\tilde{\mathbf{x}}^* = \hat{\mathbf{x}}^*$. If $\|\hat{\mathbf{x}}^*\|_2 = \sqrt{n}$, by similar argument, we have $\tilde{\mathbf{x}}^* = \hat{\mathbf{x}}^*$ as well. In a word, we have

$$\tilde{\mathbf{x}}^* = \begin{cases} \hat{\mathbf{x}}^*, & \text{if } \|\hat{\mathbf{x}}^*\|_2 \leq \sqrt{n} \\ \frac{\sqrt{n}}{\|\hat{\mathbf{x}}^*\|_2}\hat{\mathbf{x}}^*, & \text{if } \|\hat{\mathbf{x}}^*\|_2 > \sqrt{n} \end{cases},$$

which matches the expression for $P_{\mathcal{B}}(\cdot)$. Thus we complete the proof by noting $\tilde{\mathbf{x}}^* = P_{\mathcal{B}}(\hat{\mathbf{x}}^*) = P_{\mathcal{B}}(P_{\mathcal{L}}(P_{\mathcal{M}}(\mathbf{z})))$. ∎

Chapter 8

Structured Estimation for Multi-Response Linear Models

8.1 Introduction

In Chapter 3 and 4, we have studied the estimation of structured linear models using generalized Dantzig selector (GDS), and demonstrate that our statistical analysis based on geometric measures is of great applicability for general structures. In this chapter, we investigate the possibility of extending the analysis to a more complex setting, multi-response (a.k.a. multivariate) linear models. Multi-response linear models [5, 25, 78, 79] have found numerous applications in real-world problems, e.g. expression quantitative trait loci (eQTL) mapping in computational biology [95], land surface temperature prediction in climate informatics [61], neural semantic basis discovery in cognitive science [111], etc. Unlike simple linear model where each response is a scalar, one obtains a *response vector* at each observation in multi-response model, given as a (noisy) linear combinations of predictors, and the parameter (i.e., coefficient vector) to learn can be either response-specific (i.e., allowed to be different for every response), or shared by all responses. The multi-response model has been well studied under the context of the multi-task learning [38], where each response is coined as a *task*. In recent years, the multi-task learning literature have largely focused on exploring the parameter structure across tasks via convex formulations [6,51,88]. Another emphasis area in multi-response modeling is centered around the exploitation of the noise correlation among different responses [100,145,153,177,184], instead of assuming that the noise is independent for each response. To be specific, we consider the following multi-response linear models with m real-valued outputs,

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\theta}^* + \boldsymbol{\eta}_i, \qquad \boldsymbol{\eta}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_*) \ , \tag{8.1}$$

where $\mathbf{y}_i \in \mathbb{R}^m$ is the response vector, $\mathbf{X}_i \in \mathbb{R}^{m\times p}$ consists of m p-dimensional feature vectors, and $\boldsymbol{\eta}_i \in \mathbb{R}^m$ is a noise vector sampled from a multivariate zero-mean Gaussian distribution with covariance $\boldsymbol{\Sigma}_*$. For simplicity, we assume $\text{Diag}(\boldsymbol{\Sigma}_*) = \mathbf{I}_{m\times m}$ throughout the chapter. The m responses share the same underlying parameter $\boldsymbol{\theta}^* \in \mathbb{R}^p$, which corresponds to the so-called *pooled model* [64]. In fact, this seemingly restrictive setting is general enough to encompass the model with response-specific parameters, which can be realized by block-diagonalizing rows of $\mathbf{X}_i$ and stacking all coefficient vectors into a “long” vector. Under the assumption of correlated noise, the true noise covariance structure $\boldsymbol{\Sigma}_*$ is usually unknown. Therefore it is typically required to estimate the parameter $\boldsymbol{\theta}^*$ along with the covariance $\boldsymbol{\Sigma}_*$. In practice, we observe n data points, denoted by $\mathcal{D} = \{(\mathbf{X}_i, \mathbf{y}_i)\}_{i=1}^n$, and the maximum likelihood estimator (MLE) is simply as follows,

$$\left(\hat{\boldsymbol{\theta}}_{\text{MLE}}, \hat{\boldsymbol{\Sigma}}_{\text{MLE}}\right) = \underset{\boldsymbol{\theta}\in\mathbb{R}^p,\ \boldsymbol{\Sigma}\succeq 0}{\text{argmin}}\ \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2n}\sum_{i=1}^{n}\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\theta})\right\|_2^2 \tag{8.2}$$

Although being convex w.r.t. either $\boldsymbol{\theta}$ or $\boldsymbol{\Sigma}$ when the other is fixed, the optimization problem associated with the MLE is jointly *non-convex* for $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$. A popular approach to dealing with such problem is *alternating minimization* (AltMin), i.e., alternately solving for $\boldsymbol{\theta}$ (and $\boldsymbol{\Sigma}$) while keeping $\boldsymbol{\Sigma}$ (and $\boldsymbol{\theta}$) fixed. The AltMin algorithm for (8.2) iteratively performs two simple steps, solving least squares for $\boldsymbol{\theta}$ and computing empirical noise covariance for $\boldsymbol{\Sigma}$. Recent work [85] has established the non-asymptotic error bound of this approach for (8.2) with a brief extension to sparse parameter setting using iterative hard thresholding method [86]. But they did not allow more general structure of the parameter. Previous works [100, 137, 145] also considered the regularized MLE approaches for multi-response models with sparse parameters, which are solved by AltMin-type algorithms as well. Unfortunately, *none* of those works provide *finite-sample* statistical guarantees for their algorithms. AltMin technique has also been applied to many other problems, such as matrix completion [84], sparse coding [1], and mixed linear regression [180], with provable performance guarantees. Despite the success of AltMin, most existing works are dedicated to recovering unstructured sparse or low-rank parameters, with little attention paid to general structures, e.g., overlapping sparsity [80], hierarchical sparsity [91], k-support sparsity [7], etc.

In this chapter, we study the multi-response linear model in high-dimensional setting, and the structure of the coefficient vector $\boldsymbol{\theta}^*$ can be captured by a norm $\|\cdot\|$ [10]. We propose an *alternating estimation* (AltEst) procedure for finding the true parameters, which essentially alternates between estimating $\boldsymbol{\theta}$ through the GDS using norm $\|\cdot\|$ and computing the approximate empirical noise covariance for $\boldsymbol{\Sigma}$. Our analysis puts no restriction on what the norm can be, thus the AltEst framework is applicable to general structures. In contrast to AltMin, our AltEst procedure *cannot* be casted as a minimization of some joint objective function for $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$, thus is conceptually more general than AltMin. For the proposed AltEst, we provide the statistical guarantees for the iterate $\hat{\boldsymbol{\theta}}_{(t)}$ with the *resampling* assumption (see Section 8.2), which may justify the applicability of AltEst technique to other problems without joint objectives for two set of parameters. Specifically, we show that with overwhelming probability, the estimation error $\|\hat{\boldsymbol{\theta}}_{(t)} - \boldsymbol{\theta}^*\|_2$ for generally structured $\boldsymbol{\theta}^*$ converges *linearly* to a *minimum achievable error* given sub-Gaussian design under moderate sample size. With a straightforward intuition, this minimum achievable error can be tersely expressed by the aforementioned geometric measures which simply depend on the structure of $\boldsymbol{\theta}^*$. Moreover, our analysis implies the error bound for single response high-dimensional models as a by-product [41]. Note that the analysis in [85] focuses on the expected prediction error $\mathbb{E}[\boldsymbol{\Sigma}_*^{-1/2}\mathbf{X}(\hat{\boldsymbol{\theta}}_{(t)} - \boldsymbol{\theta}^*)]$ for unstructured $\boldsymbol{\theta}^*$, which is related but different from our $\|\hat{\boldsymbol{\theta}}_{(t)} - \boldsymbol{\theta}^*\|_2$ for generally structured $\boldsymbol{\theta}^*$. Compared with the error bound derived for unstructured $\boldsymbol{\theta}^*$ in [85], our result also yields better dependency on sample size by removing the $\log n$ factor, which seems unnatural to appear.

The rest of the chapter is organized as follows. We elaborate our AltEst algorithm in Section 8.2, along with the resampling assumption. In Section 8.3, we present the statistical guarantees for AltEst. We provide experimental results in Section 8.4 to support our theoretical development.

8.2 Alternating Estimation with GDS

Given the high-dimensional setting for (8.1), it is natural to consider the regularized MLE for (8.1) by adding the norm $\|\cdot\|$ to (8.2), which captures the structural information of $\boldsymbol{\theta}^*$ in (8.1),

$$\left(\hat{\boldsymbol{\theta}}_{\text{rg}},\ \hat{\boldsymbol{\Sigma}}_{\text{rg}}\right) = \underset{\boldsymbol{\theta}\in\mathbb{R}^p,\ \boldsymbol{\Sigma}\succeq 0}{\text{argmin}}\ \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2n}\sum_{i=1}^{n}\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\theta})\right\|_2^2 + \gamma_n\|\boldsymbol{\theta}\|\ , \tag{8.3}$$

where γ_n is a tuning parameter. Using AltMin the update of (8.3) can be given as

$$\hat{\boldsymbol{\theta}}_{(t)} = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\operatorname{argmin}} \ \frac{1}{2n} \sum_{i=1}^{n} \left\| \hat{\boldsymbol{\Sigma}}_{(t-1)}^{-\frac{1}{2}} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\theta}) \right\|_2^2 + \gamma_n \|\boldsymbol{\theta}\| \ , \tag{8.4}$$

$$\hat{\boldsymbol{\Sigma}}_{(t)} = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\theta}}_{(t)} \right) \left(\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\theta}}_{(t)} \right)^T \ , \tag{8.5}$$

where the subscript t denotes the t-th iteration. The update of $\hat{\boldsymbol{\theta}}_{(t)}$ is basically solving a regularized least squares problem, and the new $\hat{\boldsymbol{\Sigma}}_{(t)}$ is obtained by computing the approximated empirical covariance of the residues evaluated at $\hat{\boldsymbol{\theta}}_{(t)}$. In this work, we consider GDS as an alternative to (8.4), which is given by

$$\hat{\boldsymbol{\theta}}_{(t)} = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\operatorname{argmin}} \ \|\boldsymbol{\theta}\| \quad \text{s.t.} \quad \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i^T \hat{\boldsymbol{\Sigma}}_{(t-1)}^{-1} (\mathbf{X}_i \boldsymbol{\theta} - \mathbf{y}_i) \right\|_* \leq \gamma_n \ , \tag{8.6}$$

where $\|\cdot\|_*$ is the *dual norm* of $\|\cdot\|$. Compared with (8.4), GDS has nicer geometrical properties, which is favored in the statistical analysis. More importantly, since iteratively solving (8.6) followed by covariance estimation (8.5) no longer minimizes a specific objective function jointly, the updates go beyond the scope of AltMin, leading to our broader alternating estimation (AltEst) framework, i.e., alternately estimating one parameter by suitable approaches while keeping the other fixed. For the ease of exposition, we focus on the $m \leq n$ scenario, so that $\hat{\boldsymbol{\Sigma}}_{(t)}$ can be easily computed in closed form as shown in (8.5). When $m > n$ and $\boldsymbol{\Sigma}_*^{-1}$ is sparse, it is beneficial to directly estimate $\boldsymbol{\Sigma}_*^{-1}$ using more advanced estimators [29, 57]. Especially the CLIME estimator [29] enjoys certain desirable properties, which fits into our AltEst framework but not AltMin, and our AltEst analysis *does not* rely on the particular estimator we use to estimate noise covariance or its inverse. The algorithmic details are given in Algorithm 9, for which it is worth noting that every iteration t uses independent new samples, $\mathcal{D}_{2t-1}$ and $\mathcal{D}_{2t}$

in Step 3 and 4, respectively. This assumption is known as *resampling*, which facilitates the theoretical analysis by removing the statistical dependency between iterates. Several existing works benefit from such assumption when analyzing their AltMin-type algorithms [84, 128, 180]. Conceptually resampling can be implemented by partitioning the whole dataset into T subsets, though it is unusual to do so in practice. Loosely speaking, AltEst (AltMin) with resampling is an approximation of the practical AltEst (AltMin) with a single dataset $\mathcal{D}$ used by all iterations. For AltMin, attempts have been made to directly analyze its practical version without resampling, by studying the properties of the joint objective [159], which come at the price of invoking highly sophisticated mathematical tools. This technique, however, might fail to work for AltEst since the procedure is not even associated with a joint objective. In the next section, we will leverage such resampling assumption to show that the error of $\hat{\boldsymbol{\theta}}_{(t)}$ generated by Algorithm 9 will converge to a small value with high probability. We again emphasize that the AltEst framework may work for other suitable estimators for $(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_*)$ although (8.5) and (8.6) are considered in our analysis.

Algorithm 9 Alternating Estimation with Resampling

Input: Number of iterations T, Datasets $\mathcal{D}_1 = \{(\mathbf{X}_i, \mathbf{y}_i)\}_{i=1}^n$, ... , $\mathcal{D}_{2T} = \{(\mathbf{X}_i, \mathbf{y}_i)\}_{i=(2T-1)n+1}^{2Tn}$

1: Initialize $\hat{\boldsymbol{\Sigma}}_0 = \mathbf{I}_{m\times m}$
2: **for** t:= 1 to T **do**
3: Solve the GDS (8.6) for $\hat{\boldsymbol{\theta}}_{(t)}$ using dataset $\mathcal{D}_{2t-1}$
4: Compute $\hat{\boldsymbol{\Sigma}}_{(t)}$ according to (8.5) using dataset $\mathcal{D}_{2t}$
5: **end for**
6: **return** $\hat{\boldsymbol{\theta}}_T$

8.3 Statistical Analysis

In this section, we establish the statistical guarantees for our AltEst algorithm. The road map for the analysis is to first derive the error bounds separately for both (8.5) and (8.6), and then combine them through AltEst procedure to show the error bound of $\hat{\boldsymbol{\theta}}_{(t)}$. Throughout the analysis, the design $\mathbf{X}$ is assumed to centered, i.e., $\mathbb{E}[\mathbf{X}] = \mathbf{0}_{m\times p}$. $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ are used to denote the largest and smallest eigenvalue of a real symmetric matrix. Before presenting the results, we provide some basic but important concepts. First we give the definition of sub-Gaussian matrix $\mathbf{X}$ used in this section.

Definition 21 (sub-Gaussian matrix) $\mathbf{X} \in \mathbb{R}^{m\times p}$ is sub-Gaussian if the ψ_2-norm below is finite,

$$|||\mathbf{X}|||_{\psi_2} = \sup_{\mathbf{v}\in\mathbb{S}^{p-1},\ \mathbf{u}\in\mathbb{S}^{m-1}} \left|\left|\left| \mathbf{v}^T \boldsymbol{\Gamma}_{\mathbf{u}}^{-\frac{1}{2}} \mathbf{X}^T \mathbf{u} \right|\right|\right|_{\psi_2} \leq \kappa < +\infty \ , \tag{8.7}$$

where $\boldsymbol{\Gamma}_{\mathbf{u}} = \mathbb{E}[\mathbf{X}^T\mathbf{u}\mathbf{u}^T\mathbf{X}]$. Further we assume there exist constants $\mu_{\min}$ and $\mu_{\max}$ such that

$$0 < \mu_{\min} \leq \lambda_{\min}(\boldsymbol{\Gamma}_{\mathbf{u}}) \leq \lambda_{\max}(\boldsymbol{\Gamma}_{\mathbf{u}}) \leq \mu_{\max} < +\infty \ , \quad \forall\ \mathbf{u} \in \mathbb{S}^{m-1} \tag{8.8}$$

The definition (8.7) is also used in earlier work [85], which assumes the left end of (8.8) implicitly. Lemma 19 gives an example of sub-Gaussian $\mathbf{X}$, showing that condition (8.7) and (8.8) are reasonable.

Lemma 19 *Assume that* $\mathbf{X} \in \mathbb{R}^{m\times p}$ *has dependent anisotropic rows such that* $\mathbf{X} = \boldsymbol{\Xi}^{\frac{1}{2}}\tilde{\mathbf{X}}\boldsymbol{\Lambda}^{\frac{1}{2}}$, *where* $\boldsymbol{\Xi} \in \mathbb{R}^{m\times m}$ *encodes the dependency between rows,* $\tilde{\mathbf{X}} \in \mathbb{R}^{m\times p}$ *has independent isotropic rows, and* $\boldsymbol{\Lambda} \in \mathbb{R}^{p\times p}$ *introduces the anisotropy. In this setting, if each row of* $\tilde{\mathbf{X}}$ *satisfies* $|||\tilde{\mathbf{x}}_i|||_{\psi_2} \leq \tilde{\kappa}$, *then condition* (8.7) *and* (8.8) *hold with* $\kappa = C\tilde{\kappa}$, $\mu_{\min} = \lambda_{\min}(\boldsymbol{\Xi})\lambda_{\min}(\boldsymbol{\Lambda})$, *and* $\mu_{\max} = \lambda_{\max}(\boldsymbol{\Xi})\lambda_{\max}(\boldsymbol{\Lambda})$.

Proof: Let $\mathbf{w} = \mathbf{\Xi}^{\frac{1}{2}}\mathbf{u}$ for any $\mathbf{u} \in \mathbb{S}^{m-1}$, and we have

$$
\begin{aligned}
\mathbf{\Gamma}_{\mathbf{u}} &= \mathbb{E}\left[\mathbf{\Lambda}^{\frac{1}{2}}\tilde{\mathbf{X}}^T\mathbf{\Xi}^{\frac{1}{2}}\mathbf{u}\mathbf{u}^T\mathbf{\Xi}^{\frac{1}{2}}\tilde{\mathbf{X}}\mathbf{\Lambda}^{\frac{1}{2}}\right] \\
&= \mathbb{E}\left[\left[\mathbf{\Lambda}^{\frac{1}{2}}\tilde{\mathbf{x}}_1, \ldots, \mathbf{\Lambda}^{\frac{1}{2}}\tilde{\mathbf{x}}_m\right] \cdot \begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix} \cdot [w_1, \ldots, w_m] \cdot \begin{bmatrix} \tilde{\mathbf{x}}_1^T\mathbf{\Lambda}^{\frac{1}{2}} \\ \vdots \\ \tilde{\mathbf{x}}_m^T\mathbf{\Lambda}^{\frac{1}{2}} \end{bmatrix}\right] \\
&= \sum_{i=1}^m \sum_{j=1}^m w_i w_j \mathbb{E}\left[\mathbf{\Lambda}^{\frac{1}{2}}\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_j^T\mathbf{\Lambda}^{\frac{1}{2}}\right] = \sum_{i=1}^m w_i^2 \mathbf{\Lambda}^{\frac{1}{2}}\mathbb{E}\left[\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^T\right]\mathbf{\Lambda}^{\frac{1}{2}} = \left\|\mathbf{\Xi}^{\frac{1}{2}}\mathbf{u}\right\|_2^2 \cdot \mathbf{\Lambda}
\end{aligned}
$$

It is clear that

$$
\lambda_{\min}(\mathbf{\Xi}) \cdot \lambda_{\min}(\mathbf{\Lambda}) \leq \lambda_{\min}(\mathbf{\Gamma}_{\mathbf{u}}) \leq \lambda_{\max}(\mathbf{\Gamma}_{\mathbf{u}}) \leq \lambda_{\max}(\mathbf{\Xi}) \cdot \lambda_{\max}(\mathbf{\Lambda}) \ ,
$$

which indicates that condition (8.8) holds. If $|||\tilde{\mathbf{x}}_i|||_{\psi_2} \leq \tilde{\kappa}$, then

$$
\begin{aligned}
|||\mathbf{X}|||_{\psi_2} &= \sup_{\substack{\mathbf{v}\in\mathbb{S}^{p-1} \\ \mathbf{u}\in\mathbb{S}^{m-1}}} \left|\left|\left|\mathbf{v}^T\mathbf{\Gamma}_{\mathbf{u}}^{-\frac{1}{2}}\mathbf{X}^T\mathbf{u}\right|\right|\right|_{\psi_2} = \sup_{\substack{\mathbf{v}\in\mathbb{S}^{p-1} \\ \mathbf{u}\in\mathbb{S}^{m-1}}} \left|\left|\left|\frac{\mathbf{v}^T\mathbf{\Lambda}^{-\frac{1}{2}}}{\|\mathbf{\Xi}^{\frac{1}{2}}\mathbf{u}\|_2} \cdot \mathbf{\Lambda}^{\frac{1}{2}}\tilde{\mathbf{X}}^T\mathbf{\Xi}^{\frac{1}{2}}\mathbf{u}\right|\right|\right|_{\psi_2} \\
&= \sup_{\substack{\mathbf{v}\in\mathbb{S}^{p-1} \\ \mathbf{u}\in\mathbb{S}^{m-1}}} \left|\left|\left|\frac{\mathbf{v}^T\tilde{\mathbf{X}}^T}{\|\mathbf{\Xi}^{\frac{1}{2}}\mathbf{u}\|_2} \cdot \mathbf{\Xi}^{\frac{1}{2}}\mathbf{u}\right|\right|\right|_{\psi_2} = \sup_{\mathbf{v}\in\mathbb{S}^{p-1}} \left|\left|\left|\tilde{\mathbf{X}}\mathbf{v}\right|\right|\right|_{\psi_2} \leq C\tilde{\kappa}
\end{aligned}
$$

where the inequality follows from noting that the vector $\tilde{\mathbf{X}}\mathbf{v}$ has independent elements with ψ_2-norm bounded by $\tilde{\kappa}$, and thus $\left|\left|\left|\tilde{\mathbf{X}}\mathbf{v}\right|\right|\right|_{\psi_2} \leq C\tilde{\kappa}$ for any $\mathbf{v} \in \mathbb{S}^{p-1}$. Therefore condition (8.7) also holds with $\kappa = C\tilde{\kappa}$. ■

Similar to the analysis of GDS in Section 3.3, the recovery guarantee of multi-response GDS also relies on the *restricted eigenvalue* (RE) condition and an *admissible* tuning parameter γ_n. In multi-response setting, RE condition is defined jointly for designs $\mathbf{X}_i$ and a noise covariance $\mathbf{\Sigma}$ as follows.

Definition 22 (multi-response RE condition) The designs $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ and the covariance $\mathbf{\Sigma}$ together satisfy the RE condition for set $\mathcal{C} \subseteq \mathbb{S}^{p-1}$ with parameter $\alpha > 0$, if

$$\inf_{\mathbf{v} \in \mathcal{C}} \ \mathbf{v}^T \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{\Sigma}^{-1} \mathbf{X}_i \right) \mathbf{v} \ \geq \ \alpha \ . \tag{8.9}$$

The admissibility of tuning paramter γ_n also depends on the noise covariance $\mathbf{\Sigma}$.

Definition 23 (multi-response admissible tuning parameter) The γ_n for GDS (8.6) is said to be *admissible* if γ_n is chosen such that $\boldsymbol{\theta}^*$ belongs to the constraint set, i.e.,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{\Sigma}^{-1} (\mathbf{X}_i \boldsymbol{\theta}^* - \mathbf{y}_i) \right\|_* = \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{\Sigma}^{-1} \boldsymbol{\eta}_i \right\|_* \leq \gamma_n \tag{8.10}$$

For the rest of the chapter, we use C, C_0, C_1 and so on to denote universal constants, which are different from context to context. We will also drop the subscript $\|\cdot\|$ for the geometric measures and the related sets, unless it is referred to a specific norm.

8.3.1 Estimation of Coefficient Vector

In this subsection, we focus on estimating $\boldsymbol{\theta}^*$, i.e., Step 3 of Algorithm 9, using GDS of the form,

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\operatorname{argmin}} \ \|\boldsymbol{\theta}\| \quad \text{s.t.} \quad \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{\Sigma}^{-1} (\mathbf{X}_i \boldsymbol{\theta} - \mathbf{y}_i) \right\|_* \leq \gamma_n \ , \tag{8.11}$$

where $\mathbf{\Sigma}$ is an arbitrary but fixed input noise covariance matrix. Like Lemma 5, we first have the following result showing a *deterministic* error bound for $\hat{\boldsymbol{\theta}}$ under the RE condition and admissible γ_n defined in (8.9) and (8.10).

Lemma 20 *Suppose the RE condition* (8.9) *is satisfied by* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *and* $\mathbf{\Sigma}$ *with* $\alpha > 0$ *for the error spherical cap*

$$\mathcal{C} = \operatorname{cone} \{ \mathbf{v} \mid \ \|\boldsymbol{\theta}^* + \mathbf{v}\| \leq \|\boldsymbol{\theta}^*\| \ \} \cap \mathbb{S}^{p-1} \ . \tag{8.12}$$

If γ_n is admissible, $\hat{\boldsymbol{\theta}}$ in (8.11) *satisfies*

$$\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_2 \leq 2\Psi \cdot \frac{\gamma_n}{\alpha} \ , \tag{8.13}$$

in which $\Psi = \sup_{\mathbf{v} \in \mathcal{C}} \frac{\|\mathbf{v}\|}{\|\mathbf{v}\|_2}$ is the restricted norm compatibility.

Proof: Since $\hat{\boldsymbol{\theta}}$ is feasible and γ_n is selected to be admissible, we have

$$\begin{aligned}
\left\| \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^T \boldsymbol{\Sigma}^{-1} (\mathbf{X}_i \hat{\boldsymbol{\theta}} - \mathbf{y}_i) \right\|_* \leq \gamma_n, \quad \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^T \boldsymbol{\Sigma}^{-1} (\mathbf{X}_i \boldsymbol{\theta}^* - \mathbf{y}_i) \right\|_* \leq \gamma_n \\
\Longrightarrow \quad \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{X}_i (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \right\|_* \leq 2\gamma_n \\
\Longrightarrow \quad \left\langle \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*, \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{X}_i (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \right\rangle \leq \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| \cdot \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{X}_i (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \right\|_* \\
\Longrightarrow \quad (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T \left(\frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{X}_i \right) (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \leq 2\gamma_n \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|
\end{aligned}$$

As $\|\hat{\boldsymbol{\theta}}\| \leq \|\boldsymbol{\theta}^*\|$, we have $\frac{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*}{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2} \in \mathcal{C}$. By the assumption of RE condition, we further obtain

$$\begin{aligned}
\alpha \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2 \leq (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T \left(\frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{X}_i \right) (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \leq 2\gamma_n \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| \\
\Longrightarrow \quad \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq \frac{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|}{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2} \cdot \frac{2\gamma_n}{\alpha} \leq 2\Psi \cdot \frac{\gamma_n}{\alpha} \ ,
\end{aligned}$$

where we use the definition of restricted norm compatibility. ■

Considering the randomness of $\mathbf{X}_i$ and $\boldsymbol{\eta}_i$, now we turn to verifying the RE condition and finding the smallest admissible value of γ_n.

Restricted Eigenvalue Condition: First the following lemma characterizes the relation between the expectation and empirical mean of $\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X}$.

Lemma 21 *Given sub-Gaussian* $\mathbf{X} \in \mathbb{R}^{m \times p}$ *with its i.i.d. copies* $\mathbf{X}_1, \ldots, \mathbf{X}_n$*, and covariance* $\mathbf{\Sigma} \in \mathbb{R}^{m \times m}$ *with eigenvectors* $\mathbf{u}_1, \ldots, \mathbf{u}_m$*, let* $\mathbf{\Gamma} = \mathbb{E}[\mathbf{X}^T \mathbf{\Sigma}^{-1} \mathbf{X}]$ *and* $\hat{\mathbf{\Gamma}} = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^T \mathbf{\Sigma}^{-1} \mathbf{X}_i$*. Define the set* $\mathcal{C}_{\mathbf{\Gamma}_j}$ *for* $\mathcal{C} \subseteq \mathbb{S}^{p-1}$ *and each* $\mathbf{\Gamma}_j = \mathbb{E}[\mathbf{X}^T \mathbf{u}_j \mathbf{u}_j^T \mathbf{X}]$ *as*

$$\mathcal{C}_{\mathbf{\Gamma}_j} = \left\{ \mathbf{v} \in \mathbb{S}^{p-1} \;\middle|\; \mathbf{\Gamma}_j^{-\frac{1}{2}} \mathbf{v} \in \operatorname{cone}(\mathcal{C}) \right\} \ . \tag{8.14}$$

If $n \geq C_1 \kappa^4 \cdot \max_j \left\{ w^2(\mathcal{C}_{\mathbf{\Gamma}_j}) \right\}$*, with probability at least* $1 - m \exp\left(-\frac{C_2 n}{\kappa^4}\right)$*, we have*

$$\mathbf{v}^T \hat{\mathbf{\Gamma}} \mathbf{v} \geq \frac{1}{2} \mathbf{v}^T \mathbf{\Gamma} \mathbf{v}, \quad \forall\ \mathbf{v} \in \mathcal{C}\ . \tag{8.15}$$

Instead of $w(\mathcal{C}_{\mathbf{\Gamma}_j})$, ideally we want the condition above on n to be characterized by $w(\mathcal{C})$, which can be easier to compute in general. The next lemma accomplishes this goal.

Lemma 22 *Let* κ_0 *be the* ψ_2*-norm of standard Gaussian random vector and* $\mathbf{\Gamma}_{\mathbf{u}} = \mathbb{E}[\mathbf{X}^T \mathbf{u} \mathbf{u}^T \mathbf{X}]$*, where* $\mathbf{u} \in \mathbb{S}^{m-1}$ *is fixed. For* $\mathcal{C}_{\mathbf{\Gamma}_{\mathbf{u}}}$ *defined in Lemma 21, we have*

$$w(\mathcal{C}_{\mathbf{\Gamma}_{\mathbf{u}}}) \leq C \kappa_0 \sqrt{\frac{\mu_{\max}}{\mu_{\min}}} \cdot (w(\mathcal{C}) + 3)\ \ , \tag{8.16}$$

Lemma 22 implies that the Gaussian width $w(\mathcal{C}_{\mathbf{\Gamma}_j})$ appearing in Lemma 21 is of the same order as $w(\mathcal{C})$. Putting Lemma 21 and 22 together, we can obtain the RE condition for the analysis of GDS.

Corollary 6 *Under the notations of Lemma 21 and 22, if* $n \geq C_1 \kappa_0^2 \kappa^4 \cdot \frac{\mu_{\max}}{\mu_{\min}} \cdot (w(\mathcal{C}) + 3)^2$*, then the following inequality holds for all* $\mathbf{v} \in \mathcal{C} \subseteq \mathbb{S}^{p-1}$ *with probability at least* $1 - m \exp\left(-\frac{C_2 n}{\kappa^4}\right)$*,*

$$\mathbf{v}^T \hat{\mathbf{\Gamma}} \mathbf{v} \geq \frac{\mu_{\min}}{2} \cdot \operatorname{Tr}(\mathbf{\Sigma}^{-1}) \tag{8.17}$$

Proof: Given the definition of sub-Gaussian $\mathbf{X}$ and Lemma 21, we have

$$\begin{aligned}\mathbf{v}^T\hat{\mathbf{\Gamma}}\mathbf{v} \geq \frac{1}{2}\mathbf{v}^T\mathbf{\Gamma}\mathbf{v} &= \frac{1}{2}\mathbf{v}^T\left(\sum_{j=1}^{m}\frac{1}{\sigma_j}\cdot\mathbb{E}\left[\mathbf{X}^T\mathbf{u}_j\mathbf{u}_j^T\mathbf{X}\right]\right)\mathbf{v} \\ &\geq \frac{\mu_{\min}}{2}\cdot\mathbf{v}^T\mathbf{v}\left(\sum_{j=1}^{m}\frac{1}{\sigma_j}\right) = \frac{\mu_{\min}}{2}\operatorname{Tr}\left(\mathbf{\Sigma}^{-1}\right) \ .\end{aligned}$$

Using the bound in Lemma 22, we have

$$n \geq C_1\kappa_0^2\kappa^4\cdot\frac{\mu_{\max}}{\mu_{\min}}\cdot(w(\mathcal{C})+3)^2 \quad\Longrightarrow\quad n \geq C\kappa^4\cdot\max_j\left\{w^2(\mathcal{C}_{\mathbf{\Gamma}_j})\right\}$$

We complete the proof by combining the two equations above. ■

Admissible tuning parameter: Finding the admissible γ_n amounts to estimating $\|\frac{1}{n}\sum_{i=1}^n\mathbf{X}_i^T\mathbf{\Sigma}^{-1}\boldsymbol{\eta}_i\|_*$ in (8.10), which involves random $\mathbf{X}_i$ and $\boldsymbol{\eta}_i$. The next lemma establishes a high-probability bound for this quantity, which can be viewed as the smallest "safe" choice of γ_n.

Lemma 23 *Assume that $\mathbf{X}_i$ is sub-Gaussian and $\boldsymbol{\eta}_i\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma}_*)$. The following inequality holds with probability at least* $1-\exp\left(-\frac{n\tau^2}{2}\right)-C_2\exp\left(-\frac{C_1^2w^2(\Omega)}{4\rho^2}\right)$

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{\Sigma}^{-1}\boldsymbol{\eta}_i\right\|_* \leq \frac{C\kappa\sqrt{\mu_{\max}}}{\sqrt{n}}\cdot\sqrt{\operatorname{Tr}\left(\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_*\mathbf{\Sigma}^{-1}\right)}\cdot w(\Omega)\ , \tag{8.18}$$

where Ω denotes the unit ball of norm $\|\cdot\|$, $\rho=\sup_{\mathbf{v}\in\Omega}\|\mathbf{v}\|_2$, and $\tau=\frac{\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_*^{\frac{1}{2}}\|_F}{\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_*^{\frac{1}{2}}\|_2}$.

Estimation error of GDS: Building on Corollary 6, Lemma 20 and 23, the theorem below characterizes the estimation error of GDS for the multi-response linear model.

Theorem 25 *Under the setting of Lemma 23, if $n\geq C_1\kappa_0^2\kappa^4\cdot\frac{\mu_{\max}}{\mu_{\min}}\cdot\left(w\left(\mathcal{C}\right)+3\right)^2$, and γ_n is set to $C_2\kappa\sqrt{\frac{\mu_{\max}\operatorname{Tr}(\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_*\mathbf{\Sigma}^{-1})}{n}}\cdot w(\Omega)$, the estimation error of $\hat{\boldsymbol{\theta}}$ given by (8.11)* *satisfies*

$$\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_2 \leq C\kappa\sqrt{\frac{\mu_{\max}}{\mu_{\min}^2}} \cdot \frac{\sqrt{\operatorname{Tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_*\boldsymbol{\Sigma}^{-1}\right)}}{\operatorname{Tr}\left(\boldsymbol{\Sigma}^{-1}\right)} \cdot \frac{\Psi \cdot w(\Omega)}{\sqrt{n}} , \tag{8.19}$$

with probability at least $1 - m\exp\left(-\frac{C_3 n}{\kappa^4}\right) - \exp\left(-\frac{n\tau^2}{2}\right) - C_4\exp\left(-\frac{C_5^2 w^2(\Omega)}{4\rho^2}\right)$.

Proof: By Corollary 6, we have the RE condition hold with $\alpha = \frac{\mu_{\min}}{2} \cdot \operatorname{Tr}(\boldsymbol{\Sigma}^{-1})$ for $\mathcal{C}$. Combining Lemma 20 and 23, we get

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq 2\Psi \cdot \frac{\gamma_n}{\alpha} \leq C\kappa\sqrt{\frac{\mu_{\max}}{\mu_{\min}^2}} \cdot \frac{\sqrt{\operatorname{Tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_*\boldsymbol{\Sigma}^{-1}\right)}}{\operatorname{Tr}\left(\boldsymbol{\Sigma}^{-1}\right)} \cdot \frac{\Psi \cdot w(\Omega)}{\sqrt{n}} , \tag{8.20}$$

and the probability is computed via union bound. ■

Remark: We can see from the theorem above that the noise covariance $\boldsymbol{\Sigma}$ input to GDS plays a role in the error bound through the multiplicative factor

$$\xi\left(\boldsymbol{\Sigma}\right) = \frac{\sqrt{\operatorname{Tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_*\boldsymbol{\Sigma}^{-1}\right)}}{\operatorname{Tr}\left(\boldsymbol{\Sigma}^{-1}\right)} . \tag{8.21}$$

By taking the derivative of $\xi^2(\boldsymbol{\Sigma})$ w.r.t. $\boldsymbol{\Sigma}^{-1}$ and setting it to $\mathbf{0}$, we have

$$\frac{\partial \xi^2(\boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}^{-1}} = \frac{2\operatorname{Tr}^2\left(\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{\Sigma}_*\boldsymbol{\Sigma}^{-1} - 2\operatorname{Tr}\left(\boldsymbol{\Sigma}^{-1}\right)\operatorname{Tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_*\boldsymbol{\Sigma}^{-1}\right)\cdot \mathbf{I}_{m\times m}}{\operatorname{Tr}^4\left(\boldsymbol{\Sigma}^{-1}\right)} = \mathbf{0}$$

Then we can verify that $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_*$ is the solution to the equation above, and thus is the minimizer of $\xi(\boldsymbol{\Sigma})$ with $\xi(\boldsymbol{\Sigma}_*) = 1/\sqrt{\operatorname{Tr}(\boldsymbol{\Sigma}_*^{-1})}$. This calculation confirms that multi-response regression could benefit from taking into account the noise covariance, and the best performance is achieved when $\boldsymbol{\Sigma}_*$ is known. If we perform ordinary GDS by setting $\boldsymbol{\Sigma} = \mathbf{I}_{m\times m}$, then $\xi(\boldsymbol{\Sigma}) = 1/\sqrt{m}$. Therefore using $\boldsymbol{\Sigma}_*$ will reduce the error by a factor of $\sqrt{\frac{m}{\operatorname{Tr}(\boldsymbol{\Sigma}_*^{-1})}}$, compared with ordinary GDS.

One simple structure of $\boldsymbol{\theta}^*$ to consider for Theorem 25 is the sparsity encoded by L_1 norm. Given s-sparse $\boldsymbol{\theta}^*$, it follows from previous results [40, 127] that $\Psi_{L_1} = O(\sqrt{s})$,

$w(\mathcal{C}_{L_1}) = O(\sqrt{s \log p})$ and $w(\Omega_{L_1}) = O(\sqrt{\log p})$. Therefore if $n \geq O(s \log p)$, then with high probability we have

$$\left\| \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \right\|_2 \leq O\left(\xi(\boldsymbol{\Sigma}) \cdot \sqrt{\frac{s \log p}{n}} \right) \tag{8.22}$$

8.3.2 Estimation of Noise Covariance

In this subsection, we consider the estimation of noise covariance $\boldsymbol{\Sigma}_*$ given an arbitrary parameter vector $\boldsymbol{\theta}$. When m is small, we estimate $\boldsymbol{\Sigma}_*$ by simply using the sample covariance

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\theta}) (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\theta})^T \ . \tag{8.23}$$

Theorem 26 reveals the relation between $\hat{\boldsymbol{\Sigma}}$ and $\boldsymbol{\Sigma}_*$, which is sufficient for our AltEst analysis.

Theorem 26 *If* $n \geq C^4 m \cdot \max \left\{ 4 \left(\kappa_0 + \kappa \sqrt{\frac{\mu_{\max}}{\lambda_{\min}(\boldsymbol{\Sigma}_*)}} \left\| \boldsymbol{\theta}^* - \boldsymbol{\theta} \right\|_2 \right)^4, \kappa^4 \left(\frac{\lambda_{\max}(\boldsymbol{\Sigma}_*) \mu_{\max}}{\lambda_{\min}(\boldsymbol{\Sigma}_*) \mu_{\min}} \right)^2 \right\}$ *and* $\mathbf{X}_i$ *is sub-Gaussian, with probability at least* $1 - 2\exp(-C_1 m)$, $\hat{\boldsymbol{\Sigma}}$ *given by* (8.23) *satisfies*

$$\lambda_{\max} \left(\boldsymbol{\Sigma}_*^{-\frac{1}{2}} \hat{\boldsymbol{\Sigma}} \boldsymbol{\Sigma}_*^{-\frac{1}{2}} \right) \ \leq \ 1 + C^2 \kappa_0^2 \sqrt{\frac{m}{n}} + \frac{2\mu_{\max}}{\lambda_{\min}(\boldsymbol{\Sigma}_*)} \left\| \boldsymbol{\theta}^* - \boldsymbol{\theta} \right\|_2^2 \tag{8.24}$$

$$\lambda_{\min} \left(\boldsymbol{\Sigma}_*^{-\frac{1}{2}} \hat{\boldsymbol{\Sigma}} \boldsymbol{\Sigma}_*^{-\frac{1}{2}} \right) \ \geq \ 1 - C^2 \kappa_0^2 \sqrt{\frac{m}{n}} \tag{8.25}$$

Remark: If $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_*$, then $\lambda_{\max}(\boldsymbol{\Sigma}_*^{-\frac{1}{2}} \hat{\boldsymbol{\Sigma}} \boldsymbol{\Sigma}_*^{-\frac{1}{2}}) = \lambda_{\min}(\boldsymbol{\Sigma}_*^{-\frac{1}{2}} \hat{\boldsymbol{\Sigma}} \boldsymbol{\Sigma}_*^{-\frac{1}{2}}) = 1$. Hence $\hat{\boldsymbol{\Sigma}}$ is nearly equal to $\boldsymbol{\Sigma}_*$ when the upper and lower bounds (8.24) (8.25) are close to one. We would like to point out that there is nothing specific to the particular form of estimator (8.23), which makes AltEst work. Similar results can be obtained for other methods that estimate the inverse covariance matrix $\boldsymbol{\Sigma}_*^{-1}$ instead of $\boldsymbol{\Sigma}_*$. For instance, when $m < n$ and $\boldsymbol{\Sigma}_*^{-1}$ is sparse, we can replace (8.23) with GLasso [57] or CLIME [29], and

AltEst only requires the counterparts of (8.24) and (8.25) in order to work.

Section 8.3.1 shows that the noise covariance in GDS affects the error bound through the factor $\xi(\mathbf{\Sigma})$ defined in (8.21). In order to bound the error of $\hat{\boldsymbol{\theta}}_T$ given by AltEst, we need to further quantify how $\boldsymbol{\theta}$ affects $\xi(\hat{\mathbf{\Sigma}})$.

Lemma 24 *If $\hat{\mathbf{\Sigma}}$ is given as* (8.23) *and the condition in Theorem 26 holds, then the inequality below holds with probability at least* $1 - 2\exp(-C_1 m)$*,*

$$\xi\left(\hat{\mathbf{\Sigma}}\right) \leq \xi\left(\mathbf{\Sigma}_*\right) \cdot \left(1 + 2C\kappa_0\left(\frac{m}{n}\right)^{\frac{1}{4}} + 2\sqrt{\frac{\mu_{\max}}{\lambda_{\min}\left(\mathbf{\Sigma}_*\right)}}\left\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\right\|_2\right) \tag{8.26}$$

8.3.3 Error Bound for Alternating Estimation

Based on Lemma 24, the following theorem provides the error bound for $\hat{\boldsymbol{\theta}}_{(T)}$ given by Algorithm 9.

Theorem 27 *Let* $e_{orc} = C_1\kappa\sqrt{\frac{\mu_{\max}}{\mu_{\min}^2}} \cdot \frac{\xi(\mathbf{\Sigma}_*)\cdot\Psi w(\Omega)}{\sqrt{n}}$ *and* $e_{min} = e_{orc} \cdot \frac{1+2C\kappa_0\left(\frac{m}{n}\right)^{\frac{1}{4}}}{1-2e_{orc}\sqrt{\frac{\mu_{\max}}{\lambda_{\min}(\mathbf{\Sigma}_*)}}}$*. If* $n \geq C^4 m \cdot \max\left\{4\left(\kappa_0 + \frac{C_1}{C^2}\sqrt{\frac{\lambda_{\min}(\mathbf{\Sigma}_*)}{\lambda_{\max}^2(\mathbf{\Sigma}_*)}}\frac{\Psi w(\Omega)}{m}\right)^4, \left(\frac{2C_1\kappa\mu_{\max}}{C^2\mu_{\min}} \cdot \frac{\xi(\mathbf{\Sigma}_*)\Psi w(\Omega)}{\sqrt{m\cdot\lambda_{\min}(\mathbf{\Sigma}_*)}}\right)^2, \kappa^4\left(\frac{\lambda_{\max}(\mathbf{\Sigma}_*)\mu_{\max}}{\lambda_{\min}(\mathbf{\Sigma}_*)\mu_{\min}}\right)^2\right\}$ *and also satisfies the condition in Theorem 25, with high probability, the iterate $\hat{\boldsymbol{\theta}}_{(T)}$ returned by Algorithm 9 satisfies*

$$\left\|\hat{\boldsymbol{\theta}}_{(T)} - \boldsymbol{\theta}^*\right\|_2 \;\leq\; e_{\min} + \left(2e_{orc}\sqrt{\frac{\mu_{\max}}{\lambda_{\min}\left(\mathbf{\Sigma}_*\right)}}\right)^{T-1} \cdot \left(\left\|\hat{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}^*\right\|_2 - e_{\min}\right) \tag{8.27}$$

Remark: The three lower bounds for n inside curly braces correspond to three intuitive requirements. The first one guarantees that the covariance estimation is accurate enough, and the other two respectively ensure that e_{orc} and the initial error of $\hat{\boldsymbol{\theta}}_{(1)}$ are reasonably small , such that the subsequent errors can contract linearly. e_{orc} is the estimation error incurred by the following oracle estimator,

$$\hat{\boldsymbol{\theta}}_{\text{orc}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\text{argmin}} \ \|\boldsymbol{\theta}\| \quad \text{s.t.} \quad \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i^T \boldsymbol{\Sigma}_*^{-1} (\mathbf{X}_i \boldsymbol{\theta} - \mathbf{y}_i) \right\|_* \leq \gamma_n \ , \tag{8.28}$$

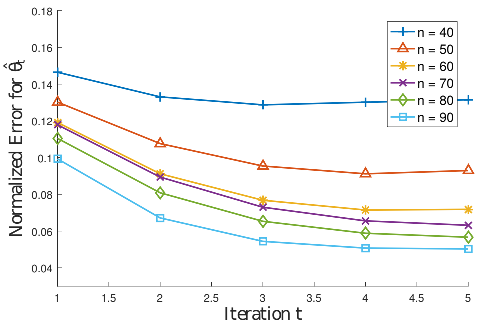
which is impossible to implement in practice. On the other hand, $e_{\min}$ is the minimum achievable error, which has an extra multiplicative factor compared with e_{orc}. The numerator of the factor compensates for the error of estimated noise covariance provided that $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ is plugged in (8.23), which merely depends on sample size. Since having $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ is also unrealistic for (8.23), the denominator further accounts for the ballpark difference between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$. As we remark after Theorem 25, if we perform ordinary GDS with $\boldsymbol{\Sigma}$ set to $\mathbf{I}_{m \times m}$ in (8.11), its error bound e_{odn} satisfies

$$e_{\text{odn}} = e_{\text{orc}} \sqrt{\frac{\text{Tr}(\boldsymbol{\Sigma}_*^{-1})}{m}} \ . \tag{8.29}$$

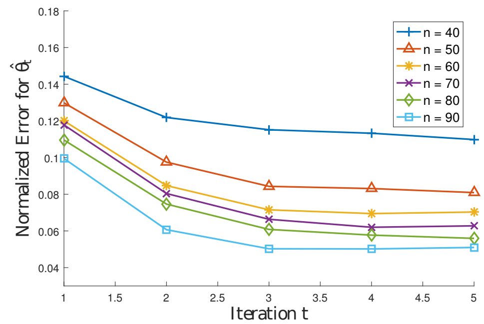
Note that this factor $\sqrt{\text{Tr}(\boldsymbol{\Sigma}_*^{-1})/m}$ is independent of n, whereas $e_{\min}$ will approach e_{orc} with increasing n as the factor between them converges to one.

8.4 Experimental Results

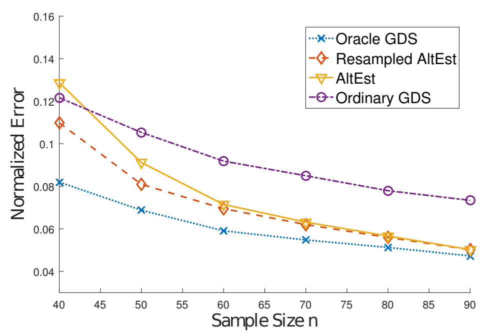
In this section, we present some experimental results to support our theoretical analysis. Specifically we focus on the sparse structure of $\boldsymbol{\theta}^*$ captured by L_1 norm. Throughout the experiment, we fix problem dimension $p = 500$, sparsity level of $\boldsymbol{\theta}^*$ $s = 20$, and number of iterations for AltEst $T = 5$. Entries of design $\mathbf{X}$ is generated by i.i.d. standard Gaussians, and $\boldsymbol{\theta}^* = [\underbrace{1, \ldots, 1}_{10}, \underbrace{-1, \ldots, -1}_{10}, \underbrace{0, \ldots, 0}_{480}]^T$. $\boldsymbol{\Sigma}_*$ is given as a block diagonal matrix with blocks $\boldsymbol{\Sigma}' = \begin{bmatrix} 1 & a \\ a & 1 \end{bmatrix}$ replicated along diagonal, and number of responses m is assumed to be even. All plots are obtained by averaging 100 trials. In the first



(a) Error for AltEst



(b) Error for resampled AltEst

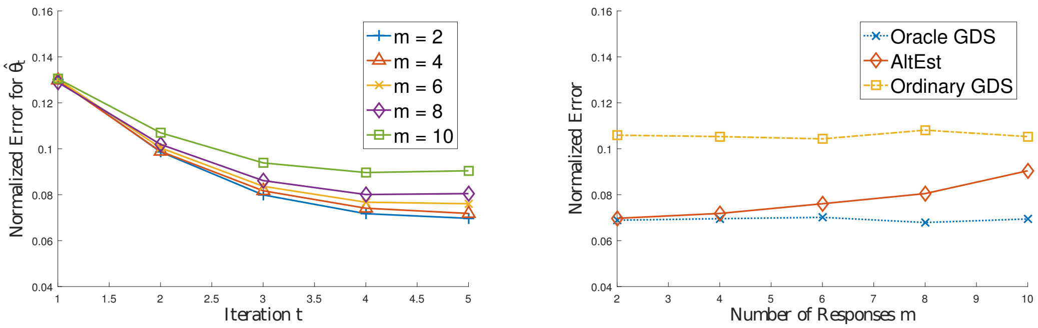


(c) Comparison of estimators

Figure 8.1: L_2-error of AltEst v.s. n. (a) When $n = 40$, AltEst is not quite stable due to the large initial error and poor quality of estimated covariance. Then the errors start to decrease for $n \geq 50$. (b) Resampld AltEst does benefit from fresh samples, and its error is slightly smaller than AltEst as well as more stable when n is small. (c) Oracle GDS outperforms the others, but the performance of AltEst is also competitive. Ordinary GDS is unable to utilize the noise correlation, thus resulting in relatively large error. By comparing the two implementations of AltEst, we can see that resampled AltEst yields smaller error especially when data is inadequate, but their errors are very close if n is suitably large.

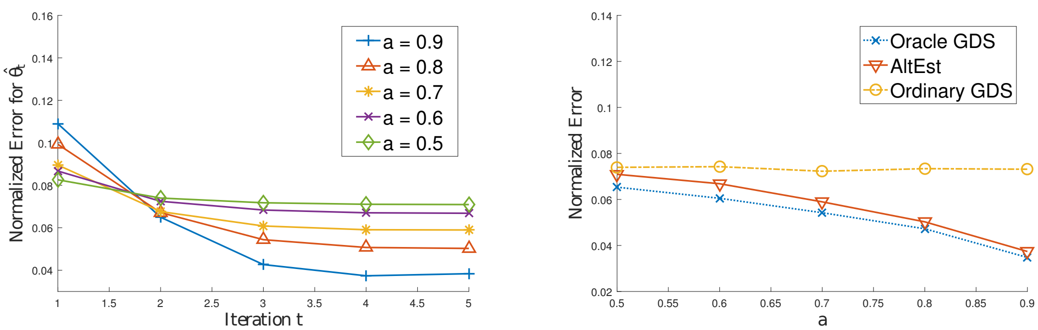
set of experiments, we set $a = 0.8$, $m = 10$ and investigate the error of $\hat{\boldsymbol{\theta}}_t$ as n varies from 40 to 90. We run AltEst (with and without resampling), the oracle GDS, and the ordinary GDS with $\boldsymbol{\Sigma} = \mathbf{I}$. The results are given in Figure 8.1.

For the second experiment, we fix the product $mn \approx 500$, and let $m = 2, 4, \ldots, 10$. For our choice of $\boldsymbol{\Sigma}_*$, the error incurred by oracle GDS e_{orc} is the same for every m. We



(a) AltEst for different m (b) Error comparison for different m

Figure 8.2: L_2-error of AltEst v.s. m. (a) Larger error comes with bigger m, which confirms that $e_{\min}$ is increasing along with m when mn is fixed. (b) The plots for oracle and ordinary GDS imply that e_{orc} and e_{odn} remain unchanged, which matches the error bounds in Theorem 25. Though $e_{\min}$ increases, AltEst still outperform the ordinary GDS by a margin.



(a) AltEst for different a (b) Error comparison for different a

Figure 8.3: L_2-error of AltEst v.s. a. (a) The error goes down when the true noise covariance becomes closer to singular, which is expected in view of Theorem 27. (b) e_{orc} also decreases as a gets larger, and the gap between $e_{\min}$ and e_{odn} widens. The definition of $e_{\min}$ in Theorem 27 indicates that the ratio between $e_{\min}$ and e_{orc} is almost a constant because both n and m are fixed. Here we observe that all the ratios at different a are between 1.05 and 1.1, which supports the theoretical results. Also, Theorem 25 suggests that e_{odn} does not change as $\mathbf{\Sigma}_*$ varies, which is verified here.

compare AltEst with both oracle and ordinary GDS, and the result is shown in Figure 8.2(a) and 8.2(b).

In the third experiment, we test AltEst under different covariance matrices $\mathbf{\Sigma}_*$ by varying a from 0.5 to 0.9. m is set to 10 and sample size n is 90. We also compare AltEst against both oracle and ordinary GDS, and the errors are reported in Figure 8.3(a) and 8.3(b).

Appendix

Appendix 8.A Proof of Statistical Guarantees for GDS

8.A.1 Proof of Lemma 21

Proof: Assume that the eigenvalue decomposition of $\mathbf{\Sigma}$ is given by $\mathbf{\Sigma} = \sum_{i=j}^{m} \sigma_i \mathbf{u}_j \mathbf{u}_j^T$. For convenience, we denote $\mathbf{z}^j = \mathbf{X}^T \mathbf{u}_j$, $\mathbf{z}_i^j = \mathbf{X}_i^T \mathbf{u}_j$, and $\hat{\mathbf{\Gamma}}_j = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{u}_j \mathbf{u}_j^T \mathbf{X}_i$. Note that $\mathbf{\Gamma}_j = \mathbb{E}[\mathbf{z}^j {\mathbf{z}^j}^T]$, $\mathbf{\Gamma} = \sum_{i=j}^{m} \frac{\mathbf{\Gamma}_j}{\sigma_j}$, $\hat{\mathbf{\Gamma}}_j = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i^j {\mathbf{z}_i^j}^T$, and $\hat{\mathbf{\Gamma}} = \sum_{j=1}^{m} \frac{\hat{\mathbf{\Gamma}}_j}{\sigma_j}$. In order to apply Lemma 3, we let (Ω_j, μ_j) be the probability measure that $\mathbf{z}^j$ is defined on, and construct the function set

$$\mathcal{H}_j = \left\{ h_{\mathbf{v}} = \left\langle \mathbf{\Gamma}_j^{-\frac{1}{2}} \mathbf{v}, \cdot \right\rangle \mid \mathbf{v} \in \mathcal{C}_{\mathbf{\Gamma}_j} \right\}$$

It is easy to see that for any $h_{\mathbf{v}} \in \mathcal{H}_j$,

$$\mathbb{E}[h_{\mathbf{v}}^2] = \mathbb{E}_{\mathbf{z}^j \sim \mu_j}\left[\mathbf{v}^T \mathbf{\Gamma}_j^{-\frac{1}{2}} \mathbf{z}^j {\mathbf{z}^j}^T \mathbf{\Gamma}_j^{-\frac{1}{2}} \mathbf{v} \right] = \mathbf{v}^T \mathbf{\Gamma}_j^{-\frac{1}{2}} \left(\mathbb{E}_{\mathbf{z}^j \sim \mu_j}\left[\mathbf{z}^j {\mathbf{z}^j}^T \right] \right) \mathbf{\Gamma}_j^{-\frac{1}{2}} \mathbf{v} = \mathbf{v}^T \mathbf{v} = 1 \ ,$$

i.e., $\mathcal{H}_j \subseteq \mathbb{S}_{L_2(\mu_j)} = \{h \mid |||h|||_{L_2(\mu_j)} = 1\}$. Based on the definition of sub-Gaussian $\mathbf{X}$, we also have for any $\mathbf{v} \in \mathcal{C}_{\mathbf{\Gamma}_j}$,

$$|||h_{\mathbf{v}}|||_{\psi_2} = \left|\left|\left| \left\langle \mathbf{\Gamma}_j^{-\frac{1}{2}} \mathbf{v}, \mathbf{z}^j \right\rangle \right|\right|\right|_{\psi_2} = \left|\left|\left| \mathbf{v}^T \mathbf{\Gamma}_j^{-\frac{1}{2}} \mathbf{X}^T \mathbf{u}_j \right|\right|\right|_{\psi_2} \leq \kappa \ ,$$

and also for any $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{C}_{\mathbf{\Gamma}_j}$, we have

$$|||h_{\mathbf{v}_1} - h_{\mathbf{v}_2}|||_{\psi_2} = \left|\left|\left|(\mathbf{v}_1 - \mathbf{v}_2)^T \mathbf{\Gamma}_j^{-\frac{1}{2}} \mathbf{z}^j\right|\right|\right|_{\psi_2} \leq \kappa \cdot \|\mathbf{v}_1 - \mathbf{v}_2\|_2 \ .$$

If we choose $\beta = \frac{1}{2}$, using (2.39), (2.40) and (2.41), then we have

$$c_1 \kappa \cdot \gamma_2(\mathcal{H}_j, |||\cdot|||_{\psi_2}) \leq c_1 \kappa^2 \cdot \gamma_2(\mathcal{C}_{\mathbf{\Gamma}_j}, \|\cdot\|_2) \leq c_1 c_4 \kappa^2 \cdot w(\mathcal{C}_{\mathbf{\Gamma}_j}) \leq \beta \sqrt{n}$$

when $n \geq C_1 \kappa^4 w^2(\mathcal{C}_{\mathbf{\Gamma}_j})$ where $C_1 = 4 c_1^2 c_4^2$. By Lemma 3, with probability at least $1 - \exp(-c_2 \beta^2 n / \kappa^4) = 1 - \exp(-C_2 n/\kappa^4)$ where $C_2 = c_2/4$, we have

$$\begin{aligned}
\sup_{h \in \mathcal{H}_j} \left| \frac{1}{n} \sum_{i=1}^{n} h^2(\mathbf{z}_i^j) - \mathbb{E}[h^2] \right| &= \sup_{\mathbf{v} \in \mathcal{C}_{\mathbf{\Gamma}_j}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{v}^T \mathbf{\Gamma}_j^{-\frac{1}{2}} \mathbf{z}_i^j \mathbf{z}_i^{jT} \mathbf{\Gamma}_j^{-\frac{1}{2}} \mathbf{v} - 1 \right| \\
&= \sup_{\mathbf{v} \in \mathcal{C}_{\mathbf{\Gamma}_j}} \left| \mathbf{v}^T \mathbf{\Gamma}_j^{-\frac{1}{2}} \hat{\mathbf{\Gamma}}_j \mathbf{\Gamma}_j^{-\frac{1}{2}} \mathbf{v} - 1 \right| \leq \frac{1}{2}
\end{aligned}$$

$$\begin{aligned}
\Longrightarrow \qquad & \mathbf{v}^T \mathbf{\Gamma}_j^{-\frac{1}{2}} \hat{\mathbf{\Gamma}}_j \mathbf{\Gamma}_j^{-\frac{1}{2}} \mathbf{v} \geq \frac{1}{2}, \quad \forall\ \mathbf{v} \in \mathcal{C}_{\mathbf{\Gamma}_j} \\
\Longrightarrow \qquad & \mathbf{v}^T \mathbf{\Gamma}_j^{-\frac{1}{2}} \hat{\mathbf{\Gamma}}_j \mathbf{\Gamma}_j^{-\frac{1}{2}} \mathbf{v} \geq \frac{1}{2} \left(\mathbf{v}^T \mathbf{\Gamma}_j^{-\frac{1}{2}} \mathbf{\Gamma}_j \mathbf{\Gamma}_j^{-\frac{1}{2}} \mathbf{v} \right), \quad \forall\ \mathbf{v} \in \mathcal{C}_{\mathbf{\Gamma}_j}
\end{aligned}$$

Let $\mathbf{w} = \mathbf{\Gamma}_j^{-\frac{1}{2}} \mathbf{v}$, and note that the inequalities above are preserved under arbitrary scaling of $\mathbf{w}$. By recalling the definition of $\mathcal{C}_{\mathbf{\Gamma}_j}$, it is not difficult to see that

$$\mathbf{w}^T \hat{\mathbf{\Gamma}}_j \mathbf{w} \geq \frac{1}{2} \mathbf{w}^T \mathbf{\Gamma}_j \mathbf{w}, \quad \forall\ \mathbf{w} \in \mathcal{C} \ . \tag{8.30}$$

Combining (8.30) for each $\mathbf{\Gamma}_j$ using union bound, we obtain

$$\mathbf{w}^T \left(\sum_{i=1}^{m} \frac{\hat{\mathbf{\Gamma}}_j}{\sigma_j} \right) \mathbf{w} \geq \frac{1}{2} \mathbf{w}^T \left(\sum_{i=1}^{m} \frac{\mathbf{\Gamma}_j}{\sigma_j} \right) \mathbf{w}, \ \ \forall\ \mathbf{w} \in \mathcal{C} \ \Longrightarrow \ \mathbf{w}^T \hat{\mathbf{\Gamma}} \mathbf{w} \geq \frac{1}{2} \mathbf{w}^T \mathbf{\Gamma} \mathbf{w}, \ \ \forall\ \mathbf{w} \in \mathcal{C} \ ,$$

which completes the proof by renaming $\mathbf{w}$ as $\mathbf{v}$. ∎

8.A.2 Proof of Lemma 22

Proof: Recall the definition of Gaussian width $w(\mathcal{C}_{\mathbf{\Gamma_u}}) = \mathbb{E}\left[\sup_{\mathbf{v}\in\mathcal{C}_{\mathbf{\Gamma_u}}} \langle \mathbf{v}, \mathbf{g}\rangle\right]$, where $\mathbf{g}$ is a standard Gaussian random vector. Given the assumption (8.8), we have $\mu_{\min} \leq \lambda_{\min}(\mathbf{\Gamma_u}) \leq \lambda_{\max}(\mathbf{\Gamma_u}) \leq \mu_{\max}$, and note that

$$\begin{aligned} \sup_{\mathbf{v}\in\mathcal{C}_{\mathbf{\Gamma_u}}} \langle \mathbf{v}, \mathbf{g}\rangle = \sup_{\mathbf{v}\in\mathcal{C}_{\mathbf{\Gamma_u}}} \left\langle \mathbf{\Gamma_u}^{-\frac{1}{2}}\mathbf{v}, \mathbf{\Gamma}_{\mathbf{u}}^{\frac{1}{2}}\mathbf{g} \right\rangle &\leq \sup_{\mathbf{v}\in\text{cone}(\mathcal{C})\cap\frac{1}{\sqrt{\mu_{\min}}}\mathbb{B}^p} \left\langle \mathbf{v}, \mathbf{\Gamma}_{\mathbf{u}}^{\frac{1}{2}}\mathbf{g} \right\rangle \\ &= \frac{1}{\sqrt{\mu_{\min}}} \cdot \sup_{\mathbf{v}\in\text{cone}(\mathcal{C})\cap\mathbb{B}^p} \left\langle \mathbf{v}, \mathbf{\Gamma}_{\mathbf{u}}^{\frac{1}{2}}\mathbf{g} \right\rangle , \end{aligned} \tag{8.31}$$

where the inequality follows from $\mathbf{\Gamma_u}^{-\frac{1}{2}}\mathbf{v} \in \text{cone}(\mathcal{C})$ and $\|\mathbf{\Gamma_u}^{-\frac{1}{2}}\mathbf{v}\|_2 \leq \frac{1}{\sqrt{\mu_{\min}}}$. Now we use generic chaining to bound the right-hand side above. Denote the set $\text{cone}(\mathcal{C}) \cap \mathbb{B}^p$ by $\mathcal{T}$, and we consider the stochastic process $\{Z_{\mathbf{v}} = \langle \mathbf{v}, \mathbf{\Gamma}_{\mathbf{u}}^{\frac{1}{2}}\mathbf{g}\rangle\}_{\mathbf{v}\in\mathcal{T}}$. For any $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{T}$, we have

$$\begin{aligned} |||Z_{\mathbf{v}_1} - Z_{\mathbf{v}_2}|||_{\psi_2} &= \left|\left|\left| \langle \mathbf{\Gamma}_{\mathbf{u}}^{\frac{1}{2}}(\mathbf{v}_1 - \mathbf{v}_2), \mathbf{g}\rangle \right|\right|\right|_{\psi_2} \\ &\leq \kappa_0 \left\| \mathbf{\Gamma}_{\mathbf{u}}^{\frac{1}{2}}(\mathbf{v}_1 - \mathbf{v}_2) \right\|_2 \\ &\leq \kappa_0 \sqrt{\mu_{\max}} \cdot \|\mathbf{v}_1 - \mathbf{v}_2\|_2 \ . \end{aligned}$$

If we define for $\mathcal{T}$ the metric $s(\mathbf{v}_1, \mathbf{v}_2) = \kappa_0\sqrt{\mu_{\max}} \cdot \|\mathbf{v}_1 - \mathbf{v}_2\|_2$, it follows from Proposition 6 that

$$\mathbb{P}\left(|Z_{\mathbf{v}_1} - Z_{\mathbf{v}_2}| \geq \epsilon\right) \leq e \cdot \exp\left(-\frac{c\epsilon^2}{\kappa_0^2 \mu_{\max}\|\mathbf{v}_1 - \mathbf{v}_2\|_2^2}\right) = e \cdot \exp\left(-\frac{c\epsilon^2}{s^2(\mathbf{v}_1, \mathbf{v}_2)}\right) \ .$$

By Lemma 2, (2.40) and (2.49), we obtain

$$\begin{aligned}\mathbb{E}\left[\sup_{\mathbf{v}\in\mathcal{T}}\langle \mathbf{v}, \boldsymbol{\Gamma}_{\mathbf{u}}^{\frac{1}{2}}\mathbf{g}\rangle\right] &= \mathbb{E}\left[\sup_{\mathbf{v}\in\mathcal{T}} Z_{\mathbf{v}}\right] \leq c_1\gamma_2(\mathcal{T}, s) \\ &= c_1\kappa_0\sqrt{\mu_{\max}}\gamma_2(\mathcal{T}, \|\cdot\|_2) \\ &\leq c_1c_2\kappa_0\sqrt{\mu_{\max}}\cdot w(\mathcal{T})\end{aligned} \tag{8.32}$$

Note that $\mathcal{T} = \text{cone}(\mathcal{C}) \cap \mathbb{B}^p \subseteq \text{conv}(\mathcal{C} \cup \{\mathbf{0}\})$. By Lemma 1, we have

$$w(\mathcal{T}) \leq w(\text{conv}(\mathcal{C}\cup\{\mathbf{0}\})) = w(\mathcal{C}\cup\{\mathbf{0}\}) \leq \max\{w(\mathcal{C}), w(\mathbf{0})\} + 2\sqrt{\ln 4} \leq w(\mathcal{C}) + 3 \ . \tag{8.33}$$

Combining (8.31), (8.32) and (8.33), we have

$$w(\mathcal{C}_{\boldsymbol{\Gamma}_{\mathbf{u}}}) = \mathbb{E}\left[\sup_{\mathbf{v}\in\mathcal{C}_{\boldsymbol{\Gamma}_{\mathbf{u}}}}\langle\mathbf{v}, \mathbf{g}\rangle\right] \leq \frac{1}{\sqrt{\mu_{\min}}}\mathbb{E}\left[\sup_{\mathbf{v}\in\mathcal{T}}\left\langle \mathbf{v}, \boldsymbol{\Gamma}_{\mathbf{u}}^{\frac{1}{2}}\mathbf{g}\right\rangle\right] \leq c_1c_2\kappa_0\sqrt{\frac{\mu_{\max}}{\mu_{\min}}}\cdot(w(\mathcal{C})+3) \ , \tag{8.34}$$

where the last inequality follows from condition (8.8). ■

8.A.3 Proof of Lemma 23

Proof: Since design $\mathbf{X}_i$ and noise $\boldsymbol{\eta}_i$ are independent, we first consider the scenario where each $\boldsymbol{\eta}_i$ is arbitrary but fixed vector. Using the definition of dual norm, we have

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}_i\right\|_* &= \frac{1}{n}\cdot\sup_{\mathbf{v}\in\Omega}\left\langle \mathbf{v},\ \sum_{i=1}^n \mathbf{X}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}_i\right\rangle \\ &= \frac{1}{n}\cdot\sup_{\mathbf{v}\in\Omega}\sum_{i=1}^n\left\langle \boldsymbol{\Lambda}_i^{\frac{1}{2}}\mathbf{v},\ \boldsymbol{\Lambda}_i^{-\frac{1}{2}}\mathbf{X}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}_i\right\rangle\end{aligned}$$

where $\mathbf{\Lambda}_i = \mathbb{E}_{\mathbf{X}_i}[\mathbf{X}_i^T \mathbf{\Sigma}^{-1} \boldsymbol{\eta}_i \boldsymbol{\eta}_i^T \mathbf{\Sigma}^{-1} \mathbf{X}_i]$. Based on the definition of sub-Gaussian $\mathbf{X}_i$, we get

$$\begin{aligned}
\left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert \mathbf{\Lambda}_i^{-\frac{1}{2}} \mathbf{X}_i^T \mathbf{\Sigma}^{-1} \boldsymbol{\eta}_i \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert_{\psi_2} &\leq \kappa \qquad \Longrightarrow \\
\left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert \sum_{i=1}^{n} \left\langle \mathbf{\Lambda}_i^{\frac{1}{2}} \mathbf{v},\ \mathbf{\Lambda}_i^{-\frac{1}{2}} \mathbf{X}_i^T \mathbf{\Sigma}^{-1} \boldsymbol{\eta}_i \right\rangle \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert_{\psi_2} &\leq c_0 \max_{1 \leq i \leq n} \left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert \mathbf{\Lambda}_i^{-\frac{1}{2}} \mathbf{X}_i^T \mathbf{\Sigma}^{-1} \boldsymbol{\eta}_i \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert_{\psi_2} \cdot \sqrt{\sum_{i=1}^{n} \left\| \mathbf{\Lambda}_i^{\frac{1}{2}} \mathbf{v} \right\|_2^2} \\
&\leq c_0 \kappa \sqrt{\sum_{i=1}^{n} \left\| \mathbf{\Lambda}_i^{\frac{1}{2}} \right\|_2^2 \|\mathbf{v}\|_2^2} \\
&\leq c_0 \kappa \sqrt{\mu_{\max}} \cdot \sqrt{\sum_{i=1}^{n} \|\mathbf{\Sigma}^{-1} \boldsymbol{\eta}_i\|_2^2} \cdot \|\mathbf{v}\|_2
\end{aligned}$$

where we use Proposition 10 in the first inequality by treating the sum of inner products as one "big" inner product. The last inequality follows from the definition of $\mu_{\max}$ in (8.8). Now we consider the stochastic process $\left\{Z_{\mathbf{v}} = \left\langle \mathbf{v},\ \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{\Sigma}^{-1} \boldsymbol{\eta}_i \right\rangle \right\}_{\mathbf{v} \in \Omega}$, where $\boldsymbol{\eta}_i$ is still fixed. For any $Z_{\mathbf{v}_1}$ and $Z_{\mathbf{v}_2}$, by the argument above and Proposition 6, we have

$$\begin{aligned}
\vert\kern-0.25ex\vert\kern-0.25ex\vert Z_{\mathbf{v}_1} - Z_{\mathbf{v}_2} \vert\kern-0.25ex\vert\kern-0.25ex\vert_{\psi_2} \leq c_0 \kappa \sqrt{\mu_{\max}} \cdot \sqrt{\sum_{i=1}^{n} \|\mathbf{\Sigma}^{-1} \boldsymbol{\eta}_i\|_2^2} \cdot \|\mathbf{v}_1 - \mathbf{v}_2\|_2 \triangleq s(\mathbf{v}_1, \mathbf{v}_2) \\
\Longrightarrow \quad \mathbb{P}\left(|Z_{\mathbf{v}_1} - Z_{\mathbf{v}_2}| > \epsilon\right) \leq e \cdot \exp\left(-\frac{C_1 \epsilon^2}{s^2(\mathbf{v}_1, \mathbf{v}_2)}\right)
\end{aligned}$$

It follows from (2.40), (2.49) and Lemma 2 that

$$\begin{aligned}
\gamma_2(\Omega, s) &= c_0 \kappa \sqrt{\mu_{\max}} \cdot \sqrt{\sum_{i=1}^{n} \|\mathbf{\Sigma}^{-1} \boldsymbol{\eta}_i\|_2^2} \cdot \gamma_2(\Omega, \|\cdot\|_2) \\
&\leq c_0 c_1 \kappa \sqrt{\mu_{\max}} \cdot \sqrt{\sum_{i=1}^{n} \|\mathbf{\Sigma}^{-1} \boldsymbol{\eta}_i\|_2^2} \cdot w(\Omega) \ ,
\end{aligned}$$

$$\mathbb{P}_{\mathbf{X}_i}\left(\sup_{\mathbf{v}_1,\mathbf{v}_2\in\Omega}|Z_{\mathbf{v}_1}-Z_{\mathbf{v}_2}|\geq c_2\left(\gamma_2(\Omega,s)+\epsilon\cdot\mathrm{diam}\left(\Omega,s\right)\right)\right)\leq c_3\exp\left(-\epsilon^2\right)$$

Combining the two inequalities above with the symmetry of Ω, we obtain

$$\mathbb{P}_{\mathbf{X}}\left(\sup_{\mathbf{v}\in\Omega}Z_{\mathbf{v}}\geq c_0c_2\kappa\sqrt{\mu_{\max}}\cdot\sqrt{\sum_{i=1}^{n}\|\mathbf{\Sigma}^{-1}\boldsymbol{\eta}_i\|_2^2}\left(\frac{c_1}{2}\cdot w(\Omega)+\epsilon\cdot\sup_{\mathbf{v}\in\Omega}\|\mathbf{v}\|_2\right)\right)$$
$$\leq c_3\exp\left(-\epsilon^2\right)$$

Letting $\rho=\sup_{\mathbf{v}\in\Omega}\|\mathbf{v}\|_2$, $\epsilon=\frac{c_1w(\Omega)}{2\rho}$, with probability at least $1-c_3\exp(-\frac{c_1^2w^2(\Omega)}{4\rho^2})$, we have

$$\sup_{\mathbf{v}\in\Omega}Z_{\mathbf{v}}=\left\|\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{\Sigma}^{-1}\boldsymbol{\eta}_i\right\|_*\leq c_0c_1c_2\kappa\sqrt{\mu_{\max}}\cdot\sqrt{\sum_{i=1}^{n}\|\mathbf{\Sigma}^{-1}\boldsymbol{\eta}_i\|_2^2}\cdot w(\Omega)\qquad(8.35)$$

for any given set of $\boldsymbol{\eta}_i$. Now we incorporate the randomness of $\boldsymbol{\eta}_i$. Essentially we need to bound

$$\sqrt{\sum_{i=1}^{n}\|\mathbf{\Sigma}^{-1}\boldsymbol{\eta}_i\|_2^2}=\sqrt{\sum_{i=1}^{n}\left\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_*^{\frac{1}{2}}\tilde{\boldsymbol{\eta}}_i\right\|_2^2},$$

where each $\tilde{\boldsymbol{\eta}}_i$ is an m-dimensional standard (isotropic) Gaussian random vector. Given $\mathbf{v}=[\mathbf{v}_1^T,\ldots,\mathbf{v}_n^T]^T\in\mathbb{R}^{mn}$, Denote $f(\mathbf{v})=\sqrt{\sum_{i=1}^{n}\left\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_*^{\frac{1}{2}}\mathbf{v}_i\right\|_2^2}$, and we have

$$|f(\mathbf{v})-f(\mathbf{w})|=\left|\sqrt{\sum_{i=1}^{n}\left\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_*^{\frac{1}{2}}\mathbf{v}_i\right\|_2^2}-\sqrt{\sum_{i=1}^{n}\left\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_*^{\frac{1}{2}}\mathbf{w}_i\right\|_2^2}\right|$$
$$\leq\sqrt{\sum_{i=1}^{n}\left(\left\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_*^{\frac{1}{2}}\mathbf{v}_i\right\|_2-\left\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_*^{\frac{1}{2}}\mathbf{w}_i\right\|_2\right)^2}$$

$$\leq \sqrt{\sum_{i=1}^{n} \left\| \mathbf{\Sigma}^{-1} \mathbf{\Sigma}_*^{\frac{1}{2}} (\mathbf{v}_i - \mathbf{w}_i) \right\|_2^2} \leq \sqrt{\sum_{i=1}^{n} \left\| \mathbf{\Sigma}^{-1} \mathbf{\Sigma}_*^{\frac{1}{2}} \right\|_2^2 \|\mathbf{v}_i - \mathbf{w}_i\|_2^2}$$
$$= \left\| \mathbf{\Sigma}^{-1} \mathbf{\Sigma}_*^{\frac{1}{2}} \right\|_2 \|\mathbf{v} - \mathbf{w}\|_2$$

which implies that f is a Lipschitz function with parameter $\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_*^{\frac{1}{2}}\|_2$. The first two inequalities use the triangular inequality for L_2 norm. Letting $\tilde{\boldsymbol{\eta}} = [\tilde{\boldsymbol{\eta}}_1^T, \ldots, \tilde{\boldsymbol{\eta}}_n^T]^T$, by the concentration inequality for Lipschitz function of Gaussian random vector (see Proposition 5.34 in [172]), we obtain

$$\mathbb{P}\left(f(\tilde{\boldsymbol{\eta}}) - \mathbb{E}f(\tilde{\boldsymbol{\eta}}) > t\right) \leq \exp\left(\frac{-t^2}{2\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_*^{\frac{1}{2}}\|_2^2}\right)$$
$$\implies \mathbb{P}\left(\sqrt{\sum_{i=1}^{n} \left\| \mathbf{\Sigma}^{-1} \mathbf{\Sigma}_*^{\frac{1}{2}} \tilde{\boldsymbol{\eta}}_i \right\|_2^2} - \mathbb{E}\sqrt{\sum_{i=1}^{n} \left\| \mathbf{\Sigma}^{-1} \mathbf{\Sigma}_*^{\frac{1}{2}} \tilde{\boldsymbol{\eta}}_i \right\|_2^2} > t\right) \leq \exp\left(\frac{-t^2}{2\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_*^{\frac{1}{2}}\|_2^2}\right)$$
$$\implies \mathbb{P}\left(\sqrt{\sum_{i=1}^{n} \|\mathbf{\Sigma}^{-1}\boldsymbol{\eta}_i\|_2^2} - \sqrt{\mathbb{E}\sum_{i=1}^{n} \mathrm{Tr}\left(\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_*^{\frac{1}{2}}\tilde{\boldsymbol{\eta}}_i\tilde{\boldsymbol{\eta}}_i^T\mathbf{\Sigma}_*^{\frac{1}{2}}\mathbf{\Sigma}^{-1}\right)} > t\right)$$
$$\leq \exp\left(\frac{-t^2}{2\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_*^{\frac{1}{2}}\|_2^2}\right)$$
$$\implies \mathbb{P}\left(\sqrt{\sum_{i=1}^{n} \|\mathbf{\Sigma}^{-1}\boldsymbol{\eta}_i\|_2^2} - \sqrt{n}\sqrt{\mathrm{Tr}\left(\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_*\mathbf{\Sigma}^{-1}\right)} > t\right) \leq \exp\left(\frac{-t^2}{2\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_*^{\frac{1}{2}}\|_2^2}\right)$$

where we use Jensen's inequality in the third step for bounding the expectation $\mathbb{E}f(\tilde{\boldsymbol{\eta}})$. Letting $t = \sqrt{\mathrm{Tr}\left(\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_*\mathbf{\Sigma}^{-1}\right) \cdot n}$ and $\tau = \|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_*^{\frac{1}{2}}\|_F / \|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_*^{\frac{1}{2}}\|_2$, with probability at least $1 - \exp\left(-\frac{n\tau^2}{2}\right)$, we have

$$\sqrt{\sum_{i=1}^{n} \|\mathbf{\Sigma}^{-1}\boldsymbol{\eta}_i\|_2^2} \leq 2\sqrt{n} \cdot \sqrt{\mathrm{Tr}\left(\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_*\mathbf{\Sigma}^{-1}\right)}\ , \tag{8.36}$$

where we use the relation $\mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_*\boldsymbol{\Sigma}^{-1}\right) = \|\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_*^{\frac{1}{2}}\|_F^2$. By applying a union bound to (8.35) and (8.36), with probability at least $1 - \exp\left(-\frac{n\tau^2}{2}\right) - c_3\exp(-\frac{c_1^2 w^2(\Omega)}{4\rho^2})$, the following inequality holds

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}_i\right\|_* \leq \frac{2c_0c_1c_2\cdot\kappa\sqrt{\mu_{\max}}}{\sqrt{n}}\cdot\sqrt{\mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_*\boldsymbol{\Sigma}^{-1}\right)}\cdot w(\Omega) \tag{8.37}$$

Finally we complete the proof by letting $C = 2c_0c_1c_2$, $C_1 = c_1$, and $C_2 = c_3$. ■

Appendix 8.B Proof of Noise Covariance Estimation

8.B.1 Proof of Theorem 26

Proof: By introducing the true parameter $\boldsymbol{\theta}^*$, $\hat{\boldsymbol{\Sigma}}$ can be rewritten as

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{\eta}_i + \mathbf{X}_i(\boldsymbol{\theta}^* - \boldsymbol{\theta})\right)\left(\boldsymbol{\eta}_i + \mathbf{X}_i(\boldsymbol{\theta}^* - \boldsymbol{\theta})\right)^T$$

And note that

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} \triangleq \mathbb{E}[\hat{\boldsymbol{\Sigma}}] = \boldsymbol{\Sigma}_* + \boldsymbol{\Delta}_{\boldsymbol{\theta}}, \quad \text{where} \quad \boldsymbol{\Delta}_{\boldsymbol{\theta}} = \mathbb{E}\left[\mathbf{X}(\boldsymbol{\theta}^* - \boldsymbol{\theta})(\boldsymbol{\theta}^* - \boldsymbol{\theta})^T\mathbf{X}^T\right].$$

The ψ_2-norm of $\boldsymbol{\Sigma}_*^{-\frac{1}{2}}\left(\boldsymbol{\eta} + \mathbf{X}(\boldsymbol{\theta}^* - \boldsymbol{\theta})\right)$ satisfies

$$\begin{aligned}
\left|\left|\left|\boldsymbol{\Sigma}_*^{-\frac{1}{2}}\left(\boldsymbol{\eta} + \mathbf{X}(\boldsymbol{\theta}^* - \boldsymbol{\theta})\right)\right|\right|\right|_{\psi_2} &\leq \left|\left|\left|\boldsymbol{\Sigma}_*^{-\frac{1}{2}}\boldsymbol{\eta}\right|\right|\right|_{\psi_2} + \left|\left|\left|\boldsymbol{\Sigma}_*^{-\frac{1}{2}}\mathbf{X}(\boldsymbol{\theta}^* - \boldsymbol{\theta})\right|\right|\right|_{\psi_2} \\
&= |||\tilde{\boldsymbol{\eta}}|||_{\psi_2} + \sup_{\mathbf{u}\in\mathbb{S}^{m-1}}\left|\left|\left|(\boldsymbol{\theta}^* - \boldsymbol{\theta})^T\boldsymbol{\Gamma}_{*\mathbf{u}}^{\frac{1}{2}}\boldsymbol{\Gamma}_{*\mathbf{u}}^{-\frac{1}{2}}\mathbf{X}^T\boldsymbol{\Sigma}_*^{-\frac{1}{2}}\mathbf{u}\right|\right|\right|_{\psi_2} \\
&\leq \kappa_0 + \sup_{\substack{\mathbf{v}\in\mathbb{S}^{p-1}\\ \mathbf{u}\in\mathbb{S}^{m-1}}}\left\|\boldsymbol{\Gamma}_{*\mathbf{u}}^{\frac{1}{2}}(\boldsymbol{\theta}^* - \boldsymbol{\theta})\right\|_2\cdot\left|\left|\left|\mathbf{v}^T\boldsymbol{\Gamma}_{*\mathbf{u}}^{-\frac{1}{2}}\mathbf{X}^T\boldsymbol{\Sigma}_*^{-\frac{1}{2}}\mathbf{u}\right|\right|\right|_{\psi_2} \\
&\leq \kappa_0 + \kappa\sup_{\mathbf{u}\in\mathbb{S}^{m-1}}\left\|\boldsymbol{\Gamma}_{*\mathbf{u}}^{\frac{1}{2}}\right\|_2\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2
\end{aligned}$$

$$\leq \kappa_0 + \kappa \sqrt{\frac{\mu_{\max}}{\lambda_{\min}\left(\mathbf{\Sigma}_*\right)}} \left\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\right\|_2$$

where $\mathbf{\Gamma}_{*\mathbf{u}} = \mathbb{E}[\mathbf{X}^T \mathbf{\Sigma}_*^{-\frac{1}{2}} \mathbf{u}\mathbf{u}^T \mathbf{\Sigma}_*^{-\frac{1}{2}} \mathbf{X}]$, and $\|\mathbf{\Gamma}_{*\mathbf{u}}\|_2^2 \leq \mu_{\max} \|\mathbf{\Sigma}_*^{-\frac{1}{2}} \mathbf{u}\|_2^2 \leq \frac{\mu_{\max}}{\lambda_{\min}(\mathbf{\Sigma}_*)}$ by the definition of sub-Gaussian $\mathbf{X}$. κ_0 is the ψ_2-norm of standard Gaussian random vector. By Theorem 5.39 and Remark 5.40 in [172], if $n \geq C_0^4 m \left(\kappa_0 + \kappa \sqrt{\frac{\mu_{\max}}{\lambda_{\min}(\mathbf{\Sigma}_*)}} \left\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\right\|_2\right)^4$, with probability at least $1 - 2\exp(-C_1 m)$, we have

$$\left\| \mathbf{\Sigma}_*^{-\frac{1}{2}} \left(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_{\boldsymbol{\theta}} \right) \mathbf{\Sigma}_*^{-\frac{1}{2}} \right\|_2 \leq C_0^2 \left(\kappa_0 + \kappa \sqrt{\frac{\mu_{\max}}{\lambda_{\min}\left(\mathbf{\Sigma}_*\right)}} \left\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\right\|_2 \right)^2 \sqrt{\frac{m}{n}} \tag{8.38}$$

Hence we have

$$\begin{aligned}
\lambda_{\max}\left(\mathbf{\Sigma}_*^{-\frac{1}{2}} \hat{\mathbf{\Sigma}} \mathbf{\Sigma}_*^{-\frac{1}{2}}\right) &= \left\| \mathbf{\Sigma}_*^{-\frac{1}{2}} \hat{\mathbf{\Sigma}} \mathbf{\Sigma}_*^{-\frac{1}{2}} \right\|_2 \\
&\leq 1 + \left\| \mathbf{\Sigma}_*^{-\frac{1}{2}} \left(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_{\boldsymbol{\theta}} \right) \mathbf{\Sigma}_*^{-\frac{1}{2}} \right\|_2 + \left\| \mathbf{\Sigma}_*^{-\frac{1}{2}} \mathbf{\Delta}_{\boldsymbol{\theta}} \mathbf{\Sigma}_*^{-\frac{1}{2}} \right\|_2 \\
&\leq 1 + C_0^2 \left(\kappa_0 + \kappa \sqrt{\frac{\mu_{\max}}{\lambda_{\min}\left(\mathbf{\Sigma}_*\right)}} \left\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\right\|_2 \right)^2 \sqrt{\frac{m}{n}} + \frac{\mu_{\max}}{\lambda_{\min}\left(\mathbf{\Sigma}_*\right)} \left\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\right\|_2^2 \\
&\overset{(a)}{\leq} 1 + 2C_0^2 \kappa_0^2 \sqrt{\frac{m}{n}} + \frac{2C_0^2 \kappa^2 \mu_{\max}}{\lambda_{\min}\left(\mathbf{\Sigma}_*\right)} \left\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\right\|_2^2 \sqrt{\frac{m}{n}} + \frac{\mu_{\max}}{\lambda_{\min}\left(\mathbf{\Sigma}_*\right)} \left\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\right\|_2^2 \\
&\leq 1 + 2C_0^2 \kappa_0^2 \sqrt{\frac{m}{n}} + \left(\frac{\mu_{\min}}{\lambda_{\max}\left(\mathbf{\Sigma}_*\right)} + \frac{\mu_{\max}}{\lambda_{\min}\left(\mathbf{\Sigma}_*\right)} \right) \left\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\right\|_2^2 \\
&\leq 1 + C^2 \kappa_0^2 \sqrt{\frac{m}{n}} + \frac{2\mu_{\max}}{\lambda_{\min}\left(\mathbf{\Sigma}_*\right)} \left\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\right\|_2^2
\end{aligned}$$

$$\begin{aligned}
\lambda_{\min}\left(\boldsymbol{\Sigma}_*^{-\frac{1}{2}}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}_*^{-\frac{1}{2}}\right) &\geq 1+\lambda_{\min}\left(\boldsymbol{\Sigma}_*^{-\frac{1}{2}}\left(\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right)\boldsymbol{\Sigma}_*^{-\frac{1}{2}}\right)+\lambda_{\min}\left(\boldsymbol{\Sigma}_*^{-\frac{1}{2}}\boldsymbol{\Delta}_{\boldsymbol{\theta}}\boldsymbol{\Sigma}_*^{-\frac{1}{2}}\right)\\
&\geq 1-\left\|\boldsymbol{\Sigma}_*^{-\frac{1}{2}}\left(\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right)\boldsymbol{\Sigma}_*^{-\frac{1}{2}}\right\|_2+\frac{\mu_{\min}}{\lambda_{\max}\left(\boldsymbol{\Sigma}_*\right)}\left\|\boldsymbol{\theta}^*-\boldsymbol{\theta}\right\|_2^2\\
&\geq 1-C_0^2\left(\kappa_0+\kappa\sqrt{\frac{\mu_{\max}}{\lambda_{\min}\left(\boldsymbol{\Sigma}_*\right)}}\left\|\boldsymbol{\theta}^*-\boldsymbol{\theta}\right\|_2\right)^2\sqrt{\frac{m}{n}}+\frac{\mu_{\min}}{\lambda_{\max}\left(\boldsymbol{\Sigma}_*\right)}\left\|\boldsymbol{\theta}^*-\boldsymbol{\theta}\right\|_2^2\\
&\overset{(b)}{\geq} 1-2C_0^2\kappa_0^2\sqrt{\frac{m}{n}}-\frac{2C_0^2\kappa^2\mu_{\max}}{\lambda_{\min}\left(\boldsymbol{\Sigma}_*\right)}\left\|\boldsymbol{\theta}^*-\boldsymbol{\theta}\right\|_2^2\sqrt{\frac{m}{n}}+\frac{\mu_{\min}}{\lambda_{\max}\left(\boldsymbol{\Sigma}_*\right)}\left\|\boldsymbol{\theta}^*-\boldsymbol{\theta}\right\|_2^2\\
&\geq 1-C^2\kappa_0^2\sqrt{\frac{m}{n}}
\end{aligned}$$

where $C^2=2C_0^2$. For (a) and (b), we use the assumption $n\geq C^4m\kappa^4\left(\frac{\lambda_{\max}(\boldsymbol{\Sigma}_*)\mu_{\max}}{\lambda_{\min}(\boldsymbol{\Sigma}_*)\mu_{\min}}\right)^2=4C_0^4m\kappa^4\left(\frac{\lambda_{\max}(\boldsymbol{\Sigma}_*)\mu_{\max}}{\lambda_{\min}(\boldsymbol{\Sigma}_*)\mu_{\min}}\right)^2$. This completes the proof. ■

8.B.2 Proof of Lemma 24

Proof: Based on the definition of $\xi(\cdot)$, we have

$$\begin{aligned}
\xi\left(\hat{\boldsymbol{\Sigma}}\right) &= \frac{\sqrt{\operatorname{Tr}\left(\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}_*\hat{\boldsymbol{\Sigma}}^{-1}\right)}}{\operatorname{Tr}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)}=\frac{1}{\sqrt{\operatorname{Tr}\left(\boldsymbol{\Sigma}_*^{-1}\right)}}\cdot\sqrt{\frac{\operatorname{Tr}\left(\boldsymbol{\Sigma}_*^{-1}\right)\cdot\operatorname{Tr}\left(\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}_*\hat{\boldsymbol{\Sigma}}^{-1}\right)}{\operatorname{Tr}^2\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)}}\\
&=\xi\left(\boldsymbol{\Sigma}_*\right)\cdot\sqrt{\frac{\operatorname{Tr}\left(\hat{\boldsymbol{\Sigma}}^{\frac{1}{2}}\boldsymbol{\Sigma}_*^{-1}\hat{\boldsymbol{\Sigma}}^{\frac{1}{2}}\hat{\boldsymbol{\Sigma}}^{-1}\right)\cdot\operatorname{Tr}\left(\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_*\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\hat{\boldsymbol{\Sigma}}^{-1}\right)}{\operatorname{Tr}^2\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)}}\\
&\leq\xi\left(\boldsymbol{\Sigma}_*\right)\cdot\sqrt{\frac{\lambda_{\max}\left(\hat{\boldsymbol{\Sigma}}^{\frac{1}{2}}\boldsymbol{\Sigma}_*^{-1}\hat{\boldsymbol{\Sigma}}^{\frac{1}{2}}\right)\operatorname{Tr}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)\cdot\lambda_{\max}\left(\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_*\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\right)\operatorname{Tr}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)}{\operatorname{Tr}^2\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)}}\\
&=\xi\left(\boldsymbol{\Sigma}_*\right)\cdot\sqrt{\lambda_{\max}\left(\hat{\boldsymbol{\Sigma}}^{\frac{1}{2}}\boldsymbol{\Sigma}_*^{-1}\hat{\boldsymbol{\Sigma}}^{\frac{1}{2}}\right)\lambda_{\max}\left(\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_*\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\right)}\\
&=\xi\left(\boldsymbol{\Sigma}_*\right)\cdot\sqrt{\frac{\lambda_{\max}\left(\boldsymbol{\Sigma}_*^{-\frac{1}{2}}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}_*^{-\frac{1}{2}}\right)}{\lambda_{\min}\left(\boldsymbol{\Sigma}_*^{-\frac{1}{2}}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}_*^{-\frac{1}{2}}\right)}}
\end{aligned}$$

where the inequality follows from von Neumann's trace inequality. Now we can bound $\xi(\hat{\mathbf{\Sigma}})$ by invoking Theorem 26,

$$\begin{aligned}
\xi\left(\hat{\mathbf{\Sigma}}\right) &\leq \xi\left(\mathbf{\Sigma}_*\right) \cdot \sqrt{\frac{1 + C^2\kappa_0^2\sqrt{\frac{m}{n}} + \frac{2\mu_{\max}}{\lambda_{\min}(\mathbf{\Sigma}_*)}\left\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\right\|_2^2}{1 - C^2\kappa_0^2\sqrt{\frac{m}{n}}}} \\
&= \xi\left(\mathbf{\Sigma}_*\right) \cdot \sqrt{1 + \frac{2C^2\kappa_0^2\sqrt{\frac{m}{n}} + \frac{2\mu_{\max}}{\lambda_{\min}(\mathbf{\Sigma}_*)}\left\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\right\|_2^2}{1 - C^2\kappa_0^2\sqrt{\frac{m}{n}}}} \\
&\leq \xi\left(\mathbf{\Sigma}_*\right) \cdot \left(1 + \frac{\sqrt{2}C\kappa_0\left(\frac{m}{n}\right)^{\frac{1}{4}} + \sqrt{\frac{2\mu_{\max}}{\lambda_{\min}(\mathbf{\Sigma}_*)}}\left\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\right\|_2}{\sqrt{1 - C^2\kappa_0^2\sqrt{\frac{m}{n}}}}\right) \\
&\leq \xi\left(\mathbf{\Sigma}_*\right) \cdot \left(1 + 2C\kappa_0\left(\frac{m}{n}\right)^{\frac{1}{4}} + 2\sqrt{\frac{\mu_{\max}}{\lambda_{\min}\left(\mathbf{\Sigma}_*\right)}}\left\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\right\|_2\right)
\end{aligned}$$

where the last inequality follows from $n \geq 4C^4 m \cdot \left(\kappa_0 + \kappa\sqrt{\frac{\mu_{\max}}{\lambda_{\min}(\mathbf{\Sigma}_*)}}\left\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\right\|_2\right)^4 \geq 4C^4 m\kappa_0^4$. ■

Appendix 8.C Proof of AltEst Procedure

8.C.1 Proof of Theorem 27

Proof: Since $n \geq C^4 m\kappa^4\left(\frac{\lambda_{\max}(\mathbf{\Sigma}_*)\mu_{\max}}{\lambda_{\min}(\mathbf{\Sigma}_*)\mu_{\min}}\right)^2$ and $\hat{\mathbf{\Sigma}}_{(0)}$ is initialized as $\hat{\mathbf{\Sigma}}_{(0)} = \mathbf{I}_{m\times m}$, by applying Theorem 25 to $\hat{\boldsymbol{\theta}}_{(1)}$, we have

$$\begin{aligned}
\left\|\hat{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}^*\right\|_2 &\leq C_1\kappa\sqrt{\frac{\mu_{\max}}{\mu_{\min}^2}} \cdot \xi\left(\hat{\mathbf{\Sigma}}_{(0)}\right) \cdot \frac{\Psi \cdot w(\Omega)}{\sqrt{m}} \\
&= C_1\kappa\sqrt{\frac{\mu_{\max}}{\mu_{\min}^2}} \cdot \frac{\Psi \cdot w(\Omega)}{\sqrt{mn}} \\
&\leq C_1\kappa\sqrt{\frac{\mu_{\max}}{\mu_{\min}^2}} \cdot \frac{\Psi \cdot w(\Omega)}{\sqrt{m}} \cdot \frac{\lambda_{\min}\left(\mathbf{\Sigma}_*\right)\mu_{\min}}{C^2\sqrt{m}\cdot\kappa^2\lambda_{\max}\left(\mathbf{\Sigma}_*\right)\mu_{\max}} \\
&= \frac{C_1}{C^2} \cdot \frac{\lambda_{\min}\left(\mathbf{\Sigma}_*\right)}{\kappa\lambda_{\max}\left(\mathbf{\Sigma}_*\right)\sqrt{\mu_{\max}}} \cdot \frac{\Psi \cdot w(\Omega)}{m}
\end{aligned}$$

It follows that

$$n \geq C^4 m \cdot 4 \left(\kappa_0 + \frac{C_1}{C^2} \sqrt{\frac{\lambda_{\min}\left(\mathbf{\Sigma}_*\right)}{\lambda_{\max}^2\left(\mathbf{\Sigma}_*\right)}} \frac{\Psi w(\Omega)}{m} \right)^4 \quad \Longrightarrow$$

$$n \geq C^4 m \cdot 4 \left(\kappa_0 + \kappa \sqrt{\frac{\mu_{\max}}{\lambda_{\min}\left(\mathbf{\Sigma}_*\right)}} \left\| \boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_{(1)} \right\|_2 \right)^4$$

By applying Lemma 24 and Theorem 25 to the second iteration,

$$\left\| \hat{\boldsymbol{\theta}}_{(2)} - \boldsymbol{\theta}^* \right\|_2 \leq e_{\text{orc}} \cdot \left(1 + 2C\kappa_0 \left(\frac{m}{n} \right)^{\frac{1}{4}} + 2\sqrt{\frac{\mu_{\max}}{\lambda_{\min}\left(\mathbf{\Sigma}_*\right)}} \left\| \hat{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}^* \right\|_2 \right) \quad \Longrightarrow$$

$$\left\| \hat{\boldsymbol{\theta}}_{(2)} - \boldsymbol{\theta}^* \right\|_2 - e_{\min} \;\leq\; 2e_{\text{orc}} \sqrt{\frac{\mu_{\max}}{\lambda_{\min}\left(\mathbf{\Sigma}_*\right)}} \cdot \left(\left\| \hat{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}^* \right\|_2 - e_{\min} \right) .$$

Since $n \geq C^4 m \cdot \left(\frac{2C_1\kappa}{C^2} \cdot \frac{\mu_{\max}}{\mu_{\min}} \cdot \frac{\xi(\mathbf{\Sigma}_*)\Psi w(\Omega)}{\sqrt{m \cdot \lambda_{\min}(\mathbf{\Sigma}_*)}} \right)^2$, we have $2e_{\text{orc}} \sqrt{\frac{\mu_{\max}}{\lambda_{\min}(\mathbf{\Sigma}_*)}} \leq 1$, which indicates that $\left\| \hat{\boldsymbol{\theta}}_{(2)} - \boldsymbol{\theta}^* \right\|_2 \leq \left\| \hat{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}^* \right\|_2$. Therefore the condition in Lemma 24 on sample size n also holds for $\hat{\boldsymbol{\theta}}_{(2)}$ and so on. By repeatedly applying Lemma 24 and Theorem 25, we have the following inequality for every $t > 0$,

$$\left\| \hat{\boldsymbol{\theta}}_{(t+1)} - \boldsymbol{\theta}^* \right\|_2 - e_{\min} \;\leq\; 2e_{\text{orc}} \sqrt{\frac{\mu_{\max}}{\lambda_{\min}\left(\mathbf{\Sigma}_*\right)}} \cdot \left(\left\| \hat{\boldsymbol{\theta}}_{(t)} - \boldsymbol{\theta}^* \right\|_2 - e_{\min} \right) \tag{8.39}$$

By combining (8.39) for every t, we obtain

$$\left\| \hat{\boldsymbol{\theta}}_{(T)} - \boldsymbol{\theta}^* \right\|_2 - e_{\min} \;\leq\; \left(2e_{\text{orc}} \sqrt{\frac{\mu_{\max}}{\lambda_{\min}\left(\mathbf{\Sigma}_*\right)}} \right)^{T-1} \cdot \left(\left\| \hat{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}^* \right\|_2 - e_{\min} \right)$$

which completes the proof. ■

Chapter 9

Improved Estimation for Structured Multi-Response Linear Models

9.1 Introduction

In this chapter, we continue to focus the multi-response linear model [5, 25, 79] with m real-valued outputs,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta}_* + \boldsymbol{\eta} \ , \quad \text{where} \ \ \boldsymbol{\eta} = \boldsymbol{\Sigma}_*^{1/2}\tilde{\boldsymbol{\eta}} \tag{9.1}$$

where $\mathbf{y} \in \mathbb{R}^m$ is the response vector and $\mathbf{X} \in \mathbb{R}^{m \times p}$ consists of m p-dimensional feature vectors. Compared with that given in Chapter 8, one difference here is the relaxed assumption on the noise vector $\boldsymbol{\eta}$. Instead of being Gaussian, $\boldsymbol{\eta}$ is now a linear transformation of an underlying zero-mean isotropic $\tilde{\boldsymbol{\eta}} \in \mathbb{R}^m$, which could be non-Gaussian. Given this relaxed model, our goal remains the estimation of the parameter

$\boldsymbol{\theta}_*$ under the unknown noise covariance $\boldsymbol{\Sigma}_*$, based on a sample $\mathcal{D} = \{(\mathbf{X}_i, \mathbf{y}_i)\}_{i=1}^n$. In this work, the true parameter $\boldsymbol{\theta}_*$ is assumed to possess certain low-complexity structure measured by some function $f : \mathbb{R}^p \mapsto \mathbb{R}_+$, which is not necessarily a norm as assumed in Chapter 8. Instead f can be even non-convex, e.g., L_0 cardinality function. In principle, we still adhere to the AltEst procedure to alternatingly estimate $\boldsymbol{\theta}_*$ and $\boldsymbol{\Sigma}_*$, but the GDS used in Chapter 8 needs to be replaced as it cannot handle the potential non-convexity of f. To this end, we switch the GDS to the constraint estimator in the AltEst framework, which gives rise to the following updates

$$\hat{\boldsymbol{\Sigma}}_{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\theta}}_{(t)}\right)\left(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\theta}}_{(t)}\right)^T , \tag{9.2}$$

$$\hat{\boldsymbol{\theta}}_{(t+1)} = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\operatorname{argmin}}\ \frac{1}{2n}\sum_{i=1}^{n}\left\|\hat{\boldsymbol{\Sigma}}_{(t+1)}^{-\frac{1}{2}}\left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\theta}\right)\right\|_2^2 \quad \text{s.t.} \quad f(\boldsymbol{\theta}) \leq \lambda\ . \tag{9.3}$$

In fact, this procedure is exactly the AltMin algorithm applied to the objective function below,

$$\left(\hat{\boldsymbol{\theta}}_{\text{cs}},\ \hat{\boldsymbol{\Sigma}}_{\text{cs}}\right) = \underset{\boldsymbol{\theta}\in\mathbb{R}^p,\ \boldsymbol{\Sigma}\succeq 0}{\operatorname{argmin}}\ \frac{1}{2n}\sum_{i=1}^{n}\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\theta}\right)\right\|_2^2 + \frac{1}{2}\log|\boldsymbol{\Sigma}| \quad \text{s.t.} \quad f(\boldsymbol{\theta}) \leq \lambda\ , \tag{9.4}$$

which corresponds to the constrained maximum likelihood estimator of $(\boldsymbol{\Sigma}, \boldsymbol{\theta})$ when the noise is Gaussian. With the replacement of the GDS, though the update (9.3) of $\hat{\boldsymbol{\theta}}_{(t+1)}$ remains non-convex if f is so, the simplicity of its objective actually favors the optimization. More precisely, the recent progress in optimization with non-convex constraints enables various algorithms to find the global minimum under mild conditions on data [15, 86, 148].

As we have discussed in Chapter 8, the current statistical understanding of AltMin (and AltEst) falls short. The statistical guarantees for non-convex AltMin procedures are often shown under the notorious *resampling* assumption [45, 84, 128, 179, 180], which assumes that each iteration receives a fresh sample. Albeit this can be achieved by partitioning the data into disjoint subsets and using different batches in each update, people seldom do so in practice, as it usually results in worse performance than using all data in every iteration. From the theoretical perspective, the resampling-based analysis is neither a satisfactory explanation for the power of AltMin, since the probability with which the statistical guarantees hold often decays as iteration goes, which is unnatural to see.

In this chapter, we aim at a better way to bound the statistical error of the above AltMin procedure for general structure-inducing f. In principal, non-asymptotic statistical analyses for high dimension typically involve bounding suprema of stochastic processes [14, 127, 135, 173]. The difficulty of analyzing AltMin lies in the dependency between the data and the obtained iterates, and the lack of independence prevents applications of various concentration inequalities to bounding the supremum of the target processes. The resampling assumption facilitates the analysis of AltMin by assuming new data that are independent of previous iterates. In contrast to resampling, we here resort to uniformity to tackle the dependency issue, which ends up dealing with more complicated stochastic processes. By carefully applying generic chaining [161], an advanced tool from probability theory, we are able to obtain the desired bounds for the processes under consideration, and eventually express the error bound in terms of Gaussian width [40, 63]. In particular, we analyze the AltMin procedure under two different choices of initialization, one with an arbitrarily initialized iterate and the other starting at a point close to $\boldsymbol{\theta}_*$. The L_2-error for both types of AltMin is shown to converge *geometrically* to certain *minimum achievable error* $e_{\min}$ with overwhelming probability, i.e.,

$$\left\|\hat{\boldsymbol{\theta}}_{(T)} - \boldsymbol{\theta}_*\right\|_2 \leq e_{\min} + \rho_n^T \cdot \left(\left\|\hat{\boldsymbol{\theta}}_{(0)} - \boldsymbol{\theta}_*\right\|_2 - e_{\min}\right) \tag{9.5}$$

where $\rho_n < 1$ is the contraction factor and $e_{\min}$ is given by

$$e_{\min} = O\left(\frac{w(\mathcal{C}) + m}{\sqrt{n}}\right) \qquad \text{(arbitrary initialization)} ,$$
$$e_{\min} = O\left(\frac{w(\mathcal{C})}{\sqrt{n}}\right) \qquad \text{(good initialization)} .$$

Here $w(\mathcal{C})$ is the Gaussian width of a set $\mathcal{C}$ related the structure of $\boldsymbol{\theta}_*$. Surprisingly the error for good initializations matches the resampling-based result up to some constant, which requires more fresh data to achieve such a bound. In summary, this work improves the results in Chapter 8 in several aspects. First, our analysis for AltMin does not rely on the resampling assumption, which can be adapted with suitable modifications to obtain resampling-free results for the original AltEst as well. Second our statistical guarantees work for general sub-Gaussian noise. Third, we allow the complexity function f to be non-convex, whereas in Chapter 8 f is required to be a norm. Moreover, our result suggests that when the amount of data is adequate the AltMin with arbitrary initialization can even achieve the same level of error as the well-initialized one, which is not discovered in the earlier study.

The rest of the chapter is organized as follows. In Section 9.2, we outline the strategies for combating non-convexity and present the algorithmic details of the AltMin procedure for structured multi-response regression. In Section 9.3, we present the deterministic statistical guarantees for the AltMin algorithm, and instantiate the error bounds under probabilistic assumptions in Section 9.4. Finally we provide some experimental results in Section 9.5. All proofs are deferred to the appendix.

9.2 Strategy to Conquer Non-Convexity

For many statistical estimation problems, we can construct the estimator of the underlying model parameter by minimizing certain loss function over the given sample

$\mathcal{D}$,

$$\hat{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}\in\mathcal{W}} \; L(\mathbf{w}; \mathcal{D}) \; . \tag{9.6}$$

For non-convex problem with the associated objective function L being non-convex, finding its global minimizer is challenging in general due to spurious local minima, which can be poor estimates of the true parameter. Thanks to the stochastic models assumed for the observed data $\mathcal{D}$, however, the scenario we face is often much more benign than the worst case that causes the failure of the optimization algorithms. Therefore it is widely believed that non-convex estimation can be done through the usual local search method with suitable initialization point. Since our ultimate goal is the statistical recovery instead of the optimization performance itself, it is reasonable to leave out the "unfriendly" data which our model is unlikely to encounter.

In order to show the recovery guarantee for non-convex estimation, there are mainly two commonly-used strategies. One strategy is to show certain local convergence in a neighborhood $\mathcal{N}$ of the global minimizer $\hat{\mathbf{w}}$ of (9.6) [34, 116, 130, 169, 189]. With a proper initialization inside $\mathcal{N}$, subsequent iterates produced by some local search might be able to converge to $\hat{\mathbf{w}}$, whose statistical error is expected to be small. This strategy is particularly suitable for the noiseless setting, as $\hat{\mathbf{w}}$ is equal to $\mathbf{w}_*$, and most of the existing works use gradient descent type or its variants as workhorse algorithms. The other strategy is to show that there is no spurious local minima of L under the assumed the statistical models, so that any optimization algorithms that provably converge to local minima will suffice for a good estimation [20, 59, 60, 105, 156, 157].

For our multi-response regression problem, however, it is difficult to apply the aforementioned strategies. First, bounding the statistical error of the global minimizer is nontrivial in the noisy setting, especially when the objective $L(\mathbf{w})$ involves more than one set of variables like the multi-response regression, let alone characterizing the equivalence of all local minima. Second, the gradient-based local search is inefficient for the problem (9.4), since the update of $\mathbf{\Sigma}$ involves matrix inversion and projection onto positive semidefinite (PSD) cone. In contrast, AltMin procedure has a closed-form solution to $\mathbf{\Sigma}$-step, which is preferred in this setting.

In this work, we consider another strategy for the non-convex estimation in which $\mathbf{w}$ is composed of two parameters ($\mathbf{a}$ and $\mathbf{b}$), and L is jointly non-convex over them but might be marginally convex w.r.t. $\mathbf{a}$ ($\mathbf{b}$) when $\mathbf{b}$ ($\mathbf{a}$) is fixed. When the marginal subproblems are easy to solve, alternating minimization procedure is appealing for the purpose of estimation, which applies to the multi-response regression. The AltMin algorithm typically executes the following updates,

$$\hat{\mathbf{a}}_{(t+1)} = \underset{\mathbf{a}\in\mathcal{A}}{\operatorname{argmin}}\ L(\mathbf{a}, \hat{\mathbf{b}}_{(t)}; \mathcal{D}) \tag{9.7}$$

$$\hat{\mathbf{b}}_{(t+1)} = \underset{\mathbf{b}\in\mathcal{B}}{\operatorname{argmin}}\ L(\hat{\mathbf{a}}_{(t+1)}, \mathbf{b}; \mathcal{D}) \tag{9.8}$$

The basic idea for showing the statistical guarantees of AltMin is to derive the error bounds for both $\mathbf{a}$- and $\mathbf{b}$-step when the other is fixed to the latest estimate. Since both subproblems (9.7) and (9.8) are usually simpler, the separate errors might be easier to characterize than considered jointly, which are ideally of the form,

$$d_1\left(\hat{\mathbf{a}}_{(t+1)},\ \mathbf{a}_*\right) \leq e_1\left(d_2\left(\hat{\mathbf{b}}_{(t)}, \mathbf{b}_*\right)\right) \tag{9.9}$$

$$d_2\left(\hat{\mathbf{b}}_{(t+1)},\ \mathbf{b}_*\right) \leq e_2\left(d_1\left(\hat{\mathbf{a}}_{(t+1)}, \mathbf{a}_*\right)\right) \tag{9.10}$$

where a_* and b_* are true underlying parameters. The function d_1 (d_2) characterizes the closeness between $\hat{\mathbf{a}}_{(t+1)}$ and $\mathbf{a}_*$ ($\hat{\mathbf{b}}_{(t+1)}$ and $\mathbf{b}_*$), which is nonnegative with $d_1(\mathbf{a}_*, \mathbf{a}_*) = 0$ ($d_2(\mathbf{b}_*, \mathbf{b}_*) = 0$) but not necessarily a metric. The upper bound e_1 (e_2) may depend on other quantities such as n, but our emphasis is the dependence on the estimation accuracy of $\mathbf{b}$ ($\mathbf{a}$). It is natural to expect that e_1 (e_2) will shrink as $\hat{\mathbf{b}}_{(t)}$ ($\hat{\mathbf{a}}_{(t)}$) moves closer to $\mathbf{b}_*$ ($\mathbf{a}_*$). Under this condition, we can get

$$d_1(\hat{\mathbf{a}}_{(T)}, \mathbf{a}_*) \leq e_1\left(d_2\left(\hat{\mathbf{b}}_{(T-1)}, \mathbf{b}_*\right)\right) \leq \ldots\ldots \leq \underbrace{e_1\left(e_2\left(\ldots e_1\left(d_2\left(\hat{\mathbf{b}}_{(0)}, \mathbf{b}_*\right)\right)\ldots\right)\right)}_{\text{composition of } T\ e_1(\cdot) \text{ and } T-1\ e_2(\cdot)} \tag{9.11}$$

$$d_2(\hat{\mathbf{b}}_{(T)}, \mathbf{b}_*) \leq e_2\left(d_1\left(\hat{\mathbf{a}}_{(T)}, \mathbf{a}_*\right)\right) \leq \ldots\ldots \leq \underbrace{e_2\left(e_1\left(\ldots e_1\left(d_2\left(\hat{\mathbf{b}}_{(0)}, \mathbf{b}_*\right)\right)\ldots\right)\right)}_{\text{composition of } T\ e_2(\cdot) \text{ and } T\ e_1(\cdot)} \tag{9.12}$$

which may imply the error of $\hat{\mathbf{a}}_{(T)}$ and $\hat{\mathbf{b}}_{(T)}$ under other metrics of interest. Compared with the previous strategies, one notable difference of our treatment is that we do not care about the *optimization* convergence of AltMin, as we neither characterize the error of any local minimizers of $L(\cdot)$ nor show any iterate convergence to those minimizers. Instead the ingredients we need are simply the *statistical* error bounds (9.9) and (9.10). Given this fact, our analysis can be extended to the alternating estimation (AltEst) procedure [45] that need not optimize a joint objective over $\mathbf{a}$ and $\mathbf{b}$, which certainly cannot be handled by the earlier strategies.

In order to get (9.9) and (9.10), the analysis for each AltMin step is often confronted with a technical challenge due to the dependency between data and the iterates obtained so far, which is bypassed by many existing analyses via the resampling assumption. Essentially the resampling-based result states that with any fixed $\hat{\mathbf{b}}_{(t)}$ ($\hat{\mathbf{a}}_{(t+1)}$), given a fresh sample $\mathcal{D}_{(t)}$ independent of $\hat{\mathbf{b}}_{(t)}$ ($\hat{\mathbf{a}}_{(t+1)}$), the next iterate $\hat{\mathbf{a}}_{(t+1)}$ ($\hat{\mathbf{b}}_{(t+1)}$) satisfies the corresponding bound in (9.9) ((9.10)) with high probability. To avoid the resampling, we leverage the idea of uniform bounds [171], which aims to show that given a sample $\mathcal{D}$, the bounds (9.9) and (9.10) hold *uniformly* with high probability for *all* possible value of the input $\hat{\mathbf{b}}_{(t)}$ and $\hat{\mathbf{a}}_{(t+1)}$. This argument asks for no fresh data in each iteration, and the probability of the error bounds being true does not deteriorate with growing number of iterations. For structured multi-response regression, we will focus on the

Algorithm 10 Alternating minimization for multi-response regression

Input: Number of iterations T, Data $\mathcal{D} = \{(\mathbf{X}_i, \mathbf{y}_i)\}_{i=1}^n$ and Tuning parameter λ

Output: Estimated $\hat{\boldsymbol{\theta}}_{(T)}$

1: Initialize $\hat{\boldsymbol{\theta}}_{(0)}$ (e.g., solving (9.3) with $\hat{\boldsymbol{\Sigma}}_{(0)} = \mathbf{I}$)
2: **for** t:= 1 to T **do**
3:    Compute $\hat{\boldsymbol{\Sigma}}_{(t)}$ according to (9.2)
4:    Compute $\hat{\boldsymbol{\theta}}_{(t)}$ by solving (9.3)
5: **end for**
6: **return** $\hat{\boldsymbol{\theta}}_{(T)}$

AltMin procedure shown in Algorithm 10. For the rest of the paper, C_0, C_1, c_0, c_1 and so on are reserved for absolute constants.

9.3 Deterministic Analysis

In this section, we apply the resampling-free analysis framework to the multi-response regression problem, for which $\mathbf{a} = \boldsymbol{\Sigma}$ and $\mathbf{b} = \boldsymbol{\theta}$. First we introduce a few notations. We denote the smallest and the largest eigenvalue of $\boldsymbol{\Sigma}_*$ as σ_*^- and σ_*^+, and assume $\text{Diag}(\boldsymbol{\Sigma}_*) = \mathbf{I}_{m\times m}$ throughout the chapter for simplicity. In addition, we drop the subscripts indexing the iteration, and analyze both $\boldsymbol{\Sigma}$-update and $\boldsymbol{\theta}$-update in a broader setting, where the other parameter is fixed as a generic input in certain regions, i.e.,

$$\hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\theta})(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\theta})^T \tag{9.13}$$

$$\hat{\boldsymbol{\theta}}(\boldsymbol{\Sigma}) = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\text{argmin}}\ \frac{1}{2n}\sum_{i=1}^{n}\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\theta})\right\|_2^2 \quad \text{s.t.} \quad f(\boldsymbol{\theta}) \le f(\boldsymbol{\theta}_*)\ , \tag{9.14}$$

Note that here the tuning parameter λ for $\boldsymbol{\theta}$-step is set as $\lambda = f(\boldsymbol{\theta}_*)$, which will be kept for the rest of the analysis. For instance, if $f = \|\cdot\|_0$, then λ has to be set to the sparsity of the true $\boldsymbol{\theta}_*$. The input regions we consider for $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ are simply given by

$$\mathcal{R} = \left\{ \boldsymbol{\theta} \in \mathbb{R}^p \;\middle|\; f(\boldsymbol{\theta}) \leq f(\boldsymbol{\theta}_*) \right\} \tag{9.15}$$

$$\mathcal{M}(e_0) = \left\{ \boldsymbol{\Sigma} \in \mathbb{R}^{m \times m} \;\middle|\; \boldsymbol{\Sigma} = \hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta}),\ f(\boldsymbol{\theta}) \leq f(\boldsymbol{\theta}_*),\ \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2 \leq e_0 \right\} \tag{9.16}$$

in which e_0 is the error tolerance to be specified for the initialization. Note that the input region $\mathcal{M}(e_0)$ implicitly depends on $\mathcal{R}$ as well as the sample $\mathcal{D} = \{(\mathbf{x}, \mathbf{y})\}_{i=1}^n$ used for computing $\hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta})$. All matrices in $\mathcal{M}(e_0)$ need to be invertible for the computation of (9.14), which will be guaranteed by the later analysis when the randomness of data is considered.

Definition 24 (distance functions) The distance function $d_1(\cdot, \cdot)$ for $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_*$ is defined as

$$d_1(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_*) = \frac{\xi(\boldsymbol{\Sigma})}{\xi(\boldsymbol{\Sigma}_*)} - 1 \ , \tag{9.17}$$

in which $\xi(\cdot)$ is given by

$$\xi(\boldsymbol{\Sigma}) = \frac{\sqrt{\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_*\boldsymbol{\Sigma}^{-1})}}{\mathrm{Tr}(\boldsymbol{\Sigma}^{-1})} \ . \tag{9.18}$$

The distance function $d_2(\cdot, \cdot)$ for $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_*$ is defined as the standard L_2-distance, i.e.,

$$d_2(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_*) = \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2 \tag{9.19}$$

It is worth noting that $\xi(\boldsymbol{\Sigma})$ is minimized at $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_*$. The error bound of $\hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta})$ relies on the definitions below.

Definition 25 (error spherical cap) For a complexity function f, its error spherical cap is defined as

$$\mathcal{C} = \mathrm{cone}\left\{ \mathbf{u} \in \mathbb{R}^p \;\middle|\; f(\boldsymbol{\theta}_* + \mathbf{u}) \leq f(\boldsymbol{\theta}_*) \right\} \ \cap\ \mathbb{S}^{p-1} \ , \tag{9.20}$$

where $\mathbb{S}^{p-1} = \{\mathbf{u} \mid \|\mathbf{u}\|_2 = 1\}$ is the unit sphere of $\mathbb{R}^p$.

The cone in the definition above is sometimes called descent cone in the literature [4], which is critical to the analysis of many high-dimensional estimation problems [40,130]. The next definition is directly extended from the notion of *restricted eigenvalue* (RE) [21,139].

Definition 26 (uniformly restricted eigenvalue) For designs $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$, the smallest uniformly restricted eigenvalue (URE) the for error spherical cap $\mathcal{C} \subseteq \mathbb{S}^{p-1}$ is defined as

$$\alpha_n^- \triangleq \inf_{\mathbf{v}\in\mathbb{S}^{m-1}} \inf_{\mathbf{u}\in\mathcal{C}} \mathbf{u}^T \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{v}\mathbf{v}^T \mathbf{X}_i \right) \mathbf{u} \tag{9.21}$$

Similarly the largest URE is given as

$$\alpha_n^+ \triangleq \sup_{\mathbf{v}\in\mathbb{S}^{m-1}} \sup_{\mathbf{u}\in\mathcal{C}} \mathbf{u}^T \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{v}\mathbf{v}^T \mathbf{X}_i \right) \mathbf{u} \tag{9.22}$$

In comparison with the standard restricted eigenvalue, the uniformity of the URE is reflected by the infimum and the supremum operation over $\mathbf{v} \in \mathbb{S}^{m-1}$ in the above definitions.

Definition 27 (type-I noise-design interaction strength) For designs $\mathbf{X}_1$, $\mathbf{X}_2$, $\ldots$, $\mathbf{X}_n$ and untransformed noises $\tilde{\boldsymbol{\eta}}_1$, $\tilde{\boldsymbol{\eta}}_2$, $\ldots$, $\tilde{\boldsymbol{\eta}}_n$, the type-I noise-design interaction (NDI) strength is defined as

$$\gamma_n \triangleq \sup_{\mathbf{u}\in\mathcal{C}} \left\| \frac{2}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{u} \tilde{\boldsymbol{\eta}}_i^T \right\|_2 \tag{9.23}$$

With the definitions presented above, we are ready to give the deterministic guarantee for (9.13).

Lemma 25 (error bound for $\mathbf{\Sigma}$-estimation) *Given data $\{(\mathbf{X}_i, \mathbf{y}_i)\}_{i=1}^n$, let $\{\delta_n\}$ be a sequence such that*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \tilde{\boldsymbol{\eta}}_i \tilde{\boldsymbol{\eta}}_i^T - \mathbf{I} \right\|_2 \leq \delta_n \ . \tag{9.24}$$

If $\frac{\delta_n \alpha_n^-}{\gamma_n^2} \geq \frac{\sigma_*^+}{4\sigma_*^-}$ and $\delta_n \leq \frac{1}{4}$, then $\hat{\mathbf{\Sigma}}(\boldsymbol{\theta})$ given by (9.13) is invertible for any $\boldsymbol{\theta} \in \mathcal{R}$ and its error satisfies

$$d_1\left(\hat{\mathbf{\Sigma}}(\boldsymbol{\theta}),\ \mathbf{\Sigma}_*\right) \leq 4\delta_n + 2\sqrt{\frac{\alpha_n^+}{\sigma_*^-}} \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2 \ . \tag{9.25}$$

To analyze the error of $\hat{\boldsymbol{\theta}}(\mathbf{\Sigma})$, we assume that the global minimum of (9.14) can be attained despite the possible non-convexity of the constraint, which is fairly reasonable in view of the recent development on non-convex optimization [15, 86]. In addition, we need the definition of another noise-design interaction strength.

Definition 28 (type-II noise-design interaction strength) For designs $\mathbf{X}_1$, $\mathbf{X}_2$, …, $\mathbf{X}_n$ and noises $\boldsymbol{\eta}_1$, $\boldsymbol{\eta}_2$, …, $\boldsymbol{\eta}_n$, the type-II noise-design interaction (NDI) strength β_n for a set of matrices $\mathcal{K}$ is defined as

$$\beta_n(\mathcal{K}) \ \triangleq \sup_{\mathbf{\Sigma}\in\mathcal{K}} \sup_{\mathbf{u}\in\mathcal{C}} \ \frac{2}{n} \sum_{i=1}^{n} \frac{\boldsymbol{\eta}_i^T \mathbf{\Sigma}^{-1} \mathbf{X}_i \mathbf{u}}{\|\mathbf{\Sigma}_*^{1/2} \mathbf{\Sigma}^{-1}\|_F} \ , \tag{9.26}$$

where the invertibility is assumed for every $\mathbf{\Sigma} \in \mathcal{K}$.

In the analysis, we specifically focus on $\beta_n(\mathcal{M}(e_0))$ as $\mathcal{M}(e_0)$ is the set of input $\mathbf{\Sigma}$ under consideration. From its definition, it is not difficult to see that $\beta_n(\mathcal{M}(e_0))$ is a monotonically increasing function of e_0, as $\mathcal{M}(e_0) \subseteq \mathcal{M}(e_0')$ for any $e_0 \leq e_0'$. In the probabilistic analysis, we will bound $\beta_n(\mathcal{M}(e_0))$ at specific values of e_0. Based on the definition of β_n, the next lemma characterizes the estimation error for (9.14).

Lemma 26 (error bound for $\boldsymbol{\theta}$-estimation) *Given data* $\{(\mathbf{X}_i, \mathbf{y}_i)\}_{i=1}^n$ *and a set* $\mathcal{K} \subseteq \mathbb{R}^{m\times m}$ *such that every* $\boldsymbol{\Sigma} \in \mathcal{K}$ *is invertible, the following error bound holds for* $\hat{\boldsymbol{\theta}}(\boldsymbol{\Sigma})$ *given by* (9.14) *with any input* $\boldsymbol{\Sigma} \in \mathcal{K}$*,*

$$d_2\left(\hat{\boldsymbol{\theta}}(\boldsymbol{\Sigma}),\ \boldsymbol{\theta}_*\right) \leq \xi(\boldsymbol{\Sigma}) \cdot \frac{\beta_n(\mathcal{K})}{\alpha_n^-}\ , \tag{9.27}$$

where $\xi(\boldsymbol{\Sigma})$ *is defined in* (9.18)*. In particular, the error for* $\hat{\boldsymbol{\theta}}(\boldsymbol{\Sigma})$ *with any input* $\boldsymbol{\Sigma} \in \mathcal{M}(e_0)$ *satisfies*

$$d_2\left(\hat{\boldsymbol{\theta}}(\boldsymbol{\Sigma}),\ \boldsymbol{\theta}_*\right) \leq \xi(\boldsymbol{\Sigma}) \cdot \frac{\beta_n(\mathcal{M}(e_0))}{\alpha_n^-}\ . \tag{9.28}$$

Remark: Apart from $\mathcal{K} = \mathcal{M}(e_0)$, other specific instantiations of this lemma also yield interesting error bounds. For example, setting $\mathcal{K} = \{\mathbf{I}\}$ bounds the error of the constrained ordinary least squares (OLS), i.e.,

$$\left\|\hat{\boldsymbol{\theta}}_{\mathrm{odn}} - \boldsymbol{\theta}_*\right\|_2 \leq \xi(\mathbf{I}) \cdot \frac{\beta_n(\{\mathbf{I}\})}{\alpha_n^-} = \frac{1}{\sqrt{m}} \cdot \frac{\beta_n(\{\mathbf{I}\})}{\alpha_n^-} \ \triangleq\ e_{\mathrm{odn}}\ , \tag{9.29}$$

where $\hat{\boldsymbol{\theta}}_{\mathrm{odn}}$ is given by

$$\hat{\boldsymbol{\theta}}_{\mathrm{odn}} = \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p} \frac{1}{2n}\sum_{i=1}^n \|\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\theta}\|_2^2 \quad \text{s.t.} \quad f(\boldsymbol{\theta}) \leq f(\boldsymbol{\theta}_*)\ . \tag{9.30}$$

If we choose $\mathcal{K} = \{\boldsymbol{\Sigma}_*\}$, the error bound corresponds to the oracle estimator with the information $\boldsymbol{\Sigma}_*$, i.e.,

$$\left\|\hat{\boldsymbol{\theta}}_{\mathrm{orc}} - \boldsymbol{\theta}_*\right\|_2 \leq \xi(\boldsymbol{\Sigma}_*) \cdot \frac{\beta_n(\{\boldsymbol{\Sigma}_*\})}{\alpha_n^-} = \frac{1}{\sqrt{\mathrm{Tr}(\boldsymbol{\Sigma}_*^{-1})}} \cdot \frac{\beta_n(\{\boldsymbol{\Sigma}_*\})}{\alpha_n^-} \ \triangleq\ e_{\mathrm{orc}}\ , \tag{9.31}$$

in which $\hat{\boldsymbol{\theta}}_{\mathrm{orc}}$ is defined as

$$\hat{\boldsymbol{\theta}}_{\mathrm{orc}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\operatorname{argmin}} \ \frac{1}{2n} \sum_{i=1}^{n} \left\| \boldsymbol{\Sigma}_*^{-\frac{1}{2}} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\theta}) \right\|_2^2 \quad \text{s.t.} \quad f(\boldsymbol{\theta}) \leq f(\boldsymbol{\theta}_*) \ . \tag{9.32}$$

Equipped with the error bounds for both $\boldsymbol{\theta}$- and $\boldsymbol{\Sigma}$-step, we have the following theorem for the whole AltMin procedure.

Theorem 28 (deterministic error bound for AltMin) *Define ε_n, ρ_n and $e_{\min}$ as*

$$\varepsilon_n = \xi(\boldsymbol{\Sigma}_*) \cdot \frac{\beta_n(\mathcal{M}(e_0))}{\alpha_n^-} \ , \qquad \rho_n = 2\varepsilon_n \sqrt{\frac{\alpha_n^+}{\sigma_*^-}} \ , \qquad e_{\min} = \varepsilon_n \cdot \frac{1 + 4\delta_n}{1 - \rho_n}$$

in which δ_n is defined in Lemma 25. Assume that the tuning parameter $\lambda = f(\boldsymbol{\theta}_*)$, and the initialization satisfies both $f(\hat{\boldsymbol{\theta}}_{(0)}) \leq f(\boldsymbol{\theta}_*)$ and $\|\hat{\boldsymbol{\theta}}_{(0)} - \boldsymbol{\theta}_*\|_2 \leq e_0$. Under the conditions that $e_{\min} < e_0$, $\rho_n < 1$, $\frac{\delta_n \alpha_n^-}{\gamma_n^2} \geq \frac{\sigma_*^+}{4\sigma_*^-}$ and $\delta_n \leq \frac{1}{4}$, then $\hat{\boldsymbol{\theta}}_{(T)}$ returned by Algorithm 10 satisfies

$$\left\| \hat{\boldsymbol{\theta}}_{(T)} - \boldsymbol{\theta}_* \right\|_2 \leq e_{\min} + \rho_n^T \cdot (e_0 - e_{\min}) \ , \tag{9.33}$$

Remark: The above inequality indicates that the upper bound of the error for AltMin procedure will decrease geometrically to the *minimum achievable error* $e_{\min}$, provided that there exists room for improvement (i.e., $e_{\min} < e_0$). Note that $e_{\min}$ is given in a multiplicative form in terms of ε_n, which is similar to the bound for the error e_{orc} incurred by the oracle estimator. The contraction factor ρ_n not only controls the convergence rate of error, but also affects the value of $e_{\min}$. The theorem also reveals the role of e_0, which is calibrating the quality of initialization. The better the initialization is, the smaller the error $e_{\min}$ is.

In the next section, we will verify the conditions in Theorem 28 under suitable stochastic assumptions, so that the above error bound is valid.

9.4 Probabilistic Analysis

9.4.1 Preliminaries

In order for the deterministic results to hold, we need the conditions stated in Theorem 28 to be satisfied. The proposition below translates those requirements into the desired individual growth rates of α_n^-, α_n^+, β_n, γ_n and δ_n, which need to hold (with high probability) when the randomness of $\mathbf{X}$ and $\mathbf{y}$ is considered.

Proposition 17 *For any fixed e_0 and an initialization with $f(\hat{\boldsymbol{\theta}}_{(0)}) \leq f(\boldsymbol{\theta}_*)$ and $\|\hat{\boldsymbol{\theta}}_{(0)} - \boldsymbol{\theta}_*\|_2 \leq e_0$, the error bound (9.33) holds with large enough n, if α_n^-, α_n^+, δ_n, γ_n and $\beta_n(\mathcal{M}(e_0))$ satisfy the following conditions,*

(i) The smallest and the largest URE: $\alpha_n^- = \Theta(1)$ and $\alpha_n^+ = \Theta(1)$

(ii) The rate of convergence for $\left\|\frac{1}{n}\sum_{i=1}^n \tilde{\boldsymbol{\eta}}_i \tilde{\boldsymbol{\eta}}_i^T - \mathbf{I}\right\|_2$: $\delta_n = o(1)$

(iii) The type-I noise-design interaction strength: $\gamma_n = o(\delta_n^{1/2})$

(iv) The type-II noise-design interaction strength: $\beta_n(\mathcal{M}(e_0)) = o(1)$

The analysis of these conditions is built upon the concept of sub-Gaussian vectors and matrices, which are defined below.

Definition 29 (sub-Gaussian vector and matrix) A vector $\mathbf{x} \in \mathbb{R}^p$ is said to be sub-Gaussian if its ψ_2-norm satisfies,

$$|||\mathbf{x}|||_{\psi_2} = \sup_{\mathbf{u} \in \mathbb{S}^{p-1}} |||\langle \mathbf{x}, \mathbf{u} \rangle|||_{\psi_2} \leq \kappa < +\infty \ , \tag{9.34}$$

where $|||\cdot|||_{\psi_2}$ is defined for a random variable $x \in \mathbb{R}$ as $|||x|||_{\psi_2} = \sup_{q \geq 1} \frac{(\mathbb{E}|x|^q)^{\frac{1}{q}}}{\sqrt{q}}$. A matrix $\mathbf{X} \in \mathbb{R}^{m \times p}$ is sub-Gaussian if the following ψ_2-norm for $\mathbf{X}$ is finite,

$$|||\mathbf{X}|||_{\psi_2} = \sup_{\substack{\mathbf{u} \in \mathbb{S}^{p-1} \\ \mathbf{v} \in \mathbb{S}^{m-1}}} \left|\left|\left| \mathbf{u}^T \mathbf{\Gamma}_{\mathbf{v}}^{-\frac{1}{2}} \mathbf{X}^T \mathbf{v} \right|\right|\right|_{\psi_2} \leq \kappa < +\infty \ , \tag{9.35}$$

where $\mathbf{\Gamma}_{\mathbf{v}} = \mathbb{E}[\mathbf{X}^T \mathbf{v} \mathbf{v}^T \mathbf{X}]$. Further, $\mathbf{\Gamma}_{\mathbf{v}}$ for any $\mathbf{v} \in \mathbb{S}^{m-1}$ is assumed to satisfy

$$0 < \mu^- \leq \lambda_{\min}(\mathbf{\Gamma}_{\mathbf{v}}) \leq \lambda_{\max}(\mathbf{\Gamma}_{\mathbf{v}}) \leq \mu^+ < +\infty \ , \tag{9.36}$$

where μ^- and μ^+ are some constants.

This definition is adopted from [85,172]. If rows of $\mathbf{X}$ are i.i.d. copies of an isotropic sub-Gaussian random vector $\mathbf{x}$ with $|||\mathbf{x}|||_{\psi_2} \leq \kappa$, it is not difficult to verify that $|||\mathbf{X}|||_{\psi_2} \leq C\kappa$ for a universal constant C, and $\mu^- = \mu^+ = 1$. Our assumptions on the randomness of $\{\mathbf{X}_i\}$ and $\{\tilde{\boldsymbol{\eta}}_i\}$ are given below.

(A1) The designs $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are i.i.d. copies of a sub-Gaussian $\mathbf{X}$ with parameter κ, μ^- and μ^+.

(A2) The isotropic noises $\tilde{\boldsymbol{\eta}}_1, \tilde{\boldsymbol{\eta}}_2, \ldots, \tilde{\boldsymbol{\eta}}_n$ are i.i.d. copies of a sub-Gaussian $\tilde{\boldsymbol{\eta}}$ with parameter τ.

Since the definitions of α_n^-, α_n^+, γ_n and β_n involve the error spherical cap $\mathcal{C}$, it is expected that certain complexity measure of $\mathcal{C}$ will show up in the analysis, which turns out to be the notion of Gaussian width given in Definition 11. In the next two subsections, we show that the conditions $(i)-(iv)$ hold with overwhelming probability by characterizing the corresponding *non-asymptotic* bounds.

9.4.2 Arbitrarily-Initialized AltMin

The lemma below justifies the claim of the condition (i).

Lemma 27 *Under the assumption (A1), if the sample size* $n \geq C_0 \max\left\{\kappa^4 \left(\frac{\mu^+}{\mu^-}\right)^2, 1\right\} \cdot \max\left\{w^2(\mathcal{C}), m\right\}$, *with probability at least* $1 - 2\exp\left(-C_1 \max\left\{w^2(\mathcal{C}), m\right\}\right)$, *the smallest and the largest URE satisfy*

$$\frac{1}{2}\mu^- \;\leq\; \alpha_n^- \;\leq\; \alpha_n^+ \;\leq\; \frac{3}{2}\mu^+ \;, \tag{9.37}$$

where $w(\mathcal{C})$ *is the Gaussian width of the error spherical cap.*

The condition (ii) is simply implied by the following bound for the convergence of sample covariance matrix, which is a direct result of Lemma 5.36 and Theorem 5.39 in [172].

Proposition 18 *Under the assumption (A2), there exist absolute constants* C_0, C_1 *and* C_2 *such that if* $n \geq C_0\tau^4 m$, *the following inequality holds with probability at least* $1 - 2\exp(-C_1 m)$,

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{\boldsymbol{\eta}}_i\tilde{\boldsymbol{\eta}}_i^T - \mathbf{I}\right\|_2 \leq C_2\tau^2\sqrt{\frac{m}{n}} \triangleq \delta_n \tag{9.38}$$

Next we show that the rate of γ_n also has a $\frac{1}{\sqrt{n}}$-dependence as δ_n, thus implying that $\gamma_n = o(\delta_n^{1/2})$ in the condition (iii).

Lemma 28 *Under the assumptions (A1) and (A2), if* $n \geq C_0 m$, *the following inequality holds with probability at least* $1 - 2\exp(-C_1 m)$ *for the type-I NDI strength* γ_n,

$$\gamma_n \leq C_2 \cdot \frac{\kappa\tau\sqrt{\mu^+}(\sqrt{m} + w(\mathcal{C}))}{\sqrt{n}} \;. \tag{9.39}$$

Lastly we verify the condition (iv). Given the statement of Theorem 28, we need to bound $\beta_n(\mathcal{M}(e_0))$ for $e_0 = +\infty$ if allowing arbitrary initializations.

Lemma 29 *Suppose that the conditions of Lemma 25 are satisfied with probability* $1 - \epsilon$ *when* $n \geq n_0$. *Under the assumptions (A1) and (A2), if sample size* $n \geq$

$\max\left\{n_0, C_0\tau^4 m\right\}$, *the type-II NDI strength for* $\mathcal{M}(e_0)$ *with* $e_0 = +\infty$ *satisfies,*

$$\beta_n(\mathcal{M}(e_0)) \leq C_3 \cdot \frac{\kappa\sqrt{\mu^+}\,(m + w(\mathcal{C}))}{\sqrt{n}} \ , \tag{9.40}$$

with probability at least $1 - \epsilon - C_2 \exp(-C_1 m)$.

Remark: The proof of Lemma 29 suggests that β_n for any singleton $\mathcal{K}$ satisfies

$$\beta_n(\mathcal{K}) \leq C_3' \cdot \frac{\kappa\sqrt{\mu^+} \cdot w(\mathcal{C})}{\sqrt{n}} \ , \tag{9.41}$$

with probability $1 - C_2' \exp(-C_1' m)$ if $n \geq C_0' \tau^4 m$. Combined with Lemma 26 and 27, this immediately implies the error of both the OLS and the oracle estimator

$$e_{\text{odn}} \leq \frac{C'\kappa\sqrt{\mu^+}}{\mu^- \sqrt{m}} \cdot \frac{w(\mathcal{C})}{\sqrt{n}} \ , \tag{9.42}$$

$$e_{\text{orc}} \leq \frac{C'\kappa\sqrt{\mu^+}}{\mu^- \sqrt{\text{Tr}(\boldsymbol{\Sigma}_*^{-1})}} \cdot \frac{w(\mathcal{C})}{\sqrt{n}} \ , \tag{9.43}$$

which indicates that the oracle estimator improves the OLS by a factor of

$$\frac{e_{\text{orc}}}{e_{\text{odn}}} = \sqrt{\frac{m}{\text{Tr}(\boldsymbol{\Sigma}_*^{-1})}} \ . \tag{9.44}$$

In practice, this improvement can be significant, when there is strong cross-correlation among the responses, such that $\boldsymbol{\Sigma}_*$ is close to singular.

Assembling the results in Lemma 27 - 29 and Proposition 18, we have the following corollary for the error of algorithm 10.

Corollary 7 *Under the assumptions (A1) and (A2), if* $n \geq C_0 \cdot \max\left\{1, \tau^4, \kappa^4 \left(\frac{\mu^+ \sigma_*^+}{\mu^- \sigma_*^-}\right)^2,\right.$

$\kappa^2 \left(\frac{\mu^+}{\mu^-}\right)^2 \left(\frac{\sigma_*^+}{\sigma_*^-}\right) \Big\} \cdot \max\left\{\frac{w^4(\mathcal{C})}{m}, m\right\}$, *with probability at least* $1 - C_2 \exp(-C_1 m)$, *the minimum achievable error* $e_{\min}$ *of Algorithm 10 with arbitrary initialization satisfies*

$$e_{\min} \leq \frac{C_3 \kappa \sqrt{\mu^+}}{\mu^- \sqrt{\mathrm{Tr}(\boldsymbol{\Sigma}_*^{-1})}} \cdot \frac{m + w(\mathcal{C})}{\sqrt{n}} \cdot \frac{1 + \delta_n}{1 - \rho_n} \ , \tag{9.45}$$

where δ_n *and* ρ_n *satisfies*

$$\delta_n = C_4 \tau^2 \sqrt{\frac{m}{n}} \leq \frac{1}{4}$$
$$\rho_n \leq \frac{C_5 \kappa \mu^+}{\mu^- \sqrt{\sigma_*^- \ \mathrm{Tr}(\boldsymbol{\Sigma}_*^{-1})}} \cdot \frac{m + w(\mathcal{C})}{\sqrt{n}} \leq \frac{1}{2}$$

Remark: Though the initialization condition $f(\hat{\boldsymbol{\theta}}_{(0)}) \leq f(\boldsymbol{\theta}_*)$ may not be true for arbitrary $\hat{\boldsymbol{\theta}}_{(0)}$, it should be satisfied by the first iterate $\hat{\boldsymbol{\theta}}_{(1)}$, from which Theorem 28 starts to work with $e_0 = +\infty$. Hence the result holds for any initialization $\hat{\boldsymbol{\theta}}_{(0)}$. Following the analysis in Chapter 8, the resampled AltMin has an error bound that matches the oracle error e_{orc} up to a constant factor. Hence the price paid for this resampling-free result is only an additive $O\left(\frac{m}{\sqrt{n}}\right)$ term. It is also worth noting that the result in Chapter 8 needs a good initialization to hold, whereas this does not.

To illustrate the error bound above, we complement it with an example, in which the complexity function f is chosen to be L_1 norm.

Example with L_1 norm: For $f = \|\cdot\|_1$ and an s-sparse $\boldsymbol{\theta}_*$, the Gaussian width of the L_1 error spherical cap satisfies

$$w(\mathcal{C}) = O\left(\sqrt{s \log p}\right)$$

according to [40]. This gives the order of $e_{\min}$ as

$$e_{\min} = O\left(\frac{m + \sqrt{s \log p}}{\sqrt{n}}\right)$$

when $n = \Omega\left(\max\left\{\frac{s^2 \log^2 p}{m}, m\right\}\right)$.

9.4.3 Well-Initialized AltMin

For well-initialized AltMin, most of the analysis stays the same, with the exception being $\beta_n(\mathcal{M}(e_0))$. With a small value of e_0, the index set $\mathcal{M}(e_0)$ in the definition of $\beta_n(\mathcal{M}(e_0))$ will shrink, so that we are able to sharpen the upper bound of $\beta_n(\mathcal{M}(e_0))$. Before presenting the results, we introduce the set called error spherical sector.

Definition 30 (error spherical sector) For a complexity function f, its error spherical sector is defined as

$$\mathcal{S} = \text{cone}\left\{\mathbf{u} \in \mathbb{R}^p \;\middle|\; f(\boldsymbol{\theta}_* + \mathbf{u}) \leq f(\boldsymbol{\theta}_*)\right\} \;\cap\; \mathbb{B}^{p-1}\ , \tag{9.46}$$

where $\mathbb{B}^p = \{\mathbf{u} \mid \|\mathbf{u}\|_2 \leq 1\}$ is the unit ball of $\mathbb{R}^p$.

Geometrically $\mathcal{S}$ is closely related to the previously defined set $\mathcal{C}$ (Definition 25), for which $\mathcal{C} \subseteq \mathcal{S}$ and $\mathcal{S} \subseteq \text{conv}(\mathcal{C} \cup \{\mathbf{0}\})$ hold. More importantly, their Gaussian widths satisfy

$$w(\mathcal{S}) \leq w(\mathcal{C}) + c \tag{9.47}$$

for some constant c. The following lemma bounds the $\beta_n(\mathcal{M}(e_0))$ at $e_0 = \sqrt{\frac{\sigma_*^-}{\mu^+}}$ using $w(\mathcal{S})$.

Lemma 30 *Suppose that the conditions of Lemma 25 are satisfied with probability* $1 - \epsilon$ *when* $n \geq n_0$*. Under the assumptions (A1) and (A2), if* $n \geq \max\Big\{n_0, C_0 \cdot$

$\max\{\tau^4, \kappa^4, 1\}\cdot \max\{w^2(\mathcal{C}), \frac{m^3}{w^2(\mathcal{C})}, m^2\}\Big\}$, *the type-II NDI strength for* $\mathcal{M}(e_0)$ *with* $e_0 = \sqrt{\frac{\sigma_*^-}{\mu^+}}$ *satisfies*

$$\beta_n\left(\mathcal{M}(e_0)\right) \leq C_3 \cdot \frac{\kappa\sqrt{\mu^+}\cdot w(\mathcal{S})}{\sqrt{n}} \tag{9.48}$$

with probability at least $1 - \epsilon - C_2 \exp\left(-C_1 \cdot \min\left\{w^2(\mathcal{S}), m\right\}\right)$.

Together with the analysis presented in the previous subsection, the improved bound for $\beta_n(\mathcal{M}(e_0))$ immediately yields the error bound of Algorithm 10 under good initialization.

Corollary 8 *Under the assumptions (A1) and (A2), if* $n \geq C_0 \cdot \max\left\{1, \tau^4, \kappa^4\left(\frac{\mu^+\sigma_*^+}{\mu^-\sigma_*^-}\right)^2,\right.$ $\left.\kappa^2\left(\frac{\mu^+}{\mu^-}\right)^2\left(\frac{\sigma_*^+}{\sigma_*^-}\right)\right\} \cdot \max\left\{\frac{w^4(\mathcal{C})}{m}, \frac{m^3}{w^2(\mathcal{C})}, m^2\right\}$ *and the initialization* $\hat{\boldsymbol{\theta}}_{(0)}$ *satisfies* $f(\hat{\boldsymbol{\theta}}_{(0)}) \leq f(\boldsymbol{\theta}_*)$ *and* $\|\hat{\boldsymbol{\theta}}_{(0)} - \boldsymbol{\theta}_*\|_2 \leq \sqrt{\frac{\sigma_*^-}{\mu^+}}$, *with probability at least* $1 - C_2 \exp\left(-C_1 \cdot \min\left\{w^2(\mathcal{S}), m\right\}\right)$, *the minimum achievable error* $e_{\min}$ *of Algorithm 10 satisfies*

$$e_{\min} \leq \frac{C_3\kappa\sqrt{\mu^+}}{\mu^-\sqrt{\mathrm{Tr}(\boldsymbol{\Sigma}_*^{-1})}} \cdot \frac{w(\mathcal{S})}{\sqrt{n}} \cdot \frac{1+\delta_n}{1-\rho_n} \tag{9.49}$$

where δ_n *and* ρ_n *satisfies*

$$\delta_n = C_4\tau^2\sqrt{\frac{m}{n}} \leq \frac{1}{4}$$
$$\rho_n \leq \frac{C_5\kappa\mu^+}{\mu^-\sqrt{\sigma_*^-\ \mathrm{Tr}(\boldsymbol{\Sigma}_*^{-1})}} \cdot \frac{w(\mathcal{S})}{\sqrt{n}} \leq \frac{1}{2}$$

Remark: Since $w(\mathcal{S})$ only differs from $w(\mathcal{C})$ by a constant, the error bound (9.49) is sharper compared with Corollary 7, matching the order of the resampling-based result and the bound for e_{orc}. A good initialization of $\hat{\boldsymbol{\theta}}_{(0)}$ can be obtained by solving (9.30), whose error is guaranteed by (9.42). Therefore the initialization condition will hold as long as the sample size is sufficiently large. On the other hand, the iterates obtained by running randomly-initialized AltMin may also satisfy the initialization requirements as Corollary 7 guarantees a moderate error for any initialization. Once the requirements are met during the iteration, the randomly-initialized AltMin can attain this sharper bound as well as the well-initialized.

Example with L_0-cardinality: For $f = \|\cdot\|_0$ and an s-sparse $\boldsymbol{\theta}_*$, the set $\mathcal{S}$ satisfies

$$\mathcal{S} \subseteq \left\{\boldsymbol{\theta} \in \mathbb{S}^{p-1} \mid \|\boldsymbol{\theta}\|_0 \leq 2s\right\} ,$$

which by simple calculation implies that

$$w(\mathcal{S}) = O\left(\sqrt{s \log p}\right) .$$

Therefore it follows from Corollary 8 that

$$e_{\min} = O\left(\sqrt{\frac{s \log p}{n}}\right)$$

if $n = \Omega\left(\max\left\{\frac{m^3}{s \log p}, \frac{s^2 \log^2 p}{m}, m^2\right\}\right)$.

9.5 Experimental Results

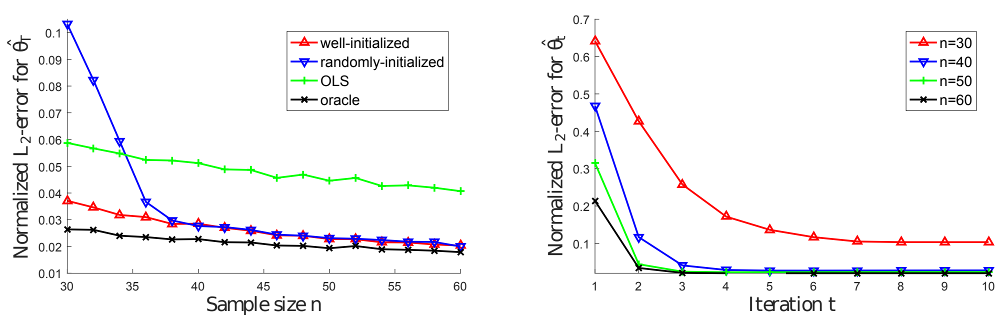
In this section, we present some experimental results to support our theoretical analysis. Specifically we focus on the sparsity structure of $\boldsymbol{\theta}_*$, and consider L_0-cardinality as complexity function f. Throughout the experiment, we fix problem dimension $p = 1000$, sparsity level of $\boldsymbol{\theta}_*$ $s = 20$, and number of iterations for AltMin $T = 10$. Entries of design $\mathbf{X}$ is generated by i.i.d. standard Gaussians, and $\boldsymbol{\theta}_* = [\underbrace{1, \ldots, 1}_{10}, \underbrace{-1, \ldots, -1}_{10}, \underbrace{0, \ldots, 0}_{980}]^T$.

$\mathbf{\Sigma}_*$ is given as

$$\mathbf{\Sigma}_* = \begin{bmatrix} \begin{matrix} 1 & a \\ a & 1 \end{matrix} & \mathbf{0}_{2\times 2} & \ldots & \mathbf{0}_{2\times 2} \\ \mathbf{0}_{2\times 2} & \begin{matrix} 1 & a \\ a & 1 \end{matrix} & \ldots & \mathbf{0}_{2\times 2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{2\times 2} & \mathbf{0}_{2\times 2} & \ldots & \begin{matrix} 1 & a \\ a & 1 \end{matrix} \end{bmatrix}.$$

The experimental results are obtained based on the average over 100 random trials.

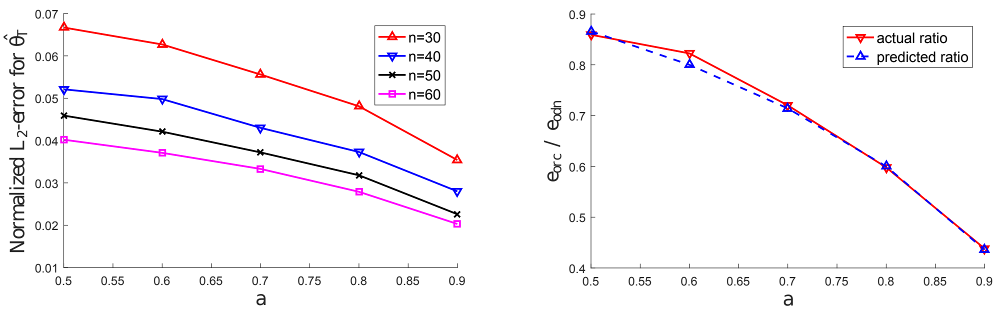
First we set $a = 0.9$, $m = 10$, and vary sample size n from 30 to 80. We run the AltMin initialized by both constrained ordinary least squares and Gaussian random vector, where $\boldsymbol{\theta}$-step is solved by the hard-thresholding pursuit (HTP) algorithm [56]. The error plots are shown in Figure 9.1.



(a) L_2-error vs. sample size n (b) L_2-error vs. iteration t

Figure 9.1: (a)A phase transition is observed for the randomly-initialized AltMin around $n = 40$, whose error is on a par with the well-initialized for $n \geq 40$. This coincides with the remark for Corollary 8. Also, the error of AltMin is close to the oracle estimator, which is significantly better than OLS. (b) Our theoretical results suggest that larger sample size leads to smaller ρ_n, thus making AltMin converge faster as shown in the plots.

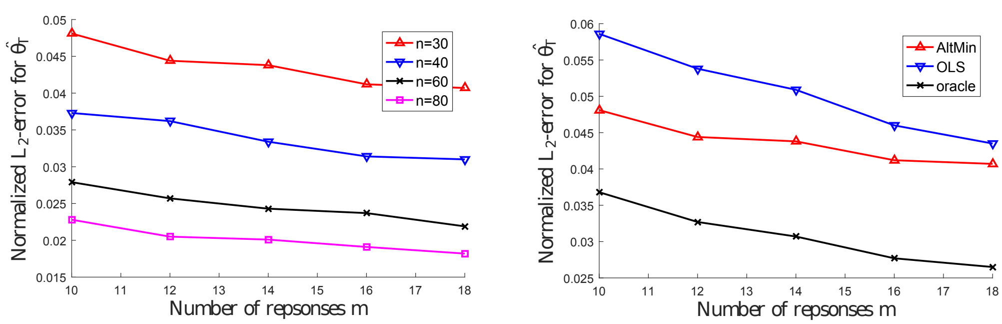
For the second set of experiments, we fix $m = 10$, and vary the parameter a in $\mathbf{\Sigma}_*$ from 0.5 to 0.9 for $n = 30$, 40, 50 and 60. The plots in Figure 9.2(a) shows the error of AltMin against a. As indicated by (9.29) and (9.31), the improvement of the oracle least squares over the ordinary one is amplified with increasingly large a. Figure 9.2(b) compares the actual ratio of e_{orc} to e_{odn} and the suggested one.



(a) L_2-error vs. a (b) $e_{\mathrm{orc}}/e_{\mathrm{odn}}$ for $n = 60$

Figure 9.2: (a) With a varying from 0.5 to 0.9, the responses become increasingly correlated and the error of AltMin reduces more quickly. (b) The actual ratio of e_{orc} to e_{odn} is very close the predicted one given by (9.44).

Finally we fix $a = 0.8$, and the number of responses m ranges from 10 to 18 for $n = 30$, 40, 50 and 60. The results are presented in Figure 9.3.



(a) L_2-error vs. number of responses m (b) L_2-error of all estimators for $n = 30$

Figure 9.3: (a) As m increases from 10 to 18, the error of AltMin does not decrease drastically. The main reason is the increasingly large error in the estimation of $\mathbf{\Sigma}_*$. (b) Compared with the error of OLS, the advantage of AltMin becomes marginal with growing m, while its gap with the oracle estimator is widened.

Appendix

Appendix 9.A Proofs for Deterministic Analysis

9.A.1 Proof of Lemma 25

Proof: We will use the shorthand notation $\hat{\mathbf{\Sigma}}$ for $\hat{\mathbf{\Sigma}}(\boldsymbol{\theta})$.

$$
\begin{aligned}
\frac{\xi(\hat{\mathbf{\Sigma}})}{\xi(\mathbf{\Sigma}_*)} &= \frac{\sqrt{\operatorname{Tr}\left(\hat{\mathbf{\Sigma}}^{-1}\mathbf{\Sigma}_*\hat{\mathbf{\Sigma}}^{-1}\right)}}{\xi(\mathbf{\Sigma}_*)\operatorname{Tr}\left(\hat{\mathbf{\Sigma}}^{-1}\right)} = \sqrt{\frac{\operatorname{Tr}\left(\mathbf{\Sigma}_*^{-1}\right)\cdot\operatorname{Tr}\left(\hat{\mathbf{\Sigma}}^{-1}\mathbf{\Sigma}_*\hat{\mathbf{\Sigma}}^{-1}\right)}{\operatorname{Tr}^2\left(\hat{\mathbf{\Sigma}}^{-1}\right)}} \\
&= \sqrt{\frac{\operatorname{Tr}\left(\hat{\mathbf{\Sigma}}^{\frac{1}{2}}\mathbf{\Sigma}_*^{-1}\hat{\mathbf{\Sigma}}^{\frac{1}{2}}\hat{\mathbf{\Sigma}}^{-1}\right)\cdot\operatorname{Tr}\left(\hat{\mathbf{\Sigma}}^{-\frac{1}{2}}\mathbf{\Sigma}_*\hat{\mathbf{\Sigma}}^{-\frac{1}{2}}\hat{\mathbf{\Sigma}}^{-1}\right)}{\operatorname{Tr}^2\left(\hat{\mathbf{\Sigma}}^{-1}\right)}} \\
&\le \sqrt{\frac{\lambda_{\max}\left(\hat{\mathbf{\Sigma}}^{\frac{1}{2}}\mathbf{\Sigma}_*^{-1}\hat{\mathbf{\Sigma}}^{\frac{1}{2}}\right)\operatorname{Tr}\left(\hat{\mathbf{\Sigma}}^{-1}\right)\cdot\lambda_{\max}\left(\hat{\mathbf{\Sigma}}^{-\frac{1}{2}}\mathbf{\Sigma}_*\hat{\mathbf{\Sigma}}^{-\frac{1}{2}}\right)\operatorname{Tr}\left(\hat{\mathbf{\Sigma}}^{-1}\right)}{\operatorname{Tr}^2\left(\hat{\mathbf{\Sigma}}^{-1}\right)}} \\
&= \sqrt{\lambda_{\max}\left(\hat{\mathbf{\Sigma}}^{\frac{1}{2}}\mathbf{\Sigma}_*^{-1}\hat{\mathbf{\Sigma}}^{\frac{1}{2}}\right)\lambda_{\max}\left(\hat{\mathbf{\Sigma}}^{-\frac{1}{2}}\mathbf{\Sigma}_*\hat{\mathbf{\Sigma}}^{-\frac{1}{2}}\right)} = \sqrt{\frac{\lambda_{\max}\left(\mathbf{\Sigma}_*^{-\frac{1}{2}}\hat{\mathbf{\Sigma}}\mathbf{\Sigma}_*^{-\frac{1}{2}}\right)}{\lambda_{\min}\left(\mathbf{\Sigma}_*^{-\frac{1}{2}}\hat{\mathbf{\Sigma}}\mathbf{\Sigma}_*^{-\frac{1}{2}}\right)}},
\end{aligned}
$$

where the inequality follows from Von Neumann's trace inequality. Now we try to bound $\lambda_{\max}\left(\mathbf{\Sigma}_*^{-\frac{1}{2}}\hat{\mathbf{\Sigma}}\mathbf{\Sigma}_*^{-\frac{1}{2}}\right)$ and $\lambda_{\min}\left(\mathbf{\Sigma}_*^{-\frac{1}{2}}\hat{\mathbf{\Sigma}}\mathbf{\Sigma}_*^{-\frac{1}{2}}\right)$ separately. Note that any $\boldsymbol{\theta}$ given by the solution of (9.14) satisfies that $\frac{\boldsymbol{\theta}-\boldsymbol{\theta}_*}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_*\|_2} \in \mathcal{C}$. By the expression for $\hat{\mathbf{\Sigma}}$ in (9.13), we have for $\lambda_{\max}\left(\mathbf{\Sigma}_*^{-\frac{1}{2}}\hat{\mathbf{\Sigma}}\mathbf{\Sigma}_*^{-\frac{1}{2}}\right)$,

$$
\begin{aligned}
\lambda_{\max}\left(\mathbf{\Sigma}_*^{-\frac{1}{2}}\hat{\mathbf{\Sigma}}\mathbf{\Sigma}_*^{-\frac{1}{2}}\right) &= 1 + \lambda_{\max}\left(\mathbf{\Sigma}_*^{-\frac{1}{2}}\hat{\mathbf{\Sigma}}\mathbf{\Sigma}_*^{-\frac{1}{2}} - \mathbf{I}\right) = 1 + \left\|\mathbf{\Sigma}_*^{-\frac{1}{2}}\hat{\mathbf{\Sigma}}\mathbf{\Sigma}_*^{-\frac{1}{2}} - \mathbf{I}\right\|_2 \\
&\le 1 + \left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{\boldsymbol{\eta}}_i\tilde{\boldsymbol{\eta}}_i^T - \mathbf{I}\right\|_2 + \left\|\frac{2}{n}\sum_{i=1}^{n}\mathbf{\Sigma}_*^{-\frac{1}{2}}\mathbf{X}_i(\boldsymbol{\theta}-\boldsymbol{\theta}_*)\tilde{\boldsymbol{\eta}}_i^T\right\|_2 \\
&\quad + \lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{\Sigma}_*^{-\frac{1}{2}}\mathbf{X}_i(\boldsymbol{\theta}-\boldsymbol{\theta}_*)(\boldsymbol{\theta}-\boldsymbol{\theta}_*)^T\mathbf{X}_i^T\mathbf{\Sigma}_*^{-\frac{1}{2}}\right)
\end{aligned}
$$

$$= 1 + \delta_n + \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2 \cdot \left\| \frac{2}{n} \sum_{i=1}^{n} \boldsymbol{\Sigma}_*^{-\frac{1}{2}} \mathbf{X}_i \cdot \frac{\boldsymbol{\theta} - \boldsymbol{\theta}_*}{\|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2} \cdot \tilde{\boldsymbol{\eta}}_i^T \right\|_2$$
$$+ \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2^2 \cdot \sup_{\mathbf{v} \in \mathbb{S}^{m-1}} \mathbf{v}^T \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\Sigma}_*^{-\frac{1}{2}} \mathbf{X}_i \cdot \frac{(\boldsymbol{\theta} - \boldsymbol{\theta}_*)(\boldsymbol{\theta} - \boldsymbol{\theta}_*)^T}{\|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2^2} \cdot \mathbf{X}_i^T \boldsymbol{\Sigma}_*^{-\frac{1}{2}} \right) \mathbf{v}$$
$$\leq 1 + \delta_n + \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2 \cdot \left\| \boldsymbol{\Sigma}_*^{-\frac{1}{2}} \right\|_2 \cdot \sup_{\mathbf{u} \in \mathcal{C}} \left\| \frac{2}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{u} \tilde{\boldsymbol{\eta}}_i^T \right\|_2$$
$$+ \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2^2 \cdot \left\| \boldsymbol{\Sigma}_*^{-1} \right\|_2 \cdot \sup_{\mathbf{v} \in \mathbb{S}^{m-1}} \sup_{\mathbf{u} \in \mathcal{C}} \mathbf{u}^T \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{v} \mathbf{v}^T \mathbf{X}_i \right) \mathbf{u}$$
$$= 1 + \delta_n + \frac{\gamma_n}{\sqrt{\sigma_*^-}} \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2 + \frac{\alpha_n^+}{\sigma_*^-} \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2^2$$

Similarly we bound $\lambda_{\min}\left(\boldsymbol{\Sigma}_*^{-\frac{1}{2}} \hat{\boldsymbol{\Sigma}} \boldsymbol{\Sigma}_*^{-\frac{1}{2}}\right)$ as follows,

$$\lambda_{\min}\left(\boldsymbol{\Sigma}_*^{-\frac{1}{2}} \hat{\boldsymbol{\Sigma}} \boldsymbol{\Sigma}_*^{-\frac{1}{2}}\right) = 1 + \lambda_{\min}\left(\boldsymbol{\Sigma}_*^{-\frac{1}{2}} \hat{\boldsymbol{\Sigma}} \boldsymbol{\Sigma}_*^{-\frac{1}{2}} - \mathbf{I}\right)$$
$$\geq 1 + \lambda_{\min}\left(\frac{1}{n} \sum_{i=1}^{n} \tilde{\boldsymbol{\eta}}_i \tilde{\boldsymbol{\eta}}_i^T - \mathbf{I} \right)$$
$$+ \lambda_{\min}\left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\Sigma}_*^{-\frac{1}{2}} \mathbf{X}_i (\boldsymbol{\theta} - \boldsymbol{\theta}_*) \tilde{\boldsymbol{\eta}}_i^T + \frac{1}{n} \sum_{i=1}^{n} \tilde{\boldsymbol{\eta}}_i (\boldsymbol{\theta} - \boldsymbol{\theta}_*)^T \mathbf{X}_i^T \boldsymbol{\Sigma}_*^{-\frac{1}{2}} \right)$$
$$+ \lambda_{\min}\left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\Sigma}_*^{-\frac{1}{2}} \mathbf{X}_i (\boldsymbol{\theta} - \boldsymbol{\theta}_*)(\boldsymbol{\theta} - \boldsymbol{\theta}_*)^T \mathbf{X}_i^T \boldsymbol{\Sigma}_*^{-\frac{1}{2}} \right)$$
$$\geq 1 - \left\| \frac{1}{n} \sum_{i=1}^{n} \tilde{\boldsymbol{\eta}}_i \tilde{\boldsymbol{\eta}}_i^T - \mathbf{I} \right\|_2 - \left\| \frac{2}{n} \sum_{i=1}^{n} \boldsymbol{\Sigma}_*^{-\frac{1}{2}} \mathbf{X}_i (\boldsymbol{\theta} - \boldsymbol{\theta}_*) \tilde{\boldsymbol{\eta}}_i^T \right\|_2$$
$$+ \lambda_{\min}\left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\Sigma}_*^{-\frac{1}{2}} \mathbf{X}_i (\boldsymbol{\theta} - \boldsymbol{\theta}_*)(\boldsymbol{\theta} - \boldsymbol{\theta}_*)^T \mathbf{X}_i^T \boldsymbol{\Sigma}_*^{-\frac{1}{2}} \right)$$
$$\geq 1 - \delta_n - \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2 \cdot \left\| \boldsymbol{\Sigma}_*^{-\frac{1}{2}} \right\|_2 \cdot \sup_{\mathbf{u} \in \mathcal{C}} \left\| \frac{2}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{u} \tilde{\boldsymbol{\eta}}_i^T \right\|_2$$
$$+ \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2^2 \cdot \lambda_{\min}(\boldsymbol{\Sigma}_*^{-1}) \inf_{\mathbf{v} \in \mathbb{S}^{m-1}} \inf_{\mathbf{u} \in \mathcal{C}} \mathbf{u} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{v} \mathbf{v}^T \mathbf{X}_i \right) \mathbf{u}$$
$$= 1 - \delta_n - \frac{\gamma_n}{\sqrt{\sigma_*^-}} \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2 + \frac{\alpha_n^-}{\sigma_*^+} \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2^2$$

Combining the inequalities above, we obtain

$$
\begin{aligned}
&\frac{\xi(\hat{\boldsymbol{\Sigma}})}{\xi(\boldsymbol{\Sigma}_*)} \\
&\le \sqrt{\frac{1+\delta_n+\frac{\gamma_n}{\sqrt{\sigma_*^-}}\|\boldsymbol{\theta}-\boldsymbol{\theta}_*\|_2+\frac{\alpha_n^+}{\sigma_*^-}\|\boldsymbol{\theta}-\boldsymbol{\theta}_*\|_2^2}{1-\delta_n-\frac{\gamma_n}{\sqrt{\sigma_*^-}}\|\boldsymbol{\theta}-\boldsymbol{\theta}_*\|_2+\frac{\alpha_n^-}{\sigma_*^+}\|\boldsymbol{\theta}-\boldsymbol{\theta}_*\|_2^2}} \\
&\le \sqrt{\frac{1+2\delta_n+\frac{\gamma_n^2}{4\sigma_*^-\delta_n}\|\boldsymbol{\theta}-\boldsymbol{\theta}_*\|_2^2+\frac{\alpha_n^+}{\sigma_*^-}\|\boldsymbol{\theta}-\boldsymbol{\theta}_*\|_2^2}{1-2\delta_n-\frac{\gamma_n^2}{4\sigma_*^-\delta_n}\|\boldsymbol{\theta}-\boldsymbol{\theta}_*\|_2^2+\frac{\alpha_n^-}{\sigma_*^+}\|\boldsymbol{\theta}-\boldsymbol{\theta}_*\|_2^2}} \qquad \left(\text{use } \sqrt{ab} \le \frac{a+b}{2} \text{ for } a,b \ge 0\right) \\
&\le \sqrt{\frac{1+2\delta_n+\frac{2\alpha_n^+}{\sigma_*^-}\|\boldsymbol{\theta}-\boldsymbol{\theta}_*\|_2^2}{1-2\delta_n}} \qquad \left(\text{use the condition } \frac{\delta_n\alpha_n^-}{\gamma_n^2} \ge \frac{\sigma_*^+}{4\sigma_*^-}\right) \\
&\le \sqrt{\frac{1+2\delta_n}{1-2\delta_n}}+\sqrt{\frac{2\alpha_n^+\|\boldsymbol{\theta}-\boldsymbol{\theta}_*\|_2^2}{(1-2\delta_n)\sigma_*^-}} \qquad \left(\text{follow from } \sqrt{a+b} \le \sqrt{a}+\sqrt{b} \text{ for } a,b \ge 0\right) \\
&\le 1+\frac{2\delta_n}{1-2\delta_n}+\sqrt{\frac{2\alpha_n^+\|\boldsymbol{\theta}-\boldsymbol{\theta}_*\|_2^2}{(1-2\delta_n)\sigma_*^-}} \qquad \left(\text{follow from } \sqrt{1+a} \le 1+\frac{a}{2} \text{ for } a \ge 0\right) \\
&\le 1+4\delta_n+2\sqrt{\frac{\alpha_n^+}{\sigma_*^-}}\cdot\|\boldsymbol{\theta}-\boldsymbol{\theta}_*\|_2 \qquad \left(\text{use the condition } \delta_n \le \frac{1}{4}\right).
\end{aligned}
$$

The invertibility of $\hat{\boldsymbol{\Sigma}}$ is guaranteed by $\lambda_{\min}(\boldsymbol{\Sigma}_*^{-\frac{1}{2}}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}_*^{-\frac{1}{2}}) > \frac{1}{2}$ following from the derivation above. ∎

9.A.2 Proof of Lemma 26

Proof: We use the shorthand notation $\hat{\boldsymbol{\theta}}$ for $\hat{\boldsymbol{\theta}}(\boldsymbol{\Sigma})$. Since the tuning parameter λ is set to $\|\boldsymbol{\theta}_*\|$, the optimality of $\hat{\boldsymbol{\theta}}$ implies that

$$
\begin{aligned}
&\frac{1}{2n}\sum_{i=1}^{n}\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}_i-\mathbf{X}_i\hat{\boldsymbol{\theta}})\right\|_2^2 \le \frac{1}{2n}\sum_{i=1}^{n}\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\theta}_*)\right\|_2^2 \\
\Longrightarrow \quad &\frac{1}{2n}\sum_{i=1}^{n}\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\theta}_*)+\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{X}_i(\boldsymbol{\theta}_*-\hat{\boldsymbol{\theta}})\right\|_2^2 \le \frac{1}{2n}\sum_{i=1}^{n}\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\theta}_*)\right\|_2^2
\end{aligned}
$$

$$\implies \quad \frac{1}{2n}\sum_{i=1}^{n}\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{X}_i(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_*)\right\|_2^2+\frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\theta}_*)^T\boldsymbol{\Sigma}^{-1}\mathbf{X}_i(\boldsymbol{\theta}_*-\hat{\boldsymbol{\theta}})\leq 0$$

$$\implies \quad \frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{X}_i(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_*)\right\|_2^2\leq \frac{2}{n}\sum_{i=1}^{n}\boldsymbol{\eta}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{X}_i(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_*)$$

$$\implies \quad \left\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_*\right\|_2\leq \frac{\frac{2}{n}\sum_{i=1}^{n}\boldsymbol{\eta}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\cdot\frac{\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_*}{\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_*\|_2}}{\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{X}_i\cdot\frac{\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_*}{\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_*\|_2}\right\|_2^2}$$

Now we try to bound the numerator and the denominator on the right-hand side. Note that $f(\hat{\boldsymbol{\theta}})\leq\lambda=f(\boldsymbol{\theta}_*)$, we thus have $\frac{\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_*}{\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_*\|_2}\in\mathcal{C}$ according to the definition of the error spherical cap. Assuming the eigenvalue decomposition $\boldsymbol{\Sigma}=\sum_{j=1}^{m}\sigma_j\mathbf{v}_j\mathbf{v}_j^T$, we further get

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{X}_i\cdot\frac{\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_*}{\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_*\|_2}\right\|_2^2 &\geq \inf_{\mathbf{u}\in\mathcal{C}}\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{X}_i\mathbf{u}\right\|_2^2\\
&=\inf_{\mathbf{u}\in\mathcal{C}}\frac{1}{n}\sum_{i=1}^{n}\mathbf{u}^T\mathbf{X}_i^T\left(\sum_{j=1}^{m}\sigma_j^{-1}\mathbf{v}_j\mathbf{v}_j^T\right)\mathbf{X}_i\mathbf{u}\\
&=\inf_{\mathbf{u}\in\mathcal{C}}\sum_{j=1}^{m}\sigma_j^{-1}\cdot\mathbf{u}^T\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{v}_j\mathbf{v}_j^T\mathbf{X}_i\right)\mathbf{u}\\
&\geq\left(\sum_{j=1}^{m}\sigma_j^{-1}\right)\cdot\inf_{\mathbf{v}\in\mathbb{S}^{m-1}}\inf_{\mathbf{u}\in\mathcal{C}}\mathbf{u}^T\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{v}\mathbf{v}^T\mathbf{X}_i\right)\mathbf{u}\\
&=\alpha_n^-\cdot\mathrm{Tr}(\boldsymbol{\Sigma}^{-1})
\end{aligned}$$

$$\begin{aligned}
\frac{2}{n}\sum_{i=1}^{n}\boldsymbol{\eta}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\cdot\frac{\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_*}{\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_*\|_2} &\leq\sup_{\mathbf{u}\in\mathcal{C}}\frac{2}{n}\sum_{i=1}^{n}\boldsymbol{\eta}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\mathbf{u}\\
&=\left\|\boldsymbol{\Sigma}_*^{1/2}\boldsymbol{\Sigma}^{-1}\right\|_F\cdot\sup_{\mathbf{u}\in\mathcal{C}}\frac{2}{n}\sum_{i=1}^{n}\frac{\boldsymbol{\eta}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\mathbf{u}}{\|\boldsymbol{\Sigma}_*^{1/2}\boldsymbol{\Sigma}^{-1}\|_F}\\
&\leq\left\|\boldsymbol{\Sigma}_*^{1/2}\boldsymbol{\Sigma}^{-1}\right\|_F\cdot\sup_{\boldsymbol{\Sigma}\in\mathcal{M}}\sup_{\mathbf{u}\in\mathcal{C}}\ \frac{2}{n}\sum_{i=1}^{n}\frac{\boldsymbol{\eta}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\mathbf{u}}{\|\boldsymbol{\Sigma}_*^{1/2}\boldsymbol{\Sigma}^{-1}\|_F}
\end{aligned}$$

$$= \beta_n \cdot \sqrt{\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_*\boldsymbol{\Sigma}^{-1})}$$

Combining the results above, we can get (9.28). ■

9.A.3 Proof of Theorem 28

Proof: Since the initialization $\hat{\boldsymbol{\theta}}_{(0)}$ satisfies $f(\hat{\boldsymbol{\theta}}_{(0)}) \leq f(\boldsymbol{\theta}_*)$ and $\|\hat{\boldsymbol{\theta}}_{(0)} - \boldsymbol{\theta}_*\|_2 \leq e_0$, we have $\hat{\boldsymbol{\Sigma}}_{(1)} \in \mathcal{M}(e_0)$ by Lemma 25 and 26, we have for the first iteration of Algorithm 10,

$$\begin{aligned}
d_1\left(\hat{\boldsymbol{\Sigma}}_{(1)},\ \boldsymbol{\Sigma}_*\right) &\leq 4\delta_n + 2\sqrt{\frac{\alpha_n^+}{\sigma_*^-}} \cdot d_2\left(\hat{\boldsymbol{\theta}}_{(0)},\ \boldsymbol{\theta}_*\right) \\
d_2\left(\hat{\boldsymbol{\theta}}_{(1)},\ \boldsymbol{\theta}_*\right) &\leq \xi(\hat{\boldsymbol{\Sigma}}_{(1)}) \cdot \frac{\beta_n(\mathcal{M}(e_0))}{\alpha_n^-} = \varepsilon_n \cdot \left(1 + d_1\left(\hat{\boldsymbol{\Sigma}}_{(1)},\ \boldsymbol{\Sigma}_*\right)\right)
\end{aligned}$$

Combining the two inequalities, we obtain the recurrence relation for the error of $\hat{\boldsymbol{\theta}}_{(1)}$ and $\hat{\boldsymbol{\theta}}_{(0)}$,

$$d_2\left(\hat{\boldsymbol{\theta}}_{(1)},\ \boldsymbol{\theta}_*\right) \leq \varepsilon_n \cdot \left(1 + 4\delta_n + 2\sqrt{\frac{\alpha_n^+}{\sigma_*^-}} \cdot d_2\left(\hat{\boldsymbol{\theta}}_{(0)},\ \boldsymbol{\theta}_*\right)\right)$$

As $\rho_n < 1$ and $e_{\min} \leq e_0$, we have $d_2(\hat{\boldsymbol{\theta}}_{(1)}, \boldsymbol{\theta}_*) \leq e_0$, thus $\hat{\boldsymbol{\Sigma}}_{(2)} \in \mathcal{M}(e_0)$. By induction, we can recursively apply the result to $t = 2, 3, \ldots, T$,

$$d_2\left(\hat{\boldsymbol{\theta}}_{(T)},\ \boldsymbol{\theta}_*\right) \leq q_T, \quad \text{where } q_t = \varepsilon_n\left(1 + 4\delta_n\right) + 2\varepsilon_n\sqrt{\frac{\alpha_n^+}{\sigma_*^-}} \cdot q_{t-1} \ \text{ and } \ q_0 \leq e_0$$

Solving the recurrence of r_t, we get

$$\begin{aligned} q_T &= \frac{\varepsilon_n\left(1+4\delta_n\right)}{1-2\varepsilon_n\sqrt{\frac{\alpha_n^+}{\sigma_*^-}}} + \left(2\varepsilon_n\sqrt{\frac{\alpha_n^+}{\sigma_*^-}}\right)^T \cdot \left(q_0 - \frac{\varepsilon_n\left(1+4\delta_n\right)}{1-2\varepsilon_n\sqrt{\frac{\alpha_n^+}{\sigma_*^-}}}\right) \\ &= e_{\min} + \rho_n^T \cdot \left(q_0 - e_{\min}\right) \\ &\leq e_{\min} + \rho_n^T \cdot \left(e_0 - e_{\min}\right) \ , \end{aligned}$$

which completes the proof. ■

Appendix 9.B Proofs for Probabilistic Analysis

9.B.1 Proof of Proposition 17

Proof: Since $\alpha_n^- = \Theta(1)$ and $\beta_n(\mathcal{M}(e_0)) = o(1)$, we have $\varepsilon_n = o(1)$. As (ii) holds, it follows from that $\delta_n \leq \frac{1}{4}$ when n is large. Due to (iii), the condition $\frac{\delta_n \alpha_n^-}{\gamma_n^2} \geq \frac{\sigma_*^+}{4\sigma_*^-}$ is true for sufficiently large n. Given that $\varepsilon_n = o(1)$ and $\alpha_n^+ = \Theta(1)$, we have $\rho_n = o(1)$. With $\delta_n = o(1)$ and $\rho_n = o(1)$, it is easy to see that $e_{\min} \leq e_0$ for large enough n. ■

9.B.2 Proof of Lemma 27

Proof: First we have

$$\begin{aligned} \alpha_n^- &= \inf_{\mathbf{v}\in\mathbb{S}^{m-1}} \inf_{\mathbf{u}\in\mathcal{C}} \mathbf{u}^T \left(\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i^T \mathbf{v}\mathbf{v}^T \mathbf{X}_i\right)\mathbf{u} \\ &\geq \inf_{\mathbf{v}\in\mathbb{S}^{m-1}} \inf_{\mathbf{u}\in\mathcal{C}} \mathbf{u}^T \left(\mathbb{E}\left[\mathbf{X}^T\mathbf{v}\mathbf{v}^T\mathbf{X}\right]\right)\mathbf{u} \\ &\quad + \inf_{\mathbf{v}\in\mathbb{S}^{m-1}} \inf_{\mathbf{u}\in\mathcal{C}} \mathbf{u}^T \left(\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i^T \mathbf{v}\mathbf{v}^T \mathbf{X}_i - \mathbb{E}\left[\mathbf{X}^T\mathbf{v}\mathbf{v}^T\mathbf{X}\right]\right)\mathbf{u} \end{aligned}$$

$$
\begin{aligned}
&\geq \inf_{\mathbf{v}\in\mathbb{S}^{m-1}} \inf_{\mathbf{u}\in\mathcal{C}} \mathbf{u}^T \left(\mathbb{E}\left[\mathbf{X}^T\mathbf{v}\mathbf{v}^T\mathbf{X}\right]\right)\mathbf{u} - \sup_{\mathbf{v}\in\mathbb{S}^{m-1}} \sup_{\mathbf{u}\in\mathcal{C}} \left|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{u}^T\mathbf{X}_i^T\mathbf{v})^2 - \mathbb{E}(\mathbf{u}^T\mathbf{X}^T\mathbf{v})^2\right| \\
&\geq \mu^- - \sup_{\mathbf{v}\in\mathbb{S}^{m-1}} \sup_{\mathbf{u}\in\mathcal{C}} \left|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{u}^T\mathbf{X}_i^T\mathbf{v})^2 - \mathbb{E}(\mathbf{u}^T\mathbf{X}^T\mathbf{v})^2\right| \\
\alpha_n^+ &= \sup_{\mathbf{v}\in\mathbb{S}^{m-1}} \sup_{\mathbf{u}\in\mathcal{C}} \mathbf{u}^T \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{v}\mathbf{v}^T\mathbf{X}_i\right)\mathbf{u} \\
&\leq \sup_{\mathbf{v}\in\mathbb{S}^{m-1}} \sup_{\mathbf{u}\in\mathcal{C}} \mathbf{u}^T \left(\mathbb{E}\left[\mathbf{X}^T\mathbf{v}\mathbf{v}^T\mathbf{X}\right]\right)\mathbf{u} \\
&\quad + \sup_{\mathbf{v}\in\mathbb{S}^{m-1}} \sup_{\mathbf{u}\in\mathcal{C}} \mathbf{u}^T \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{v}\mathbf{v}^T\mathbf{X}_i - \mathbb{E}\left[\mathbf{X}^T\mathbf{v}\mathbf{v}^T\mathbf{X}\right]\right)\mathbf{u} \\
&\leq \sup_{\mathbf{v}\in\mathbb{S}^{m-1}} \sup_{\mathbf{u}\in\mathcal{C}} \mathbf{u}^T \left(\mathbb{E}\left[\mathbf{X}^T\mathbf{v}\mathbf{v}^T\mathbf{X}\right]\right)\mathbf{u} + \sup_{\mathbf{v}\in\mathbb{S}^{m-1}} \sup_{\mathbf{u}\in\mathcal{C}} \left|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{u}^T\mathbf{X}_i^T\mathbf{v})^2 - \mathbb{E}(\mathbf{u}^T\mathbf{X}^T\mathbf{v})^2\right| \\
&\leq \mu^+ + \sup_{\mathbf{v}\in\mathbb{S}^{m-1}} \sup_{\mathbf{u}\in\mathcal{C}} \left|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{u}^T\mathbf{X}_i^T\mathbf{v})^2 - \mathbb{E}(\mathbf{u}^T\mathbf{X}^T\mathbf{v})^2\right|
\end{aligned}
$$

Now the goal is to bound $\sup_{\mathbf{v}\in\mathbb{S}^{m-1}} \sup_{\mathbf{u}\in\mathcal{C}} \left|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{u}^T\mathbf{X}_i^T\mathbf{v})^2 - \mathbb{E}(\mathbf{u}^T\mathbf{X}^T\mathbf{v})^2\right|$. In order to apply Corollary 1, we let $\mathcal{A} = \mathbb{S}^{m-1}\times\mathcal{C} \subset \mathbb{R}^{m+p}$, $\mathbf{a} = (\mathbf{v}, \mathbf{u})$, and the function class $\mathcal{F} = \{f_{\mathbf{a}} = \mathbf{u}^T\mathbf{X}^T\mathbf{v}\}_{\mathbf{a}\in\mathcal{A}}$. We then verify the conditions required by Corollary 1 for $\mathcal{F}$ and $\mathcal{A}$.

$$
\begin{aligned}
\sup_{f\in\mathcal{F}} |||f|||_{\psi_2} &= \sup_{\mathbf{v}\in\mathbb{S}^{m-1}} \sup_{\mathbf{u}\in\mathcal{C}} \left|\left|\left|\mathbf{u}^T\mathbf{X}^T\mathbf{v}\right|\right|\right|_{\psi_2} \\
&= \sup_{\mathbf{v}\in\mathbb{S}^{m-1}} \sup_{\mathbf{u}\in\mathcal{C}} \left|\left|\left|\mathbf{u}^T\Gamma_{\mathbf{v}}^{1/2}\Gamma_{\mathbf{v}}^{-1/2}\mathbf{X}^T\mathbf{v}\right|\right|\right|_{\psi_2} \\
&\leq \kappa\cdot \sup_{\mathbf{v}\in\mathbb{S}^{m-1}} \sup_{\mathbf{u}\in\mathcal{C}} \left\|\Gamma_{\mathbf{v}}^{1/2}\mathbf{u}\right\|_2 \\
&\leq \kappa\cdot \sup_{\mathbf{v}\in\mathbb{S}^{m-1}} \left\|\Gamma_{\mathbf{v}}^{1/2}\right\|_2 \leq \kappa\sqrt{\mu^+} \qquad \Longrightarrow \qquad R_{\mathcal{F}} = \kappa\sqrt{\mu^+}
\end{aligned}
$$

$$
\begin{aligned}
\forall\ \mathbf{a}, \mathbf{a}' \in \mathcal{A}, \qquad & \left\lvert\!\left\lvert\!\left\lvert f_{\mathbf{a}} - f_{\mathbf{a}'} \right\rvert\!\right\rvert\!\right\rvert_{\psi_2} \\
&= \left\lvert\!\left\lvert\!\left\lvert \mathbf{u}^T \mathbf{X}^T \mathbf{v} - \mathbf{u}'^T \mathbf{X}^T \mathbf{v}' \right\rvert\!\right\rvert\!\right\rvert_{\psi_2} \\
&= \left\lvert\!\left\lvert\!\left\lvert (\mathbf{u} - \mathbf{u}')^T \mathbf{X}^T \mathbf{v} + \mathbf{u}'^T \mathbf{X}^T (\mathbf{v} - \mathbf{v}') \right\rvert\!\right\rvert\!\right\rvert_{\psi_2} \\
&\le \left\| \mathbf{u} - \mathbf{u}' \right\|_2 \left\lvert\!\left\lvert\!\left\lvert \frac{(\mathbf{u} - \mathbf{u}')^T}{\|\mathbf{u} - \mathbf{u}'\|_2} \mathbf{X}^T \mathbf{v} \right\rvert\!\right\rvert\!\right\rvert_{\psi_2} + \left\| \mathbf{v} - \mathbf{v}' \right\|_2 \left\lvert\!\left\lvert\!\left\lvert \mathbf{u}'^T \mathbf{X}^T \frac{(\mathbf{v} - \mathbf{v}')}{\|\mathbf{v} - \mathbf{v}'\|_2} \right\rvert\!\right\rvert\!\right\rvert_{\psi_2} \\
&\le \kappa \sqrt{\mu^+} \left(\left\| \mathbf{u} - \mathbf{u}' \right\|_2 + \left\| \mathbf{v} - \mathbf{v}' \right\|_2 \right) \\
&\le \sqrt{2} \kappa \sqrt{\mu^+} \cdot \sqrt{\|\mathbf{u} - \mathbf{u}'\|_2^2 + \|\mathbf{v} - \mathbf{v}'\|_2^2} \\
&= \sqrt{2} \kappa \sqrt{\mu^+} \left\| \mathbf{a} - \mathbf{a}' \right\|_2 \qquad \Longrightarrow \qquad K_{\mathcal{F}} = \sqrt{2} \kappa \sqrt{\mu^+}
\end{aligned}
$$

It follows from Corollary 1 that if $n \ge c_0 w^2(\mathcal{A})$, the following result holds with probability at least $1 - 2\exp\left(-c_1 w^2(\mathcal{A})\right)$,

$$
\sup_{\mathbf{v} \in \mathbb{S}^{m-1}} \sup_{\mathbf{u} \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} (\mathbf{u}^T \mathbf{X}_i^T \mathbf{v})^2 - \mathbb{E}(\mathbf{u}^T \mathbf{X}^T \mathbf{v})^2 \right| \ \le\ c_2 \cdot \frac{\kappa^2 \mu^+ \cdot w(\mathcal{A})}{\sqrt{n}} \tag{9.50}
$$

If n further satisfies $n \ge 4 c_2^2 \kappa^4 \left(\frac{\mu^+}{\mu^-} \right)^2 w^2(\mathcal{A})$, then

$$
\begin{aligned}
&\sup_{\mathbf{v} \in \mathbb{S}^{m-1}} \sup_{\mathbf{u} \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} (\mathbf{u}^T \mathbf{X}_i^T \mathbf{v})^2 - \mathbb{E}(\mathbf{u}^T \mathbf{X}^T \mathbf{v})^2 \right| \ \le\ \frac{1}{2} \mu^- \\
\Longrightarrow \quad & \alpha_n^- \ge \mu^- - \frac{1}{2}\mu^- = \frac{1}{2}\mu^-, \quad \alpha_n^+ \le \mu^+ + \frac{1}{2}\mu^- \le \frac{3}{2}\mu^+
\end{aligned}
$$

Finally we note that

$$
\begin{aligned}
w(\mathcal{A}) &= \mathbb{E}\left[\sup_{\mathbf{a} \in \mathcal{A}} \langle \mathbf{a}, \mathbf{g}_{m+p} \rangle \right] = \mathbb{E}\left[\sup_{\mathbf{u} \in \mathbb{S}^{m-1}} \langle \mathbf{u}, \mathbf{g}_m \rangle + \sup_{\mathbf{v} \in \mathcal{C}} \langle \mathbf{v}, \mathbf{g}_p \rangle \right] \\
&= \mathbb{E}\left[\|\mathbf{g}_m\|_2 \right] + w(\mathcal{C}) = \Theta\left(\sqrt{m} \right) + w(\mathcal{C})
\end{aligned}
$$

By renaming the constants, we finish the proof. ■

9.B.3 Proof of Lemma 28

Proof: First we have

$$\gamma_n = \sup_{\mathbf{u}\in\mathcal{C}} \left\| \frac{2}{n}\sum_{i=1}^{n} \mathbf{X}_i \mathbf{u} \tilde{\boldsymbol{\eta}}_i^T \right\|_2 = 2 \sup_{\mathbf{u}\in\mathcal{C}} \sup_{\mathbf{v}\in\mathbb{S}^{m-1}} \sup_{\mathbf{b}\in\mathbb{S}^{m-1}} \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{v}^T\mathbf{X}_i\mathbf{u}\right)\left(\tilde{\boldsymbol{\eta}}_i^T\mathbf{b}\right)$$
$$= 2 \sup_{\mathbf{u}\in\mathcal{C}} \sup_{\mathbf{v}\in\mathbb{S}^{m-1}} \sup_{\mathbf{b}\in\mathbb{S}^{m-1}} \left| \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{v}^T\mathbf{X}_i\mathbf{u}\right)\left(\tilde{\boldsymbol{\eta}}_i^T\mathbf{b}\right) - \mathbb{E}\left[\mathbf{v}^T\mathbf{X}\mathbf{u}\tilde{\boldsymbol{\eta}}^T\mathbf{b}\right] \right|$$

Next we use Theorem 2 to bound the stochastic process above. Let $\mathcal{A} = \mathbb{S}^{m-1} \times \mathcal{C} \subset \mathbb{R}^{m+p}$, $\mathbf{a} = (\mathbf{v}, \mathbf{u})$ and $\mathcal{B} = \mathbb{S}^{m-1}$. Construct $\mathcal{F} = \{f_{\mathbf{a}} = \mathbf{v}^T\mathbf{X}\mathbf{u}\}_{\mathbf{a}\in\mathcal{A}}$ and $\mathcal{H} = \{h_{\mathbf{b}} = \tilde{\boldsymbol{\eta}}^T\mathbf{b}\}_{\mathbf{b}\in\mathcal{B}}$. We start by verifying the assumptions. Note that

$$\begin{aligned}
\sup_{f\in\mathcal{F}} |||f|||_{\psi_2} &= \sup_{\mathbf{u}\in\mathcal{C}} \sup_{\mathbf{v}\in\mathbb{S}^{m-1}} \left|\left|\left| \mathbf{u}^T\mathbf{X}^T\mathbf{v} \right|\right|\right|_{\psi_2} \\
&\le \sup_{\mathbf{u}\in\mathcal{C}} \sup_{\mathbf{v}\in\mathbb{S}^{m-1}} \left|\left|\left| \mathbf{u}^T\mathbf{X}^T\mathbf{v} \right|\right|\right|_{\psi_2} \\
&= \sup_{\mathbf{u}\in\mathcal{C}} \sup_{\mathbf{v}\in\mathbb{S}^{m-1}} \left|\left|\left| \mathbf{u}^T\boldsymbol{\Gamma}_{\mathbf{v}}^{1/2}\boldsymbol{\Gamma}_{\mathbf{v}}^{-1/2}\mathbf{X}^T\mathbf{v} \right|\right|\right|_{\psi_2} \\
&\le \sup_{\mathbf{u}\in\mathcal{C}} \sup_{\mathbf{v}\in\mathbb{S}^{m-1}} \kappa \left\| \boldsymbol{\Gamma}_{\mathbf{v}}^{1/2}\mathbf{u} \right\|_2 \\
&\le \kappa\sqrt{\mu^+} \qquad \Longrightarrow \qquad R_{\mathcal{F}} = \kappa\sqrt{\mu^+} \\
\sup_{h\in\mathcal{H}} |||h|||_{\psi_2} &= \sup_{\mathbf{b}\in\mathbb{S}^{m-1}} \left|\left|\left| \tilde{\boldsymbol{\eta}}^T\mathbf{b} \right|\right|\right|_{\psi_2} \le \tau \qquad \Longrightarrow \qquad R_{\mathcal{H}} = \tau
\end{aligned}$$

Similar to the proof for Lemma 27, we have

$$\begin{aligned}
\forall\ \mathbf{a}, \mathbf{a}' \in \mathcal{A}, \qquad & |||f_{\mathbf{a}} - f_{\mathbf{a}'}|||_{\psi_2} \\
&= \left|\left|\left| \mathbf{v}^T\mathbf{X}^T\boldsymbol{\Sigma}_*^{-1/2}\mathbf{u} - \mathbf{v}'^T\mathbf{X}^T\boldsymbol{\Sigma}_*^{-1/2}\mathbf{u}' \right|\right|\right|_{\psi_2} \\
&\le \left|\left|\left| (\mathbf{v}-\mathbf{v}')^T\mathbf{X}^T\mathbf{u} + \mathbf{v}'^T\mathbf{X}^T(\mathbf{u}-\mathbf{u}') \right|\right|\right|_{\psi_2} \\
&\le \|\mathbf{v}-\mathbf{v}'\|_2 \left|\left|\left| \frac{(\mathbf{v}-\mathbf{v}')^T}{\|\mathbf{v}-\mathbf{v}'\|_2}\mathbf{X}^T\mathbf{u} \right|\right|\right|_{\psi_2} + \|\mathbf{u}-\mathbf{u}'\|_2 \left|\left|\left| \mathbf{v}'^T\mathbf{X}^T\frac{(\mathbf{u}-\mathbf{u}')}{\|\mathbf{u}-\mathbf{u}'\|_2} \right|\right|\right|_{\psi_2}
\end{aligned}$$

$$
\begin{aligned}
&\leq \kappa\sqrt{\mu^+}\left(\left\|\mathbf{v}-\mathbf{v}'\right\|_2 + \left\|\mathbf{u}-\mathbf{u}'\right\|_2\right)\\
&\leq \sqrt{2}\kappa\sqrt{\mu^+}\cdot\sqrt{\|\mathbf{v}-\mathbf{v}'\|_2^2 + \|\mathbf{u}-\mathbf{u}'\|_2^2}\\
&= \sqrt{2}\kappa\sqrt{\mu^+}\left\|\mathbf{a}-\mathbf{a}'\right\|_2 \qquad \Longrightarrow \qquad K_{\mathcal{F}} = \sqrt{2}\kappa\sqrt{\mu^+}
\end{aligned}
$$

$$
\forall\ \mathbf{b},\mathbf{b}'\in\mathcal{B},\ \ |||h_{\mathbf{b}} - h_{\mathbf{b}'}|||_{\psi_2} = \left|\left|\left|\tilde{\boldsymbol{\eta}}^T(\mathbf{b}-\mathbf{b}')\right|\right|\right|_{\psi_2} \leq \tau\left\|\mathbf{b}-\mathbf{b}'\right\|_2 \qquad \Longrightarrow \qquad K_{\mathcal{H}} = \tau
$$

By invoking Theorem 2 and noting that $w(\mathbb{S}^{m-1}) = \Theta(\sqrt{m})$, $w(\mathcal{A}) = w(\mathbb{S}^{m-1}) + w(\mathcal{C}) \geq w(\mathcal{B})$, if $n \geq c_0 m$, we get

$$
\begin{aligned}
\gamma_n &\leq 2\sup_{\mathbf{u}\in\mathcal{C}}\sup_{\mathbf{v}\in\mathbb{S}^{m-1}}\sup_{\mathbf{b}\in\mathbb{S}^{m-1}}\left|\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{v}^T\mathbf{X}_i\mathbf{u}\right)\left(\tilde{\boldsymbol{\eta}}_i^T\mathbf{b}\right) - \mathbb{E}\left[\mathbf{v}^T\mathbf{X}\mathbf{u}\tilde{\boldsymbol{\eta}}^T\mathbf{b}\right]\right|\\
&\leq c_2\cdot\kappa\tau\sqrt{\mu^+}\cdot\frac{\sqrt{m}+w(\mathcal{C})}{\sqrt{n}}
\end{aligned}
$$

with probability at least $1-2\exp(-c_1 m)$. The proof is completed by renaming the constants. ■

9.B.4 Proof of Lemma 29

Proof: When the conditions of Lemma 25 is satisfied, the invertibility holds for all $\boldsymbol{\Sigma}\in\mathcal{M}$. using the relation $\boldsymbol{\eta} = \boldsymbol{\Sigma}_*^{1/2}\tilde{\boldsymbol{\eta}}$, we have

$$
\begin{aligned}
\beta_n &= \sup_{\boldsymbol{\Sigma}\in\mathcal{M}}\sup_{\mathbf{u}\in\mathcal{C}}\ \frac{2}{n}\sum_{i=1}^{n}\frac{\boldsymbol{\eta}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\mathbf{u}}{\|\boldsymbol{\Sigma}_*^{1/2}\boldsymbol{\Sigma}^{-1}\|_F}\\
&= \sup_{\boldsymbol{\Sigma}\in\mathcal{M}}\sup_{\mathbf{u}\in\mathcal{C}}\ \frac{2}{n}\sum_{i=1}^{n}\frac{\tilde{\boldsymbol{\eta}}_i^T\boldsymbol{\Sigma}_*^{1/2}\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\mathbf{u}}{\|\boldsymbol{\Sigma}_*^{1/2}\boldsymbol{\Sigma}^{-1}\|_F}\\
&\leq \underbrace{\sup_{\boldsymbol{\Lambda}\in\mathbb{S}^{m\times m-1}}\sup_{\mathbf{u}\in\mathcal{C}}\frac{2}{n}\sum_{i=1}^{n}\tilde{\boldsymbol{\eta}}_i^T\boldsymbol{\Lambda}\mathbf{X}_i\mathbf{u}}_{\nu_n}
\end{aligned}
$$

Therefore we just need to bound ν_n. Since the design and noise are independent, we will consider their randomness in a sequential fashion. The proof proceeds in two steps. First we show that the noises $\tilde{\boldsymbol{\eta}}_1, \tilde{\boldsymbol{\eta}}_2, \ldots, \tilde{\boldsymbol{\eta}}_n$ will behave "well" with high probability. By the word "well", we mean that the following event is true,

$$\mathcal{E} = \left\{ \{\tilde{\boldsymbol{\eta}}_i\} \;\middle|\; \sup_{\boldsymbol{\Lambda} \in \mathbb{S}^{m \times m-1}} \frac{1}{n} \sum_{i=1}^{n} \left\| \boldsymbol{\Lambda}^T \tilde{\boldsymbol{\eta}}_i \right\|_2^2 \leq 2 \right\} . \tag{9.51}$$

Denoting the columns of $\boldsymbol{\Lambda}$ by $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_m$, we have

$$\begin{aligned}
\sup_{\boldsymbol{\Lambda} \in \mathbb{S}^{m \times m-1}} \frac{1}{n} \sum_{i=1}^{n} \left\| \boldsymbol{\Lambda}^T \tilde{\boldsymbol{\eta}}_i \right\|_2^2 &= \sup_{\boldsymbol{\Lambda} \in \mathbb{S}^{m \times m-1}} \frac{1}{n} \sum_{i=1}^{n} \operatorname{Tr}\left(\boldsymbol{\Lambda}^T \tilde{\boldsymbol{\eta}}_i \tilde{\boldsymbol{\eta}}_i^T \boldsymbol{\Lambda} \right) \\
&= \sup_{\boldsymbol{\Lambda} \in \mathbb{S}^{m \times m-1}} \sum_{j=1}^{m} \boldsymbol{\lambda}_j^T \left(\frac{1}{n} \sum_{i=1}^{n} \tilde{\boldsymbol{\eta}}_i \tilde{\boldsymbol{\eta}}_i^T \right) \boldsymbol{\lambda}_j \\
&= \sup_{\boldsymbol{\Lambda} \in \mathbb{S}^{m \times m-1}} \sum_{j=1}^{m} \|\boldsymbol{\lambda}_j\|_2^2 \cdot \left\| \frac{1}{n} \sum_{i=1}^{n} \tilde{\boldsymbol{\eta}}_i \tilde{\boldsymbol{\eta}}_i^T \right\|_2 \\
&= \left\| \frac{1}{n} \sum_{i=1}^{n} \tilde{\boldsymbol{\eta}}_i \tilde{\boldsymbol{\eta}}_i^T \right\|_2
\end{aligned}$$

By Proposition 18, if $n \geq c_0 \tau^4 m$, we have

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \tilde{\boldsymbol{\eta}}_i \tilde{\boldsymbol{\eta}}_i^T \right\|_2 \leq 1 + \left\| \frac{1}{n} \sum_{i=1}^{n} \tilde{\boldsymbol{\eta}}_i \tilde{\boldsymbol{\eta}}_i^T - \mathbf{I} \right\|_2 \leq 2$$

with probability at least $1 - 2\exp(-c_1 m)$.

Next we consider the randomness of $\mathbf{X}_i$ given that $\tilde{\boldsymbol{\eta}}_i$'s are fixed and $\mathcal{E}$ is true. Construct the stochastic process $\left\{ Z_{\mathbf{t}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\boldsymbol{\eta}}_i^T \boldsymbol{\Lambda} \mathbf{X}_i \mathbf{u} \right\}_{\mathbf{t} \in \mathcal{T}}$, where $\mathcal{T} = \mathbb{S}^{m \times m-1} \times \mathcal{C} \subset \mathbb{R}^{m \times m+p}$ and $\mathbf{t} = (\operatorname{vec}(\boldsymbol{\Lambda}), \mathbf{u})$. Note that

$$\forall\, \mathbf{t}, \mathbf{t}' \in \mathcal{T}, \quad \left\| \mathbf{t} - \mathbf{t}' \right\|_2 = \sqrt{\|\boldsymbol{\Lambda} - \boldsymbol{\Lambda}'\|_F^2 + \|\mathbf{u} - \mathbf{u}'\|_2^2} \leq 2\sqrt{2} \quad \Longrightarrow \quad \operatorname{diam}(\mathcal{T}) \leq 2\sqrt{2}$$

In order to apply Theorem 1 to $\{Z_{\mathbf{t}}\}$, we first verify the required condition.

$$
\begin{aligned}
\forall\ \mathbf{t},\mathbf{t}' \in \mathcal{T}, \qquad & \left\|\left\| Z_{\mathbf{t}} - Z_{\mathbf{t}'} \right\|\right\|_{\psi_2} \\
&= \left\|\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\boldsymbol{\eta}}_i^T \boldsymbol{\Lambda} \mathbf{X}_i \mathbf{u} - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\boldsymbol{\eta}}_i^T \boldsymbol{\Lambda}' \mathbf{X}_i \mathbf{u}' \right\|\right\|_{\psi_2} \\
&\leq \left\|\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\boldsymbol{\eta}}_i^T (\boldsymbol{\Lambda} - \boldsymbol{\Lambda}') \mathbf{X}_i \mathbf{u} \right\|\right\|_{\psi_2} + \left\|\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\boldsymbol{\eta}}_i^T \boldsymbol{\Lambda}' \mathbf{X}_i (\mathbf{u} - \mathbf{u}') \right\|\right\|_{\psi_2} \\
&\overset{(a)}{\leq} c_2 \sqrt{\frac{1}{n} \sum_{i=1}^{n} \|(\boldsymbol{\Lambda} - \boldsymbol{\Lambda}')^T \tilde{\boldsymbol{\eta}}_i\|_2^2} \cdot \sup_{\mathbf{v} \in \mathbb{S}^{m-1}} \left\|\left\| \mathbf{v}^T \mathbf{X} \mathbf{u} \right\|\right\|_{\psi_2} \\
&\qquad + c_2 \sqrt{\frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{\Lambda}'^T \tilde{\boldsymbol{\eta}}_i\|_2^2} \cdot \left\| \mathbf{u} - \mathbf{u}' \right\|_2 \cdot \sup_{\mathbf{v} \in \mathbb{S}^{m-1}} \left\|\left\| \mathbf{v}^T \mathbf{X} \frac{\mathbf{u} - \mathbf{u}'}{\|\mathbf{u} - \mathbf{u}'\|_2} \right\|\right\|_{\psi_2} \\
&\leq \sqrt{2} c_2 \kappa \sqrt{\mu^+} \left(\left\| \boldsymbol{\Lambda} - \boldsymbol{\Lambda}' \right\|_F + \left\| \mathbf{u} - \mathbf{u}' \right\|_2 \right) \\
&\leq 2 c_2 \kappa \sqrt{\mu^+} \left\| \begin{bmatrix} \operatorname{vec}(\boldsymbol{\Lambda}) \\ \mathbf{u} \end{bmatrix} - \begin{bmatrix} \operatorname{vec}(\boldsymbol{\Lambda}') \\ \mathbf{u}' \end{bmatrix} \right\|_2 \qquad \Longrightarrow \qquad K = 2 c_2 \kappa \sqrt{\mu^+}\ ,
\end{aligned}
$$

where step (a) follows from Proposition 10. By Theorem 1, we have for fixed $\{\tilde{\boldsymbol{\eta}}_i\}$ under event $\mathcal{E}$,

$$
\nu_n = \frac{2}{\sqrt{n}} \cdot \sup_{\mathbf{t} \in \mathcal{T}} Z_{\mathbf{t}} = \frac{1}{\sqrt{n}} \cdot \sup_{\mathbf{t},\mathbf{t}' \in \mathcal{T}} |Z_{\mathbf{t}} - Z_{\mathbf{t}'}| \leq c_3 \cdot \frac{\kappa \sqrt{\mu^+} \cdot w(\mathcal{T})}{\sqrt{n}}
$$

with probability at least $1 - c_4 \exp\left(-\frac{w^2(\mathcal{T})}{\operatorname{diam}^2(\mathcal{T})}\right) \geq 1 - c_4 \exp\left(-\frac{w^2(\mathcal{T})}{8}\right)$. Now we combine the randomness of $\mathbf{X}_i$ and $\tilde{\boldsymbol{\eta}}_i$, and get

$$
\begin{aligned}
& \mathbb{P}_{\mathbf{X},\tilde{\boldsymbol{\eta}}} \left(\nu_n \leq c_3 \cdot \frac{\kappa \sqrt{\mu^+} \cdot w(\mathcal{T})}{\sqrt{n}} \right) \\
= & \int \mathbb{P}_{\mathbf{X}} \left(\nu_n \leq c_3 \cdot \frac{\kappa \sqrt{\mu^+} \cdot w(\mathcal{T})}{\sqrt{n}} \ \middle|\ \{\tilde{\boldsymbol{\eta}}_i\} \right) p\left(\tilde{\boldsymbol{\eta}}_1, \ldots, \tilde{\boldsymbol{\eta}}_n\right) d\tilde{\boldsymbol{\eta}}_1 \ldots d\tilde{\boldsymbol{\eta}}_n
\end{aligned}
$$

$$\geq \int_{\mathcal{E}} \mathbb{P}_{\mathbf{X}} \left(\nu_n \leq c_3 \cdot \frac{\kappa\sqrt{\mu^+} \cdot w(\mathcal{T})}{\sqrt{n}} \;\middle|\; \{\tilde{\boldsymbol{\eta}}_i\} \right) p\left(\tilde{\boldsymbol{\eta}}_1, \ldots, \tilde{\boldsymbol{\eta}}_n\right) d\tilde{\boldsymbol{\eta}}_1 \ldots d\tilde{\boldsymbol{\eta}}_n$$
$$\geq \left(1 - c_4 \exp\left(-\frac{w^2(\mathcal{T})}{8}\right)\right) \cdot \mathbb{P}\left(\mathcal{E}\right)$$
$$\geq \left(1 - c_4 \exp\left(-\frac{w^2(\mathcal{T})}{8}\right)\right) \left(1 - 2\exp\left(-c_1 m\right)\right)$$
$$\geq 1 - 2\exp\left(-c_1 m\right) - c_4 \exp\left(-\frac{w^2(\mathcal{T})}{8}\right)$$
$$\geq 1 - c_5 \exp(-c_6 m) \ ,$$

where the last step follows from $w(\mathcal{T}) = w(\mathbb{S}^{m\times m-1} \times \mathcal{C}) = w(\mathbb{S}^{m\times m-1}) + w(\mathcal{C}) = \Theta(m) + w(\mathcal{C})$. Since the invertibility for $\mathcal{M}$ is implied by the conditions of Lemma 25, we have that if $n \geq \max\{n_0, C_0\tau^4 m\}$,

$$\beta_n \leq c_7 \cdot \frac{\kappa\sqrt{\mu^+}(m + w(\mathcal{C}))}{\sqrt{n}}$$

with probability at least $1-\epsilon-c_5\exp(-c_6 m)$. Finally we complete the proof by renaming the constants. ■

9.B.5 Proof of Lemma 30

Proof: Throughout the proof, e_0 is set as $\sqrt{\frac{\sigma_*^-}{\mu^+}}$, and we will use the shorthand notation β_n and $\mathcal{M}$ for $\beta_n(\mathcal{M}(e_0))$ and $\mathcal{M}(e_0)$. First we introduce the following notations

$$\mathcal{S}' = e_0 \cdot \mathcal{S} = \{e_0\mathbf{u} \mid \mathbf{u} \in \mathcal{S}\}$$
$$\boldsymbol{\Gamma}_{\mathbf{w}} = \mathbb{E}\left[\mathbf{X}\mathbf{w}\mathbf{w}^T\mathbf{X}^T\right]$$
$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\boldsymbol{\theta}-\boldsymbol{\theta}_*}$$
$$\hat{\boldsymbol{\Gamma}}_{\mathbf{w}} = -\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i\mathbf{w}\boldsymbol{\eta}_i^T - \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\eta}_i\mathbf{w}^T\mathbf{X}_i + \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i\mathbf{w}\mathbf{w}^T\mathbf{X}_i^T$$

$$\hat{\mathbf{\Sigma}}_{\boldsymbol{\theta}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\eta}_i \boldsymbol{\eta}_i^T + \hat{\mathbf{\Gamma}}_{\boldsymbol{\theta}-\boldsymbol{\theta}_*} = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\theta}\right)\left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\theta}\right)^T$$

Note that $\mu^- \leq \lambda_{\min}(\mathbf{\Gamma}_{\mathbf{w}}) \leq \lambda_{\max}(\mathbf{\Gamma}_{\mathbf{w}}) \leq \mu^+$ for any $\mathbf{w} \in \mathbb{S}^{p-1}$, $\mathbf{\Gamma}_{\mathbf{w}} = \mathbb{E}[\hat{\mathbf{\Gamma}}_{\mathbf{w}}]$, $\mathbf{\Sigma}_{\boldsymbol{\theta}} = \mathbb{E}[\hat{\mathbf{\Sigma}}_{\boldsymbol{\theta}}]$ and $\mathcal{M} \subseteq \{\hat{\mathbf{\Sigma}}_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \mathcal{S}' + \boldsymbol{\theta}_*\}$. Then we decompose β_n as

$$\begin{aligned}
\beta_n &= \sup_{\mathbf{\Sigma}\in\mathcal{M}} \sup_{\mathbf{u}\in\mathcal{C}} \frac{2}{n}\sum_{i=1}^{n} \frac{\boldsymbol{\eta}_i^T \mathbf{\Sigma}^{-1}\mathbf{X}_i\mathbf{u}}{\|\mathbf{\Sigma}_*^{1/2}\mathbf{\Sigma}^{-1}\|_F} = \sup_{\mathbf{\Sigma}\in\mathcal{M}} \sup_{\mathbf{u}\in\mathcal{C}} \frac{2}{n}\sum_{i=1}^{n} \frac{\tilde{\boldsymbol{\eta}}_i^T \mathbf{\Sigma}_*^{1/2}\mathbf{\Sigma}^{-1}\mathbf{X}_i\mathbf{u}}{\|\mathbf{\Sigma}_*^{1/2}\mathbf{\Sigma}^{-1}\|_F} \\
&\leq \sup_{\boldsymbol{\theta}\in\mathcal{S}'+\boldsymbol{\theta}_*} \sup_{\mathbf{u}\in\mathcal{C}} \frac{2}{n}\sum_{i=1}^{n} \tilde{\boldsymbol{\eta}}_i^T \left(\frac{\mathbf{\Sigma}_*^{1/2}\hat{\mathbf{\Sigma}}_{\boldsymbol{\theta}}^{-1}}{\|\mathbf{\Sigma}_*^{1/2}\hat{\mathbf{\Sigma}}_{\boldsymbol{\theta}}^{-1}\|_F} - \frac{\mathbf{\Sigma}_*^{1/2}\mathbf{\Sigma}_{\boldsymbol{\theta}}^{-1}}{\|\mathbf{\Sigma}_*^{1/2}\mathbf{\Sigma}_{\boldsymbol{\theta}}^{-1}\|_F} \right) \mathbf{X}_i\mathbf{u} \\
&\quad + \sup_{\boldsymbol{\theta}\in\mathcal{S}'+\boldsymbol{\theta}_*} \sup_{\mathbf{u}\in\mathcal{C}} \frac{2}{n}\sum_{i=1}^{n} \frac{\tilde{\boldsymbol{\eta}}_i^T \mathbf{\Sigma}_*^{1/2}\mathbf{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{X}_i\mathbf{u}}{\|\mathbf{\Sigma}_*^{1/2}\mathbf{\Sigma}_{\boldsymbol{\theta}}^{-1}\|_F} \\
&\leq \underbrace{\sup_{\mathbf{\Lambda}\in\mathbb{S}^{m\times m-1}} \sup_{\mathbf{u}\in\mathcal{C}} \frac{2}{n}\sum_{i=1}^{n} \tilde{\boldsymbol{\eta}}_i^T \mathbf{\Lambda}\mathbf{X}_i\mathbf{u}}_{\nu_n} \cdot \underbrace{\sup_{\boldsymbol{\theta}\in\mathcal{S}'+\boldsymbol{\theta}_*} \left\| \frac{\mathbf{\Sigma}_*^{1/2}\hat{\mathbf{\Sigma}}_{\boldsymbol{\theta}}^{-1}}{\|\mathbf{\Sigma}_*^{1/2}\hat{\mathbf{\Sigma}}_{\boldsymbol{\theta}}^{-1}\|_F} - \frac{\mathbf{\Sigma}_*^{1/2}\mathbf{\Sigma}_{\boldsymbol{\theta}}^{-1}}{\|\mathbf{\Sigma}_*^{1/2}\mathbf{\Sigma}_{\boldsymbol{\theta}}^{-1}\|_F} \right\|_F}_{\zeta_n} \\
&\quad + \underbrace{\sup_{\boldsymbol{\theta}\in\mathcal{S}'+\boldsymbol{\theta}_*} \sup_{\mathbf{u}\in\mathcal{C}} \frac{2}{n}\sum_{i=1}^{n} \frac{\tilde{\boldsymbol{\eta}}_i^T \mathbf{\Sigma}_*^{1/2}\mathbf{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{X}_i\mathbf{u}}{\|\mathbf{\Sigma}_*^{1/2}\mathbf{\Sigma}_{\boldsymbol{\theta}}^{-1}\|_F}}_{\phi_n}
\end{aligned}$$

where ν_n is analyzed in the proof of Lemma 29. Therefore we focus on bounding ζ_n and ϕ_n. We first try to bound ζ_n,

$$\begin{aligned}
\zeta_n &= \sup_{\boldsymbol{\theta}\in\mathcal{S}'+\boldsymbol{\theta}_*} \left\| \frac{\mathbf{\Sigma}_*^{1/2}\hat{\mathbf{\Sigma}}_{\boldsymbol{\theta}}^{-1}}{\|\mathbf{\Sigma}_*^{1/2}\hat{\mathbf{\Sigma}}_{\boldsymbol{\theta}}^{-1}\|_F} - \frac{\mathbf{\Sigma}_*^{1/2}\mathbf{\Sigma}_{\boldsymbol{\theta}}^{-1}}{\|\mathbf{\Sigma}_*^{1/2}\mathbf{\Sigma}_{\boldsymbol{\theta}}^{-1}\|_F} \right\|_F \\
&\leq \sup_{\boldsymbol{\theta}\in\mathcal{S}'+\boldsymbol{\theta}_*} \left\| \frac{\mathbf{\Sigma}_*^{1/2}\hat{\mathbf{\Sigma}}_{\boldsymbol{\theta}}^{-1}}{\|\mathbf{\Sigma}_*^{1/2}\hat{\mathbf{\Sigma}}_{\boldsymbol{\theta}}^{-1}\|_F} - \frac{\mathbf{\Sigma}_*^{1/2}\hat{\mathbf{\Sigma}}_{\boldsymbol{\theta}}^{-1}}{\|\mathbf{\Sigma}_*^{1/2}\mathbf{\Sigma}_{\boldsymbol{\theta}}^{-1}\|_F} \right\|_F \\
&\quad + \sup_{\boldsymbol{\theta}\in\mathcal{S}'+\boldsymbol{\theta}_*} \left\| \frac{\mathbf{\Sigma}_*^{1/2}\hat{\mathbf{\Sigma}}_{\boldsymbol{\theta}}^{-1}}{\|\mathbf{\Sigma}_*^{1/2}\mathbf{\Sigma}_{\boldsymbol{\theta}}^{-1}\|_F} - \frac{\mathbf{\Sigma}_*^{1/2}\mathbf{\Sigma}_{\boldsymbol{\theta}}^{-1}}{\|\mathbf{\Sigma}_*^{1/2}\mathbf{\Sigma}_{\boldsymbol{\theta}}^{-1}\|_F} \right\|_F \\
&\leq \sup_{\boldsymbol{\theta}\in\mathcal{S}'+\boldsymbol{\theta}_*} \left| \frac{\|\mathbf{\Sigma}_*^{1/2}\hat{\mathbf{\Sigma}}_{\boldsymbol{\theta}}^{-1}\|_F - \|\mathbf{\Sigma}_*^{1/2}\mathbf{\Sigma}_{\boldsymbol{\theta}}^{-1}\|_F}{\|\mathbf{\Sigma}_*^{1/2}\mathbf{\Sigma}_{\boldsymbol{\theta}}^{-1}\|_F} \right| + \sup_{\boldsymbol{\theta}\in\mathcal{S}'+\boldsymbol{\theta}_*} \frac{\left\|\mathbf{\Sigma}_*^{1/2}\hat{\mathbf{\Sigma}}_{\boldsymbol{\theta}}^{-1} - \mathbf{\Sigma}_*^{1/2}\mathbf{\Sigma}_{\boldsymbol{\theta}}^{-1}\right\|_F}{\|\mathbf{\Sigma}_*^{1/2}\mathbf{\Sigma}_{\boldsymbol{\theta}}^{-1}\|_F}
\end{aligned}$$

$$
\begin{aligned}
&\leq 2 \sup_{\boldsymbol{\theta} \in \mathcal{S}' + \boldsymbol{\theta}_*} \frac{\left\| \boldsymbol{\Sigma}_*^{1/2} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{-1} - \boldsymbol{\Sigma}_*^{1/2} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \right\|_F}{\| \boldsymbol{\Sigma}_*^{1/2} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \|_F} \\
&\leq 2 \sup_{\boldsymbol{\theta} \in \mathcal{S}' + \boldsymbol{\theta}_*} \frac{\left\| \boldsymbol{\Sigma}_*^{1/2} (\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}) \boldsymbol{\Sigma}_*^{1/2} \right\|_2 \cdot \left\| \boldsymbol{\Sigma}_*^{-1/2} \right\|_F}{\lambda_{\min} \left(\boldsymbol{\Sigma}_*^{1/2} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \boldsymbol{\Sigma}_*^{1/2} \right) \cdot \left\| \boldsymbol{\Sigma}_*^{-1/2} \right\|_F} \\
&\leq 2 \sup_{\boldsymbol{\theta} \in \mathcal{S}' + \boldsymbol{\theta}_*} \frac{\left\| \boldsymbol{\Sigma}_*^{-1/2} (\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) \boldsymbol{\Sigma}_*^{-1/2} \right\|_2 \cdot \left\| \boldsymbol{\Sigma}_*^{1/2} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{-1} \boldsymbol{\Sigma}_*^{1/2} \right\|_2 \cdot \left\| \boldsymbol{\Sigma}_*^{1/2} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \boldsymbol{\Sigma}_*^{1/2} \right\|_2}{\lambda_{\min} \left(\boldsymbol{\Sigma}_*^{1/2} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \boldsymbol{\Sigma}_*^{1/2} \right)} \\
&= 2 \sup_{\boldsymbol{\theta} \in \mathcal{S}' + \boldsymbol{\theta}_*} \frac{\left\| \boldsymbol{\Sigma}_*^{-1/2} (\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) \boldsymbol{\Sigma}_*^{-1/2} \right\|_2 \cdot \lambda_{\max} \left(\boldsymbol{\Sigma}_*^{-1/2} \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{\Sigma}_*^{-1/2} \right)}{\lambda_{\min} \left(\boldsymbol{\Sigma}_*^{-1/2} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}} \boldsymbol{\Sigma}_*^{-1/2} \right) \cdot \lambda_{\min} \left(\boldsymbol{\Sigma}_*^{-1/2} \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{\Sigma}_*^{-1/2} \right)} \\
&\leq \frac{2 \sup_{\boldsymbol{\theta} \in \mathcal{S}' + \boldsymbol{\theta}_*} \left\| \boldsymbol{\Sigma}_*^{-1/2} (\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) \boldsymbol{\Sigma}_*^{-1/2} \right\|_2 \cdot \sup_{\mathbf{w} \in \mathcal{S}'} \lambda_{\max} \left(\boldsymbol{\Sigma}_*^{-1/2} (\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}}) \boldsymbol{\Sigma}_*^{-1/2} \right)}{\inf_{\boldsymbol{\theta} \in \mathcal{S}' + \boldsymbol{\theta}_*} \lambda_{\min} \left(\boldsymbol{\Sigma}_*^{-1/2} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}} \boldsymbol{\Sigma}_*^{-1/2} \right) \cdot \inf_{\mathbf{w} \in \mathcal{S}'} \lambda_{\min} \left(\boldsymbol{\Sigma}_*^{-1/2} (\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}}) \boldsymbol{\Sigma}_*^{-1/2} \right)} \\
&\leq \frac{2 \sup_{\boldsymbol{\theta} \in \mathcal{S}' + \boldsymbol{\theta}_*} \left\| \boldsymbol{\Sigma}_*^{-1/2} (\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) \boldsymbol{\Sigma}_*^{-1/2} \right\|_2 \cdot \left(1 + \frac{\mu^+}{\sigma_*^-} \cdot \sup_{\mathbf{w} \in \mathcal{S}'} \|\mathbf{w}\|_2^2 \right)}{(1 - 2\delta_n) \cdot \left(1 + \frac{\mu^-}{\sigma_*^+} \cdot \inf_{\mathbf{w} \in \mathcal{S}'} \|\mathbf{w}\|_2^2 \right)} \\
&\leq 8 \sup_{\boldsymbol{\theta} \in \mathcal{S}' + \boldsymbol{\theta}_*} \left\| \boldsymbol{\Sigma}_*^{-1/2} (\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) \boldsymbol{\Sigma}_*^{-1/2} \right\|_2
\end{aligned}
$$

where the last two steps use the conditions in Lemma 25 and borrow some derivations from its proof. The last term can be further bounded as follows,

$$
\begin{aligned}
&\sup_{\boldsymbol{\theta} \in \mathcal{S}' + \boldsymbol{\theta}_*} \left\| \boldsymbol{\Sigma}_*^{-\frac{1}{2}} (\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) \boldsymbol{\Sigma}_*^{-\frac{1}{2}} \right\|_2 \\
&= \sup_{\mathbf{w} \in \mathcal{S}'} \left\| \boldsymbol{\Sigma}_*^{-\frac{1}{2}} \left(\frac{1}{n} \sum_{i=1}^n \boldsymbol{\eta}_i \boldsymbol{\eta}_i^T + \hat{\boldsymbol{\Gamma}}_{\mathbf{w}} - \boldsymbol{\Sigma}_* - \boldsymbol{\Gamma}_{\mathbf{w}} \right) \boldsymbol{\Sigma}_*^{-\frac{1}{2}} \right\|_2 \\
&\leq \left\| \frac{1}{n} \sum_{i=1}^n \tilde{\boldsymbol{\eta}}_i \tilde{\boldsymbol{\eta}}_i^T - \mathbf{I} \right\|_2 + \sup_{\mathbf{w} \in \mathcal{S}'} \left(\left\| \frac{1}{n} \sum_{i=1}^n \boldsymbol{\Sigma}_*^{-\frac{1}{2}} \mathbf{X}_i \mathbf{w} \tilde{\boldsymbol{\eta}}_i^T \right\|_2 + \left\| \frac{1}{n} \sum_{i=1}^n \tilde{\boldsymbol{\eta}}_i \mathbf{w}^T {\mathbf{X}_i}^T \boldsymbol{\Sigma}_*^{-\frac{1}{2}} \right\|_2 \right) \\
&\quad + \sup_{\mathbf{w} \in \mathcal{S}'} \left\| \boldsymbol{\Sigma}_*^{-1/2} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \mathbf{w} \mathbf{w}^T {\mathbf{X}_i}^T - \boldsymbol{\Gamma}_{\mathbf{w}} \right) \boldsymbol{\Sigma}_*^{-1/2} \right\|_2 \\
&\leq \delta_n + \frac{e_0}{\sqrt{\sigma_*^-}} \cdot \sup_{\mathbf{w} \in \mathcal{C}} \left\| \frac{2}{n} \sum_{i=1}^n \mathbf{X}_i \mathbf{w} \tilde{\boldsymbol{\eta}}_i^T \right\|_2 + \frac{e_0^2}{\sigma_*^-} \cdot \sup_{\mathbf{w} \in \mathcal{C}} \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \mathbf{w} \mathbf{w}^T {\mathbf{X}_i}^T - \boldsymbol{\Gamma}_{\mathbf{w}} \right\|_2
\end{aligned}
$$

$$
\begin{aligned}
&\le \delta_n + \frac{e_0\gamma_n}{\sqrt{\sigma_*^-}} + \frac{e_0^2}{\sigma_*^-}\cdot \sup_{\mathbf{v}\in\mathbb{S}^{m-1}}\sup_{\mathbf{w}\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^n(\mathbf{w}^T\mathbf{X}_i^T\mathbf{v})^2 - \mathbb{E}(\mathbf{w}^T\mathbf{X}^T\mathbf{v})^2\right| \\
&\le c_1\tau^2\sqrt{\frac{m}{n}} + \frac{c_2\kappa\tau(\sqrt{m}+w(\mathcal{C}))}{\sqrt{n}} + \frac{c_3\kappa^2(\sqrt{m}+w(\mathcal{C}))}{\sqrt{n}}
\end{aligned}
$$

which holds with probability at least $1 - c_4\exp(-c_5 m)$ when $n \ge c_6\max\left\{\tau^4, 1\right\}\cdot \max\left\{w^2(\mathcal{C}), m\right\}$. The last step follows from Proposition 18, Lemma 28 and intermediate results in the proof of Lemma 27. Hence ζ_n can be bounded by

$$
\zeta_n \le c_7\cdot\max\left\{\tau^2,\kappa^2\right\}\cdot\frac{\sqrt{m}+w(\mathcal{C})}{\sqrt{n}}
$$

Now we turn to bounding ϕ_n. Following the idea for proving Lemma 29, we also consider the randomness of $\{\tilde{\boldsymbol{\eta}}_i\}$ and $\{\mathbf{X}_i\}$ sequentially. For $\{\tilde{\boldsymbol{\eta}}_i\}$, we first have that the event

$$
\mathcal{E} = \left\{\{\tilde{\boldsymbol{\eta}}_i\}\ \middle|\ \sup_{\boldsymbol{\Lambda}\in\mathbb{S}^{m\times m-1}}\frac{1}{n}\sum_{i=1}^n\left\|\boldsymbol{\Lambda}^T\tilde{\boldsymbol{\eta}}_i\right\|_2^2 \le 2\right\}
$$

holds with probability at least $1 - 2\exp(-c_1' m)$ if $n \ge c_0'\tau^4 m$, which is shown in the proof of Lemma 29. Now we consider the randomness of $\{\mathbf{X}_i\}$ under any fixed $\{\boldsymbol{\eta}_i\}\in\mathcal{E}$. We have

$$
\begin{aligned}
\phi_n &= \sup_{\boldsymbol{\theta}\in\mathcal{S}'+\boldsymbol{\theta}_*}\sup_{\mathbf{u}\in\mathcal{C}}\frac{2}{n}\sum_{i=1}^n\frac{\tilde{\boldsymbol{\eta}}_i^T\boldsymbol{\Sigma}_*^{1/2}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{X}_i\mathbf{u}}{\|\boldsymbol{\Sigma}_*^{1/2}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\|_F} \\
&\le \frac{1}{e_0}\cdot\sup_{\mathbf{w}\in\mathcal{S}'}\sup_{\mathbf{u}\in\mathcal{S}'}\frac{2}{n}\sum_{i=1}^n\frac{\tilde{\boldsymbol{\eta}}_i^T\boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_*+\boldsymbol{\Gamma}_{\mathbf{w}})^{-1}\mathbf{X}_i\mathbf{u}}{\|\boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_*+\boldsymbol{\Gamma}_{\mathbf{w}})^{-1}\|_F} = \frac{2}{e_0\sqrt{n}}\cdot\sup_{\mathbf{t}\in\mathcal{T}}Z_{\mathbf{t}}\ ,
\end{aligned}
$$

where $Z_{\mathbf{t}} = \frac{1}{\sqrt{n}}\sum_{i=1}^n\frac{\tilde{\boldsymbol{\eta}}_i^T\boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_*+\boldsymbol{\Gamma}_{\mathbf{w}})^{-1}\mathbf{X}_i\mathbf{u}}{\|\boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_*+\boldsymbol{\Gamma}_{\mathbf{w}})^{-1}\|_F}$, $\mathbf{t} = (\mathbf{w},\mathbf{u})$ and $\mathcal{T} = \mathcal{S}'\times\mathcal{S}'$. Note that

$$
\forall\ \mathbf{t},\mathbf{t}'\in\mathcal{T},\quad \left\|\mathbf{t}-\mathbf{t}'\right\|_2 = \sqrt{\|\mathbf{w}-\mathbf{w}'\|_F^2 + \|\mathbf{u}-\mathbf{u}'\|_2^2} \le 2\sqrt{2}e_0 \quad\Longrightarrow\quad \operatorname{diam}(\mathcal{T}) \le 2\sqrt{2}e_0
$$

Then we try to bound the stochastic process $\{Z_{\mathbf{t}}\}_{\mathbf{t}\in\mathcal{T}}$ using Theorem 1. We start with verifying the required condition.

$$
\begin{aligned}
&\forall\ \mathbf{t},\mathbf{t}'\in\mathcal{T},\\
&\quad \left\|\left\| Z_{\mathbf{t}} - Z_{\mathbf{t}'} \right\|\right\|_{\psi_2}\\
&= \left\|\left\| \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{\tilde{\boldsymbol{\eta}}_i^T \boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}})^{-1}\mathbf{X}_i\mathbf{u}}{\|\boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}})^{-1}\|_F} - \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{\tilde{\boldsymbol{\eta}}_i^T \boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}'})^{-1}\mathbf{X}_i\mathbf{u}'}{\|\boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}'})^{-1}\|_F} \right\|\right\|_{\psi_2}\\
&\le \left\|\left\| \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \tilde{\boldsymbol{\eta}}_i^T \left(\frac{\boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}})^{-1}}{\|\boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}})^{-1}\|_F} - \frac{\boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}'})^{-1}}{\|\boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}'})^{-1}\|_F} \right) \mathbf{X}_i\mathbf{u} \right\|\right\|_{\psi_2}\\
&\quad + \left\|\left\| \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{\tilde{\boldsymbol{\eta}}_i^T \boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}'})^{-1}\mathbf{X}_i(\mathbf{u}-\mathbf{u}')}{\|\boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}'})^{-1}\|_F} \right\|\right\|_{\psi_2}\\
&\overset{(a)}{\le} c_2' \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left\| \left(\frac{\boldsymbol{\Sigma}_*^{\frac{1}{2}}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}})^{-1}}{\|\boldsymbol{\Sigma}_*^{\frac{1}{2}}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}})^{-1}\|_F} - \frac{\boldsymbol{\Sigma}_*^{\frac{1}{2}}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}'})^{-1}}{\|\boldsymbol{\Sigma}_*^{\frac{1}{2}}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}'})^{-1}\|_F} \right)^T \tilde{\boldsymbol{\eta}}_i \right\|_2^2}\\
&\qquad \times \sup_{\mathbf{v}\in\mathbb{S}^{m-1}} \left\|\left\| \mathbf{v}^T\mathbf{X}\mathbf{u} \right\|\right\|_{\psi_2} +\\
&\quad c_2' \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left\| \left(\frac{\boldsymbol{\Sigma}_*^{\frac{1}{2}}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}'})^{-1}}{\|\boldsymbol{\Sigma}_*^{\frac{1}{2}}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}'})^{-1}\|_F} \right)^T \tilde{\boldsymbol{\eta}}_i \right\|_2^2} \cdot \left\|\mathbf{u}-\mathbf{u}'\right\|_2 \cdot \sup_{\mathbf{v}\in\mathbb{S}^{m-1}} \left\|\left\| \mathbf{v}^T\mathbf{X}\frac{\mathbf{u}-\mathbf{u}'}{\|\mathbf{u}-\mathbf{u}'\|_2} \right\|\right\|_{\psi_2}\\
&\overset{(b)}{\le} \sqrt{2}c_2'\kappa\sqrt{\mu^+} \left(e_0 \left\| \frac{\boldsymbol{\Sigma}_*^{\frac{1}{2}}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}})^{-1}}{\|\boldsymbol{\Sigma}_*^{\frac{1}{2}}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}})^{-1}\|_F} - \frac{\boldsymbol{\Sigma}_*^{\frac{1}{2}}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}'})^{-1}}{\|\boldsymbol{\Sigma}_*^{\frac{1}{2}}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}'})^{-1}\|_F} \right\|_F + \left\|\mathbf{u}-\mathbf{u}'\right\|_2 \right)\\
&\overset{(c)}{\le} \sqrt{2}c_2'\kappa\sqrt{\mu^+} \left(8\left\|\mathbf{w}-\mathbf{w}'\right\|_2 + \left\|\mathbf{u}-\mathbf{u}'\right\|_2 \right)\\
&\le 16c_2'\kappa\sqrt{\mu^+} \left\| \begin{bmatrix}\mathbf{w}\\ \mathbf{u}\end{bmatrix} - \begin{bmatrix}\mathbf{w}'\\ \mathbf{u}'\end{bmatrix} \right\|_2 \qquad \Longrightarrow \qquad K = 16c_2'\kappa\sqrt{\mu^+}\ ,
\end{aligned}
$$

where step (a) follows from Proposition 10 and step (b) follows from the event $\mathcal{E}$. Step (c) follows from the calculation below (similar to bounding ζ_n),

$$
\begin{aligned}
&\left\| \frac{\boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}})^{-1}}{\|\boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}})^{-1}\|_F} - \frac{\boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}'})^{-1}}{\|\boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}'})^{-1}\|_F} \right\|_F \\
&\leq \left\| \frac{\boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}})^{-1}}{\|\boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}})^{-1}\|_F} - \frac{\boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}'})^{-1}}{\|\boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}})^{-1}\|_F} \right\|_F \\
&\quad + \left\| \frac{\boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}'})^{-1}}{\|\boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}})^{-1}\|_F} - \frac{\boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}'})^{-1}}{\|\boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}'})^{-1}\|_F} \right\|_F \\
&\leq \frac{2\left\|\boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}})^{-1} - \boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}'})^{-1}\right\|_F}{\left\|\boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}})^{-1}\right\|_F} \\
&\leq \frac{2\left\|\boldsymbol{\Sigma}_*^{1/2}\left((\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}})^{-1} - (\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}'})^{-1}\right)\boldsymbol{\Sigma}_*^{1/2}\right\|_2}{\lambda_{\min}\left(\boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}})^{-1}\boldsymbol{\Sigma}_*^{1/2}\right)} \\
&\leq \frac{2\left\|\boldsymbol{\Sigma}_*^{-\frac{1}{2}}\left(\boldsymbol{\Gamma}_{\mathbf{w}} - \boldsymbol{\Gamma}_{\mathbf{w}'}\right)\boldsymbol{\Sigma}_*^{-\frac{1}{2}}\right\|_2 \left\|\boldsymbol{\Sigma}_*^{\frac{1}{2}}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}})^{-1}\boldsymbol{\Sigma}_*^{\frac{1}{2}}\right\|_2 \left\|\boldsymbol{\Sigma}_*^{\frac{1}{2}}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}'})^{-1}\boldsymbol{\Sigma}_*^{\frac{1}{2}}\right\|_2}{\lambda_{\min}\left(\boldsymbol{\Sigma}_*^{\frac{1}{2}}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}})^{-1}\boldsymbol{\Sigma}_*^{\frac{1}{2}}\right)} \\
&= \frac{2\left\|\boldsymbol{\Sigma}_*^{-1/2}\left(\boldsymbol{\Gamma}_{\mathbf{w}} - \boldsymbol{\Gamma}_{\mathbf{w}'}\right)\boldsymbol{\Sigma}_*^{-1/2}\right\|_2 \cdot \lambda_{\max}\left(\boldsymbol{\Sigma}_*^{-1/2}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}})\boldsymbol{\Sigma}_*^{-1/2}\right)}{\lambda_{\min}\left(\boldsymbol{\Sigma}_*^{-1/2}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}'})\boldsymbol{\Sigma}_*^{-1/2}\right) \cdot \lambda_{\min}\left(\boldsymbol{\Sigma}_*^{-1/2}(\boldsymbol{\Sigma}_* + \boldsymbol{\Gamma}_{\mathbf{w}})\boldsymbol{\Sigma}_*^{-1/2}\right)} \\
&\leq \frac{2\left\|\boldsymbol{\Gamma}_{\mathbf{w}} - \boldsymbol{\Gamma}_{\mathbf{w}'}\right\|_2 \cdot \left(1 + \frac{\mu^+}{\sigma_*^-}\|\mathbf{w}\|_2^2\right)}{\sigma_*^-\left(1 + \frac{\mu^-}{\sigma_*^+}\|\mathbf{w}'\|_2^2\right) \cdot \left(1 + \frac{\mu^-}{\sigma_*^+}\|\mathbf{w}\|_2^2\right)} \\
&\leq \frac{4}{\sigma_*^-}\left\|\mathbb{E}\left[\mathbf{X}\mathbf{w}\mathbf{w}^T\mathbf{X}^T\right] - \mathbb{E}\left[\mathbf{X}\mathbf{w}'\mathbf{w}'^T\mathbf{X}^T\right]\right\|_2 \\
&\leq \frac{4}{\sigma_*^-} \cdot \sup_{\mathbf{v} \in \mathbb{S}^{m-1}} \left|\mathbf{v}^T\left(\mathbb{E}\left[\mathbf{X}\mathbf{w}\mathbf{w}^T\mathbf{X}^T\right] - \mathbb{E}\left[\mathbf{X}\mathbf{w}'\mathbf{w}'^T\mathbf{X}^T\right]\right)\mathbf{v}\right| \\
&\leq \frac{4}{\sigma_*^-}\left(\sup_{\mathbf{v} \in \mathbb{S}^{m-1}} \left|\mathbf{v}^T\mathbb{E}\left[\mathbf{X}\mathbf{w}(\mathbf{w} - \mathbf{w}')^T\mathbf{X}^T\right]\mathbf{v}\right| + \sup_{\mathbf{v} \in \mathbb{S}^{m-1}} \left|\mathbf{v}^T\mathbb{E}\left[\mathbf{X}(\mathbf{w} - \mathbf{w}')\mathbf{w}'^T\mathbf{X}^T\right]\mathbf{v}\right|\right) \\
&\leq \frac{8}{\sigma_*^-} \cdot \left\|\mathbf{w} - \mathbf{w}'\right\|_2 \cdot \sup_{\mathbf{v} \in \mathbb{S}^{m-1}} \sup_{\mathbf{z} \in \mathbb{S}^{p-1}} \sup_{\mathbf{r} \in \mathcal{S}'} \mathbf{v}^T\mathbb{E}\left[\mathbf{X}\mathbf{r}\mathbf{z}^T\mathbf{X}^T\right]\mathbf{v}
\end{aligned}
$$

$$\leq \frac{8e_0}{\sigma_*^-} \cdot \left\|\mathbf{w} - \mathbf{w}'\right\|_2 \cdot \sup_{\mathbf{v}\in\mathbb{S}^{m-1}} \sup_{\mathbf{z}\in\mathbb{S}^{p-1}} \sup_{\mathbf{r}\in\mathcal{S}'} \frac{\mathbb{E}\left(\frac{\mathbf{v}^T\mathbf{X}\mathbf{r}}{e_0}\right)^2 + \mathbb{E}\left(\mathbf{v}^T\mathbf{X}\mathbf{z}\right)^2}{2}$$
$$\leq \frac{8e_0}{\sigma_*^-} \cdot \left\|\mathbf{w} - \mathbf{w}'\right\|_2 \cdot \mu^+ = \frac{8}{e_0} \cdot \left\|\mathbf{w} - \mathbf{w}'\right\|_2$$

By invoking Theorem 1, we have for ϕ_n with any fixed $\{\tilde{\boldsymbol{\eta}}_i\} \in \mathcal{E}$,

$$\phi_n = \frac{2}{e_0\sqrt{n}} \cdot \sup_{\mathbf{t}\in\mathcal{T}} Z_{\mathbf{t}} \leq \frac{2}{e_0\sqrt{n}} \cdot \sup_{\mathbf{t},\mathbf{t}'\in\mathcal{T}} |Z_{\mathbf{t}} - Z_{\mathbf{t}'}|$$
$$\leq \frac{2c_3'}{e_0} \cdot \frac{\kappa\sqrt{\mu^+} \cdot w(\mathcal{T})}{\sqrt{n}} = 4c_3' \cdot \frac{\kappa\sqrt{\mu^+} \cdot w(\mathcal{S})}{\sqrt{n}}$$

with probability at least $1 - c_4' \exp\left(-\frac{w^2(\mathcal{T})}{\operatorname{diam}^2(\mathcal{T})}\right) \geq 1 - c_4' \exp\left(-\frac{w^2(\mathcal{S})}{2}\right)$. Now we combine the randomness of $\mathbf{X}_i$ and $\tilde{\boldsymbol{\eta}}_i$, and get

$$\mathbb{P}_{\mathbf{X},\tilde{\boldsymbol{\eta}}}\left(\phi_n \leq 4c_3' \cdot \frac{\kappa\sqrt{\mu^+} \cdot w(\mathcal{S})}{\sqrt{n}}\right)$$
$$= \int \mathbb{P}_{\mathbf{X}}\left(\phi_n \leq 4c_3' \cdot \frac{\kappa\sqrt{\mu^+} \cdot w(\mathcal{S})}{\sqrt{n}} \;\middle|\; \{\tilde{\boldsymbol{\eta}}_i\}\right) p\left(\tilde{\boldsymbol{\eta}}_1, \ldots, \tilde{\boldsymbol{\eta}}_n\right) d\tilde{\boldsymbol{\eta}}_1 \ldots d\tilde{\boldsymbol{\eta}}_n$$
$$\geq \int_{\mathcal{E}} \mathbb{P}_{\mathbf{X}}\left(\phi_n \leq 4c_3' \cdot \frac{\kappa\sqrt{\mu^+} \cdot w(\mathcal{S})}{\sqrt{n}} \;\middle|\; \{\tilde{\boldsymbol{\eta}}_i\}\right) p\left(\tilde{\boldsymbol{\eta}}_1, \ldots, \tilde{\boldsymbol{\eta}}_n\right) d\tilde{\boldsymbol{\eta}}_1 \ldots d\tilde{\boldsymbol{\eta}}_n$$
$$\geq \left(1 - c_4' \exp\left(-\frac{w^2(\mathcal{S})}{2}\right)\right) \cdot \mathbb{P}\left(\mathcal{E}\right)$$
$$\geq \left(1 - c_4' \exp\left(-\frac{w^2(\mathcal{S})}{2}\right)\right)\left(1 - 2\exp\left(-c_1' m\right)\right)$$
$$\geq 1 - 2\exp\left(-c_1' m\right) - c_4' \exp\left(-\frac{w^2(\mathcal{S})}{2}\right)$$

We obtain the final bound by assembling everything above. If $n \geq \max\left\{n_0, C_0' \cdot \max\left\{\tau^4, 1\right\} \cdot \max\left\{w^2(\mathcal{C}), m\right\}\right\}$, with probability at least $1-\epsilon-C_1' \exp\left(-C_2' \min\{w^2(\mathcal{S}), m\}\right)$, we have

$$\begin{aligned}
\beta_n &\leq \sqrt{m}\gamma_n\zeta_n + \phi_n \\
&\leq C_3' \cdot \max\{\tau^2, \kappa^2\} \cdot \frac{\kappa\sqrt{\mu^+}(m + w(\mathcal{C}))(\sqrt{m} + w(\mathcal{C}))}{n} + C_4' \cdot \frac{\kappa\sqrt{\mu^+} \cdot w(\mathcal{S})}{\sqrt{n}} \ ,
\end{aligned}$$

In particular, if the sample size also satisfies $n \geq C_5' \cdot \max\left\{\tau^4, \kappa^4\right\} \cdot \max\left\{\frac{m^3}{w^2(\mathcal{C})}, m^2, w^2(\mathcal{C})\right\} \geq C_6' \cdot \max\left\{\tau^4, \kappa^4\right\} \cdot \left(\frac{(m+w(\mathcal{C}))(\sqrt{m}+w(\mathcal{C}))}{w(\mathcal{S})}\right)^2$, we further have

$$\beta_n \leq C_7' \cdot \frac{\kappa\sqrt{\mu^+} \cdot w(\mathcal{S})}{\sqrt{n}} \ ,$$

which completes the proof. ■

Chapter 10

Conclusions

In this thesis, we present our research on both computational and statistical aspects of some high-dimensional estimation problems, with a focus on general structures. The problems we consider have covered a set of models that are widely used in practice, from vector to matrix, linear to nonlinear, parametric to semi-parametric, and convex to non-convex. The main contributions of this thesis are two-fold. On one hand, the thesis establishes computational frameworks for estimating the model parameters in those problems, which are allowed to possess diverse structures. On the other hand, it also provides unified views into the corresponding statistical guarantees. At the heart of the statistical analyses are the geometric measures, which can tersely characterize the recovery error of the estimators.

In Chapter 3, we start with the estimation of high-dimensional linear models, and propose the generalized Dantzig selector (GDS) to incorporate the structure information of the parameter. With an ADMM-type optimization algorithm, we can efficiently compute the GDS, whose statistical error is later shown to be conveniently bounded by certain geometric measures, such as Gaussian width and restricted norm compatibility.

In Chapter 4, we are committed to a comprehensive study of the geometric measures introduced in Chapter 3. For a broad class of structures that can be captured by *atomic norms*, we can further bound the geometric measures using simple information of the structure, which are believed to facilitate the statistical analysis for new structures.

In Chapter 5, we extend the GDS to the matrix setting, which yields similar type of results as obtained in Chapter 3 and 4. The general bounds derived in this chapter apply to a broad family of matrix structures, which can be encoded by the unitarily invariant norms.

In Chapter 6, we move from the parametric linear model to a semi-parametric nonlinear extension, which is the single-index model (SIM). Based on U-statistics, we propose two simple estimators for the model parameter estimation, which are robust to heavy-tailed noise. The statistical guarantees of both estimators are built on similar geometric measures as in Chapter 3. Instantiated for both one-bit compressed sensing and the monotone transfer setting, the estimators lead to novel algorithms with provably guarantees.

In Chapter 7, we continue to focus on semi-parametric models. Specifically we propose a new model called sparse linear isotonic model (SLIM), in order to introduce nonlinearity in the features of sparse linear models. The model is parameterized by a vector (as in linear models) as well as a set of unknown monotone functions applied on features. Computationally a two-step algorithm is designed for the sequential estimation of the sparse parameter and the monotone functions, which avoids the specification of the monotonicity compared with other related models. Statistically we show that the algorithm can recover the parameter with provably small error.

In Chapter 8, we switch our attention to non-convex problems. We propose an alternating estimation (AltEst) procedure for solving the structured multi-response linear models in high dimension. The procedure uses the GDS in the estimation of the parameter vector, and its statistical guarantee is determined by the geometric measures under the idealized resampling assumption. By leveraging the noise correlation among responses, AltEst can achieve significantly smaller estimation error than ignoring the noise structure.

Lastly, in Chapter 9, we present several extensions to the results in Chapter 8. With the GDS substituted by the constrained estimator in AltEst framework, we allow non-convex characterizations of the parameter structure. In the statistical analysis, we are able to relax the Gaussian assumption imposed on the noise, and show the recovery guarantee without the resampling assumption. The theoretical result yields a new discovery that randomly-initialized AltEst could also have great statistical performance, which is confirmed by the empirical study.

References

[1] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon. Learning sparsely used overcomplete dictionaries via alternating minimization. *CoRR*, abs/1310.7991, 2013.

[2] A. Ahmed and E. P Xing. Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences*, 106(29):11878–11883, 2009.

[3] P. Alquier and G. Biau. Sparse single-index model. *Journal of Machine Learning Research*, 14:243–280, 2013.

[4] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: Phase transitions in convex programs with random data. *Inform. Inference*, 3(3):224–294, 2014.

[5] T. W. Anderson. *An introduction to multivariate statistical analysis.* Wiley-Interscience, 2003.

[6] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[7] A. Argyriou, R. Foygel, and N. Srebro. Sparse prediction with the k-support norm. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[8] S. Arora, R. Ge, R. Kannan, and A. Moitra. Computing a nonnegative matrix factorization–provably. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 145–162. ACM, 2012.

[9] P. Bacchetti. Additive isotonic model. *Journal of the American Statistical Association*, 84(405):289–294, 1989.

[10] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Convex optimization with sparsity-inducing norms. *Optimization for Machine Learning*, 5, 2011.

[11] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trendss® in Machine Learning*, 4(1), 2012.

[12] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Structured sparsity through convex optimization. *Statist. Sci.*, 27(4), 2012.

[13] F. R. Bach. Structured sparsity-inducing norms through submodular functions. In *Advances in Neural Information Processing Systems*, pages 118–126, 2010.

[14] A. Banerjee, S. Chen, F. Fazayeli, and V. Sivakumar. Estimation with norm regularization. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[15] R. Barber and W. Ha. Gradient descent with nonconvex constraints: local concavity determines convergence. *arXiv preprint arXiv:1703.07755*, 2017.

[16] R.E. Barlow. *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression.* J. Wiley, 1972.

[17] A. Beck and L. Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization*, 23(4):2037–2060, 2013.

[18] R. Bellman. On the approximation of curves by line segments using dynamic programming. *Communications of the ACM*, 4(6):284, 1961.

[19] R. Bhatia. *Matrix Analysis*. Springer, 1997.

[20] S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016.

[21] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[22] M. Bogdan, E. van den Berg, W. Su, and E. Candes. Statistical estimation and testing via the sorted L1 norm. *arXiv:1310.1969*, 2013.

[23] P. T Boufounos and R. G Baraniuk. 1-bit compressive sensing. In *Information Sciences and Systems, 2008. CISS 2008. 42nd Annual Conference on*, 2008.

[24] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[25] L. Breiman and J. H. Friedman. Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(1):3–54, 1997.

[26] S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[27] P. Buhlmann and S. van de Geer. *Statistics for High Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer, 2011.

[28] T. T. Cai, T. Liang, and A. Rakhlin. Geometrizing local rates of convergence for high-dimensional linear inverse problems. *arXiv:1404.4408*, 2014.

[29] T. T. Cai, W. Liu, and X. Luo. A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

[30] T. T. Cai and A. Zhang. Rop: Matrix recovery via rank-one projections. *The Annals of Statistics*, 43(1):102–138, 2015.

[31] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2012.

[32] E. Candes and T Tao. The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

[33] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):11:1–11:37, 2011.

[34] E. J Candes, X. Li, and M. Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[35] E. J. Candès and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

[36] E. J. Candès and B. Recht. Simple bounds for recovering low-complexity models. *Math. Program.*, 141(1-2):577–589, 2013.

[37] E. J Candes, T. Strohmer, and V. Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[38] R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[39] G. Casella and R. L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[40] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

[41] S. Chatterjee, S. Chen, and A. Banerjee. Generalized dantzig selector: Application to the k-support norm. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[42] S. Chen and A. Banerjee. One-bit compressed sensing with the k-support norm. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

[43] S. Chen and A. Banerjee. Structured estimation with atomic norms: General bounds and applications. In *Advances in Neural Information Processing Systems*, 2015.

[44] S. Chen and A. Banerjee. Structured matrix recovery via the generalized dantzig selector. In *Advances in Neural Information Processing Systems*, 2016.

[45] S. Chen and A. Banerjee. Alternating estimation for structured high-dimensional multi-response models. In *Advances in Neural Information Processing Systems*, pages 2835–2844, 2017.

[46] S. Chen and A. Banerjee. Robust structured estimation with single-index models. In *Proceedings of the 34th International Conference on Machine Learning*, pages 712–721, 2017.

[47] S. Chen and A. Banerjee. Sparse linear isotonic models. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, pages 1270–1279, 2018.

[48] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Rev.*, 43(1):129–159, 2001.

[49] G. Cheng. Semiparametric additive isotonic regression. *J. Stat. Plan. Inference*, 139(6):1980–1991, 2009.

[50] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990.

[51] T. Evgeniou and M. Pontil. Regularized multi–task learning. In *KDD*, pages 109–117, 2004.

[52] K. Fang, S. Kotz, and K. Ng. *Symmetric multivariate and related distributions.* Number 36 in Monographs on statistics and applied probability. Chapman & Hall, 1990.

[53] Z. Fang and N. Meinshausen. LASSO isotone for high-dimensional additive isotonic regression. *Journal of Computational and Graphical Statistics*, 21(1):72–91, 2012.

[54] M. A. T. Figueiredo and R. D. Nowak. Sparse estimation with strongly correlated variables using ordered weighted l1 regularization. *arXiv:1409.4005*, 2014.

[55] M. A. T. Figueiredo and R. D. Nowak. Ordered weighted l1 regularized regression with strongly correlated covariates: Theoretical aspects. In *AISTATS*, 2016.

[56] S. Foucart. Hard thresholding pursuit: an algorithm for compressive sensing. *SIAM Journal on Numerical Analysis*, 49(6):2543–2563, 2011.

[57] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[58] R. Ganti, N. Rao, R. M Willett, and R. Nowak. Learning single index models in high dimensions. *arXiv preprint arXiv:1506.08910*, 2015.

[59] R. Ge, C. Jin, and Y. Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *arXiv preprint arXiv:1704.00708*, 2017.

[60] R. Ge, J. D Lee, and T. Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

[61] A. R. Goncalves, P. Das, S. Chatterjee, V. Sivakumar, F. J. Von Zuben, and A. Banerjee. Multi-task sparse structure learning. In *CIKM*, pages 451–460, 2014.

[62] S. Gopi, P. Netrapalli, P. Jain, and A. Nori. One-bit compressed sensing: Provable support and vector recovery. In *Proceedings of The 30th International Conference on Machine Learning*, 2013.

[63] Y. Gordon. Some inequalities for gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289, 1985.

[64] W. H. Greene. *Econometric Analysis.* Prentice Hall, 7. edition, 2011.

[65] S. J. Grotzinger and C. Witzgall. Projections onto order simplexes. *Applied Mathematics and Optimization*, 12(1):247–270, 1984.

[66] Q. Gu and A. Banerjee. High dimensional structured superposition models. In *Advances in Neural Information Processing Systems*, pages 3684–3692, 2016.

[67] S. Gunasekar, A. Banerjee, and J. Ghosh. Unified view of matrix completion under general structural constraints. In *NIPS*, pages 1180–1188, 2015.

[68] S. Gunasekar, P. Ravikumar, and J. Ghosh. Exponential family matrix completion under structural constraints. In *International Conference on Machine Learning (ICML)*, 2014.

[69] F. Han and H. Liu. Optimal rates of convergence for latent generalized correlation matrix estimation in transelliptical distribution. *arXiv:1305.6916*, 2013.

[70] F. Han and H. Liu. Scale-invariant sparse PCA on high-dimensional meta-elliptical data. *Journal of the American Statistical Association*, 109(505):275–287, 2014.

[71] T. J. Hastie and R. J. Tibshirani. *Generalized additive models*. London: Chapman & Hall, 1990.

[72] W. Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[73] J. L. Horowitz and W. Hardle. Direct semiparametric estimation of single-index models with discrete covariates. *Journal of the American Statistical Association*, 91(436):1632–1640, 1996.

[74] J. L. Horowitz and E. Mammen. Nonparametric estimation of an additive model with a link function. *Ann. Statist.*, 32(6):2412–2443, 2004.

[75] J. Huang and T. Zhang. The benefit of group sparsity. *arXiv preprint arXiv:0901.2962*, 2009.

[76] J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. *Journal of Machine Learning Research*, 12:3371–3412, 2011.

[77] H. Ichimura. Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58:71–120, 1993.

[78] A. J. Izenman. Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 5(2):248–264, 1975.

[79] A. J. Izenman. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer, 2008.

[80] L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *International Conference on Machine Learning (ICML)*, 2009.

[81] L. Jacques, J. N Laska, P. T Boufounos, and R. G Baraniuk. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *IEEE Transactions on Information Theory*, 59(4):2082–2102, 2013.

[82] P. Jain and P. Kar. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–336, 2017.

[83] P. Jain, R. Meka, and I. Dhillon. Guaranteed rank minimization via singular value projection. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[84] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *STOC*, pages 665–674, 2013.

[85] P. Jain and A. Tewari. Alternating minimization for regression problems with vector-valued outputs. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1126–1134, 2015.

[86] P. Jain, A. Tewari, and P. Kar. On iterative hard thresholding methods for high-dimensional m-estimation. In *NIPS*, pages 685–693, 2014.

[87] A. Jalali, P. Ravikumar, S. Sanghavi, and C. Ruan. A dirty model for multi-task learning. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, pages 964–972, 2010.

[88] A. Jalali, S. Sanghavi, C. Ruan, and P. K. Ravikumar. A dirty model for multi-task learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 964–972, 2010.

[89] G. M. James and P. Radchenko. A generalized dantzig selector with shrinkage tuning. *Biometrika*, 96(2):323–337, 2009.

[90] G. M. James, P. Radchenko, and J. Lv. Dasso: connections between the dantzig selector and lasso. *Journal of the Royal Statistical Society Series B*, 71(1):127–142, 2009.

[91] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for hierarchical sparse coding. *J. Mach. Learn. Res.*, 12:2297–2334, 2011.

[92] S. M Kakade, V. Kanade, O. Shamir, and A. Kalai. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems*, 2011.

[93] A. T Kalai and R. Sastry. The isotron algorithm: High-dimensional isotonic regression. In *COLT*, 2009.

[94] M.G. Kendall. *Rank correlation methods*. C. Griffin, 1948.

[95] S. Kim and E. P. Xing. Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eqtl mapping. *Ann. Appl. Stat.*, 6(3):1095–1117, 2012.

[96] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[97] C. L Lawson and R. J Hanson. *Solving least squares problems*, volume 15. Siam, 1995.

[98] A. J. Lee. *U-Statistics: Theory and Practice*. Taylor & Francis, 1990.

[99] S. Lee, D. Bryzski, and M. Bogdan. Fast Saddle-Point Algorithm for Generalized Dantzig Selector and FDR Control with the Ordered l1-Norm. In *AISTATS*, 2016.

[100] W. Lee and Y. Liu. Simultaneous multiple response regression and inverse covariance matrix estimation via penalized gaussian maximum likelihood. *J. Multivar. Anal.*, 111:241–255, 2012.

[101] E.L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer Verlag, 1998.

[102] A. S. Lewis. The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2(1-2):173–183, 1995.

[103] B. Li and S. CH Hoi. Online portfolio selection: A survey. *ACM Computing Surveys (CSUR)*, 46(3):35, 2014.

[104] P. Li. One scan 1-bit compressed sensing. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 2016.

[105] Q. Li and G. Tang. The nonconvex geometry of low-rank matrix optimizations with general objective functions. *arXiv preprint arXiv:1611.03060*, 2016.

[106] Y. Lin and H. Zhang. Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.*, 34(5):2272–2297, 2006.

[107] F. Lindskog, A. McNeil, and U. Schmock. Kendalls tau for elliptical distributions. In *Credit Risk*, Contributions to Economics, pages 149–156. Physica-Verlag HD, 2003.

[108] H. Liu, F. Han, M. Yuan, J. Lafferty, and L. Wasserman. High-dimensional semiparametric gaussian copula graphical models. *Ann. Statist.*, 40(4):2293–2326, 2012.

[109] H. Liu, F. Han, and C. H. Zhang. Transelliptical graphical models. In *NIPS*, pages 809–817, 2012.

[110] H. Liu, J. Lafferty, and L. Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.*, 10:2295–2328, 2009.

[111] H. Liu, M. Palatucci, and J. Zhang. Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In *ICML*, pages 649–656, 2009.

[112] H. Liu, J. Zhang, X. Jiang, and J. Liu. The group dantzig selector. In *AISTATS*, 2010.

[113] J. Liu, L. Yuan, and J. Ye. Guaranteed sparse recovery under linear transformation. In *International Conference on Machine Learning*, pages 91–99, 2013.

[114] Z. Lu, T. K. Pong, and Y. Zhang. An alternating direction method for finding dantzig selectors. *Computational Statistics & Data Analysis*, 56(12):4037 – 4046, 2012.

[115] Z. Q. Luo and P. Tseng. On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, 72(1):7–35, 1992.

[116] C. Ma, K. Wang, Y. Chi, and Y. Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution. *arXiv preprint arXiv:1711.10467*, 2017.

[117] E. Mammen and K. Yu. Additive isotone regression. *IMS Lecture Notes-Monograph Series Asymptotics: Particles, Processes and Inverse Problems*, 55:179–195, 2007.

[118] A. Maurer, M. Pontil, and B. Romera-Paredes. An inequality with applications to structured sparsity and multitask dictionary learning. In *Conference on Learning Theory (COLT)*, 2014.

[119] P. McCullagh. Generalized linear models. *European Journal of Operational Research*, 16(3):285–292, 1984.

[120] A. M. McDonald, M. Pontil, and D. Stamos. New perspectives on k-support and cluster norms. *ArXiv e-prints*, 2014.

[121] A. M. McDonald, M. Pontil, and D. Stamos. Spectral k-support norm regularization. In *NIPS*, 2014.

[122] L. Meier, S. Van de Geer, and P. Bhlmann. High-dimensional additive modeling. *Ann. Statist.*, 37(6B):3779–3821, 2009.

[123] N. Meinshausen and P. Bhlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 2006.

[124] S. Mendelson. Upper bounds on product and multiplier empirical processes. *S-tochastic Processes and their Applications*, 126(12):3652 – 3680, 2016.

[125] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Reconstruction and sub-Gaussian operators in asymptotic geometric analysis. *Geometric and Functional Analysis*, 17:1248–1282, 2007.

[126] D. Needell and J. A Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.

[127] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for the analysis of regularized M-estimators. *Statistical Science*, 27(4):538–557, 2012.

[128] P. Netrapalli, P. Jain, and S. Sanghavi. Phase retrieval using alternating minimization. In *NIPS*, 2013.

[129] M. Neykov, J. S. Liu, and T. Cai. L1-regularized least squares for support recovery of high dimensional single index models with gaussian designs. *J. Mach. Learn. Res.*, 17(1):2976–3012, 2016.

[130] S. Oymak, B. Recht, and M. Soltanolkotabi. Sharp time–data tradeoffs for linear inverse problems. *arXiv preprint arXiv:1507.04793*, 2015.

[131] S. Oymak and M. Soltanolkotabi. Fast and reliable parameter estimation from nonlinear observations. *arXiv preprint arXiv:1610.07108*, 2016.

[132] S. Oymak, C. Thrampoulidis, and B. Hassibi. The squared-error of generalized Lasso: A precise analysis. *arXiv:1311.0830*, 2013.

[133] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3), 2014.

[134] Y. Plan and R. Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, 2013.

[135] Y. Plan, R. Vershynin, and E. Yudovina. High-dimensional estimation with geometric constraints. *Information and Inference*, 2016.

[136] P. Radchenko. High dimensional single index models. *Journal of Multivariate Analysis*, 139:266–282, 2015.

[137] P. Rai, A. Kumar, and H. Daume. Simultaneously leveraging output and task structures for multiple-output regression. In *NIPS*, pages 3185–3193, 2012.

[138] N. Rao, B. Recht, and R. Nowak. Universal measurement bounds for structured sparse signal recovery. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

[139] G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010.

[140] P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.

[141] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[142] F. Ricci, L. Rokach, and B. Shapira. Introduction to recommender systems handbook. In *Recommender systems handbook*, pages 1–35. Springer, 2011.

[143] E. Richard, P. Savalle, and N. Vayatis. Estimation of simultaneously sparse and low rank matrices. In *ICML*, 2012.

[144] R. T. Rockafellar. *Convex Analysis*. Princetion University Press, 1970.

[145] A. J. Rothman, E. Levina, and J. Zhu. Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, 19(4):947–962, 2010.

[146] M. Rudelson and R. Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians*, pages 1576–1602, 2010.

[147] P. Sajda. Machine learning for detection and diagnosis of disease. *Annu. Rev. Biomed. Eng.*, 8:537–565, 2006.

[148] J. Shen and P. Li. On the iteration complexity of support recovery via hard thresholding pursuit. In *International Conference on Machine Learning*, pages 3115–3124, 2017.

[149] D. Silver, A. Huang, C. J Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[150] V. Sivakumar, A. Banerjee, and P. Ravikumar. Beyond sub-gaussian measurements: High-dimensional structured estimation with sub-exponential designs. In *NIPS*, pages 2206–2214, 2015.

[151] M. Slawski and P. Li. b-bit marginal regression. In *Advances in Neural Information Processing Systems*, 2015.

[152] M. Slawski and P. Li. Linear signal recovery from b-bit-quantized linear measurements: precise analysis of the trade-off between bit depth and number of measurements. *arXiv preprint arXiv:1607.02649*, 2016.

[153] K.-A. Sohn and S. Kim. Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. In *AISTATS*, pages 1081–1089, 2012.

[154] M. Soltanolkotabi. Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization. *arXiv preprint arXiv:1702.06175*, 2017.

[155] I. Stewart. *Galois Theory, Third Edition.* Chapman Hall/CRC Mathematics Series. Taylor & Francis, 2003.

[156] J. Sun, Q. Qu, and J. Wright. A geometric analysis of phase retrieval. *arXiv preprint arXiv:1602.06664*, 2016.

[157] J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere i: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2017.

[158] R. Sun and M. Hong. Improved iteration complexity bounds of cyclic block coordinate descent for convex problems. *arXiv:1512.04680*, 2015.

[159] R. Sun and Z.-Q. Luo. Guaranteed matrix completion via nonconvex factorization. In *FOCS*, 2015.

[160] M. Talagrand. A simple proof of the majorizing measure theorem. *Geometric & Functional Analysis GAFA*, 2(1):118–125, 1992.

[161] M. Talagrand. *The Generic Chaining.* Springer, 2005.

[162] M. Talagrand. *Upper and Lower Bounds for Stochastic Processes.* Springer, 2014.

[163] W. M Thorburn. The myth of occam's razor. *Mind*, 27(107):345–353, 1918.

[164] Z. Tian, H. Zhang, and R. Kuang. Sparse group selection on fused lasso components for identifying group-specific dna copy number variations. In *IEEE 12th International Conference on Data Mining*, pages 665–674, 2012.

[165] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

[166] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

[167] J. A Tropp and A. C Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on information theory*, 53(12):4655–4666, 2007.

[168] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001.

[169] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via procrustes flow. *arXiv preprint arXiv:1507.03566*, 2015.

[170] A W Van der Vaart. *Asymptotic statistics*. Cambridge university press, 1998.

[171] V. Vapnik. *Statistical learning theory*. Wiley New York, 1998.

[172] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok, editors, *Compressed Sensing*, chapter 5, pages 210–268. Cambridge University Press, 2012.

[173] R. Vershynin. *Estimation in High Dimensions: A Geometric Perspective*, pages 3–66. Springer International Publishing, 2015.

[174] M. J. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming(Lasso). *IEEE Transactions on Information Theory*, 55:2183–2202, 2009.

[175] H. Wang and A. Banerjee. Bregman alternating direction method of multipliers. In *NIPS*, 2014.

[176] X. Wang and X. Yuan. The linearized alternating direction method of multipliers for dantzig selector. *SIAM J. Scientific Computing*, 34(5), 2012.

[177] M. Wytock and Z. Kolter. Sparse gaussian conditional random fields: Algorithms, theory, and application to energy forecasting. In *International conference on machine learning*, pages 1265–1273, 2013.

[178] L. Xue and H. Zou. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann. Statist.*, 40(5):2541–2571, 2012.

[179] X. Yi and C. Caramanis. Regularized EM algorithms: A unified framework and statistical guarantees. In *NIPS*, pages 1567–1575, 2015.

[180] X. Yi, C. Caramanis, and S. Sanghavi. Alternating minimization for mixed linear regression. In *ICML*, pages 613–621, 2014.

[181] X. Yi, Z. Wang, C. Caramanis, and H. Liu. Optimal linear estimation under unknown nonlinear transform. In *Advances in Neural Information Processing Systems*, 2015.

[182] M. Yuan. High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11:2261–2286, 2010.

[183] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006.

[184] X.-T. Yuan and T. Zhang. Partial gaussian graphical model estimation. *IEEE Transactions on Information Theory*, 60:1673–1687, 2014.

[185] X. Zeng and M. A. T. Figueiredo. The ordered weighted ℓ_1 norm: atomic formulation, projections, and algorithms. *arXiv:1409.4271*, 2014.

[186] L. Zhang, J. Yi, and R. Jin. Efficient algorithms for robust one-bit compressive sensing. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014.

[187] X. Zhang, Y. Yu, and D. Schuurmans. Polar operators for structured sparse estimation. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[188] P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2567, November 2006.

[189] Q. Zheng and J. Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. In *Advances in Neural Information Processing Systems*, pages 109–117, 2015.

[190] S. Zhou. Restricted eigenvalue conditions on subgaussian random matrices. Technical report, Department of Mathematics, ETH Zurich, December 2009.

[191] R. Zhu and Q. Gu. Towards a lower sample complexity for robust one-bit compressed sensing. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

[192] O. Zuk and A. Wagner. Low-rank matrix recovery from row-and-column affine measurements. In *International Conference on Machine Learning (ICML)*, 2015.