II. HIGH DIMENSIONAL ESTIMATION UNDER WEAK MOMENT ASSUMPTIONS:
STRUCTURED RECOVERY AND MATRIX ESTIMATION

by

Xiaohan Wei

Presented to the
FACULTY OF THE USC GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(ELECTRICAL ENGINEERING)

Dec 2019

Copyright 2019

Xiaohan Wei
Approved by

Professor Stanislav Minsker,
Committee Chair,
Department of Mathematics,
*University of Southern California.*

Professor Michael Neely,
Committee Chair,
Department of Electrical Engineering,
*University of Southern California.*

Professor Larry Goldstein,
Department of Mathematics,
*University of Southern California.*

Professor Mihailo Jovanovic,
Department of Electrical Engineering,
*University of Southern California.*

Professor Ashutosh Nayyar,
Department of Electrical Engineering,
*University of Southern California.*
Dedication

To my parents and my wife, Yuhong, who supported me both mentally and financially over the years.
Acknowledgements

First, I would like to thank my advisor professor Michael J. Neely for guiding me throughout the PhD journey since Summer 2013. He is a man of accuracy and rigorousness, always passionate about discussing concrete research problems, and willing to roll up the sleeves and grind through technical details with me. His way of treating research topics significantly impacts me. Rather than blindly following existing works and doing incremental works when trying to get into a new area, I learned to ask fundamental mathematical questions, making connections to the tools and theories we already familiar with and be not afraid of getting my hands dirty. His blazing new ideas are my morale boost when grasping in the dark.

Next, I would like to thank professor Stanislav Minsker, who is the advisor on my high-dimensional statistics research. I got to know him during the Math-547 statistical learning course Fall 2015. Though not much senior than me, he is already extremely knowledgable on the statistical learning area and has been widely recognized for his works on robust high-dimensional statistics. He is a quick thinker and can always point out meaningful new directions hiding rather deeply which eventually lead to high-quality publications. I would have published no paper on this area should I never met with him. Along the way, he also teaches me how to sell my works and helps me practicing my seminar talks, which lead to impressive presentations and Ming-Hsieh scholarships.

Also, I would like to thank professor Larry Goldstein, whom I met during a small paper reading group Spring 2016. He is an expert on Stein’s method and, as a senior professor, surprisingly accessible to PhD students and active on various research areas. Together with Prof. Minsker, we had quite a few fruitful discussions and made some nice progress on robust statistics.

I would also like to thank professor Mihailo Jovanovic, Ashutosh Nayyar for discussing research problems with me and siting on my qualifying exam committee. I appreciate them for their valuable comments and suggestions.
Moreover, I thank my senior lab mates Hao Yu and Sucha Supittayapornpong who were always accessible to discussing problems with me and came up with new research ideas. Also, Ruda Zhang, Lang Wang, and Jie Ruan studied various math courses and interesting math problems with me and helped me clear up the hurdles on different stages, for which I really appreciate. Special thanks to professor Qing Ling, who was my undergraduate advisor, but continuously influences me on various aspects of my academic career.

Last but not least, I would like to take the chance to express my gratitude for folks who made contribution on various stages of my research. In particular, I thank Zhuoran Yang, for lighting up new areas and expanding my research horizon, Dongsheng Ding, who brings idea from control perspective and is always passionate to try out research ideas with me, Sheng Chen for sharing with me his perspective on robust LASSO problems, professor Jason D. Lee for working on the geometric median problem with me, and Jianshu Chen from Tencent AI who introduced me to the area of reinforcement learning.
# Table of Contents

Dedication iii  
Acknowledgements iv  
Abstract viii  

1 Introduction and a Heavy-tailed Framework  
1.1 Background 1  
1.1.1 From least square to supremum of an empirical process 1  
1.1.2 Supremum of an empirical process: binary functions 3  
1.1.3 Supremum of an empirical process: General cases 7  
1.1.4 Other key inequalities 12  
1.1.5 Gordon’s theorem and bounds on the estimation error 14  
1.1.6 Theorem 1.1.10 is restrictive 16  
1.2 Small-ball Method 18  
1.2.1 A general theorem 18  
1.2.2 Application to least squares ERM 19  
1.3 Organization of the Thesis 21  

2 Optimal Statistical Rate in Generalized Linear Models under Weak Moment Assumptions  
2.1 Introduction 23  
2.1.1 Related works 26  
2.2 Main Results 27  
2.2.1 Optimal Estimation in $\ell_1$-ball 27  
2.2.2 Optimal Estimation of Bounded Sparse Vectors 29  
2.3 Proof of Theorems: A Heavy-tailed Framework 31  
2.4 Proof of Theorem 2.2.2 Computing Local Complexities 38  
2.4.1 Bounding $r_Q$: Preliminary estimates 38  
2.4.2 Weak small-ball estimates for small $N$ 42  
2.4.3 Applying Mendelson’s small-ball method for large $N$ 49  
2.4.4 Bounding $r_M$ via Montgomery-Smith inequality 51  
2.4.5 Bounding the radius $r_V$ 61  
2.4.6 Putting everything together 66  
2.5 Proof of Theorem 2.2.5 Computing Local Complexities 67  
2.5.1 Bounding radius $r_Q$ 67  
2.5.2 Bounding $r_M$ and $r_V$ 69  
2.5.3 Putting everything together 71
3 Structured Recovery from Non-linear and Heavy-tailed Measurements

3.1 Introduction ......................................................... 73
3.2 Definitions and Background Material. ................................ 76
    3.2.1 Elliptically symmetric distributions. .......................... 76
    3.2.2 Geometry ..................................................... 78
3.3 Main Results ....................................................... 79
    3.3.1 Description of the proposed estimator ......................... 79
    3.3.2 Estimator performance guarantees ............................ 82
    3.3.3 Examples ..................................................... 84
3.4 Numerical Experiments ............................................ 85
3.5 Proofs .............................................................. 87
    3.5.1 Preliminaries ............................................... 87
    3.5.2 Roadmap of the proof of Theorem 3.3.1 ...................... 90
    3.5.3 Roadmap of the proof of Theorem 3.3.2 ...................... 92
    3.5.4 Bias of the truncated mean .................................. 93
    3.5.5 Concentration via generic chaining .......................... 96
3.6 Technical Results .................................................. 110
3.7 Decomposable Norms and Restricted Compatibility ................. 112

4 Estimation of the Covariance Structure of Heavy-tailed Distributions

4.1 Introduction ......................................................... 115
    4.1.1 Notation ...................................................... 116
    4.1.2 Problem formulation and overview of the existing work .... 117
4.2 Main Results ....................................................... 119
    4.2.1 Robust mean estimation ...................................... 120
    4.2.2 Robust covariance estimation ................................. 120
    4.2.3 Bounds in terms of intrinsic dimension ...................... 122
4.3 Applications: Low-rank Covariance Estimation ..................... 123
4.4 Proofs .............................................................. 124
    4.4.1 Proof of Lemma 4.2.1 ....................................... 124
    4.4.2 Proof of Lemma 4.4.1 ....................................... 125
    4.4.3 Proof of Theorem 4.2.1 ..................................... 127
    4.4.4 Proof of Theorem 4.3.1 ..................................... 127
4.5 Proof of Additional Technical Lemmas ................................ 128
    4.5.1 Preliminaries ............................................... 128
    4.5.2 Additional computation in the proof of Lemma 4.2.1 ......... 130
    4.5.3 Proof of Lemma 4.4.2 ....................................... 131
    4.5.4 Proof of Lemma 4.5.5 ....................................... 132
    4.5.5 Proof of Lemma 4.5.6 ....................................... 133
    4.5.6 Proof of Lemma 4.2.2 ....................................... 133
    4.5.7 Proof of Lemma 4.2.3 ....................................... 140
Abstract

The purpose of this thesis is to develop new theories on high-dimensional structured signal recovery under a rather weak assumption on the measurements that only a finite number of moments exists. High-dimensional recovery has been one of the emerging topics in the last decade partly due to the celebrated work of Candes, Romberg and Tao (e.g. [CRT06, CRT04]). The original analysis there (and the works thereafter) necessitates a strong concentration argument (namely, the restricted isometry property), which only holds for a rather restricted class of measurements with light-tailed distributions. It had long been conjectured that high-dimensional recovery is possible even if restricted isometry type conditions do not hold, but the general theory was beyond the grasp until very recently, when the works [Men14a, KM15] propose a new “small-ball method”. In these two papers, the authors initiated a new analysis framework for general empirical risk minimization (ERM) problems with respect to the square loss, which is “robust” and can potentially allow heavy-tailed loss functions. The materials in this thesis are partly inspired by [Men14a], but are of a different mindset: rather than directly analyzing the existing ERMs for signal recovery for which it is difficult to avoid strong moment assumptions, we show that, in many circumstances, by carefully re-designing the ERMs to start with, one can still achieve the minimax optimal statistical rate of signal recovery with very high probability under much weaker assumptions than existing works.
Chapter 1

Introduction and a Heavy-tailed Framework

The main focus of this thesis is to study robust recovery and estimation in the presence of heavy-tailed design or noises. In the analysis of regression models and matrix estimation procedures, it is common to assume that the data satisfy an certain model along with a set of assumptions such as i.i.d. observations from a Gaussian distribution. However, the data in practical world often violate such assumptions due to noise and outliers. One of the viable ways to model noisy data and outliers is to assume that the observations are generated by a heavy-tailed distribution\footnote{Throughout the thesis, a distribution is “heavy-tailed” if and only if finite number of moments exists.}. Therefore, the practical significance of this research is to relax the strong assumptions ubiquitous in previous high-dimensional recovery and estimation works, thereby reducing the gap between mathematical theories and the real world problems.

1.1 Background

1.1.1 From least square to supremum of an empirical process

Our main focus is the high-dimensional empirical risk minimization (ERM). We start by considering the classical least squares ERM, which is easy to understand and serves as a foundation for all subsequent development of this thesis. Let $\Theta$ be a measurable subset of $\mathbb{R}^d$, let $\mathbf{x} \in \mathbb{R}^d$ be a random vector, and let $y \in \mathbb{R}$ be a target response variable. One would like to find some vector $\theta^* \in \Theta$ so that $\langle \mathbf{x}, \theta^* \rangle$ and $y$ are as close as possible. A classical way of measuring the distance is to consider the square loss function $(\langle \mathbf{x}, \theta \rangle - y)^2$, and one hopes to select this $\theta^* \in \Theta$ so as to
minimize the expected loss:

\[
L(\theta) = \mathbb{E}((x, \theta) - y)^2 = \theta^T \mathbb{E}[xx^T] \theta - 2 \mathbb{E}[yx^T] \theta + \mathbb{E}[y^2].
\]

The term \( \mathbb{E}[y^2] \) is irrelevant in terms of minimization. However, it should be noted that in most cases, the expectations \( \mathbb{E}[xx^T] \) and \( \mathbb{E}[yx^T] \) are not known. Instead, we only have access to the i.i.d. samples \( \{x_i, y_i\}_{i=1}^N \) of \( \{x, y\} \). Thus, we instead aim to find \( \hat{\theta}_N \in \Theta \) minimizing the empirical loss:

\[
L_N(\theta) = \frac{1}{N} \sum_{i=1}^N \theta^T x_i x_i^T \theta - \frac{2}{N} \sum_{i=1}^N y_i x_i^T \theta \quad (1.1)
\]

It should also be noted that there are two aspects of this problem. One aspect is the estimation problem which aims to find some \( \theta_N \) so that \( \|\theta^* - \theta_N\|_2 \) is as small as possible. The other aspect is the prediction problem, namely, given an estimator \( \hat{\theta}_N \), we would like to know how it performs on future data compared to \( \theta^* \), i.e.

\[
\mathbb{E}\left[\left(\langle x, \hat{\theta}_N \rangle - y \right)^2 - (\langle x, \theta^* \rangle - y)^2 \mid \{x_i, y_i\}_{i=1}^N\right].
\]

This is also known as the “generalization error” of \( \hat{\theta}_N \). Throughout the thesis, we mainly focus on the estimation problem.

The classical way one analyzes the performance of (1.1) is as follows ([BBM+05]): since \( \hat{\theta}_N \in \Theta \) minimizes (1.1), it must satisfy:

\[
\frac{1}{N} \sum_{i=1}^N (\hat{\theta}_N - \theta^*)^T x_i x_i^T (\hat{\theta}_N - \theta^*) - \frac{2}{N} \sum_{i=1}^N y_i x_i^T (\hat{\theta}_N - \theta^*) \leq 0.
\]

Rearranging the terms gives:

\[
\frac{1}{N} \sum_{i=1}^N (\hat{\theta}_N - \theta^*)^T x_i x_i^T (\hat{\theta}_N - \theta^*) - \frac{2}{N} \sum_{i=1}^N (y_i - x_i^T \theta^*) x_i^T (\hat{\theta}_N - \theta^*) \leq 0.
\]

Thus, it follows that:

\[
\frac{1}{N} \sum_{i=1}^N (\hat{\theta}_N - \theta^*)^T x_i x_i^T (\hat{\theta}_N - \theta^*) \leq \frac{2}{N} \sum_{i=1}^N (y_i - x_i^T \theta^*) x_i^T (\hat{\theta}_N - \theta^*) - 2 \mathbb{E}\left[\left((y_i - x_i^T \theta^*)x_i^T (\hat{\theta}_N - \theta^*)\right)\right] + 2 \mathbb{E}\left[\left((y_i - x_i^T \theta^*)x_i^T (\hat{\theta}_N - \theta^*)\right)\right].
\]

(1.2)
The right hand side corresponds to the classical “bias-variance decomposition”. When \( \mathbb{E}[y_i] = \mathbb{E}[x_i^T \theta^*] \), the last term (which is the bias) is 0 and we only have the variance term. It should be kept in mind though that in general this bias term can be non-zero and increasing the bias in some sense can actually help us control the variance, which will be discussed in more details later.

If one believes that the matrix \( \frac{1}{N} \sum_{i=1}^{N} x_i x_i^T \) is invertible in the range of \( \Theta - \Theta := \{ \theta_1 - \theta_2 : \theta_1, \theta_2 \in \Theta \} \), i.e.

\[
\inf_{\theta_1, \theta_2 \in \Theta} \frac{1}{N} \sum_{i=1}^{N} (\theta_1 - \theta_2)^T x_i x_i^T (\theta_1 - \theta_2) \geq \sigma_{\min}
\]

for some absolute constant \( \sigma_{\min} > 0 \) and

\[
\sup_{\theta_1, \theta_2 \in \Theta} \left| \frac{1}{N} \sum_{i=1}^{N} (y_i - x_i^T \theta^*) x_i^T (\theta_1 - \theta_2) - 2\mathbb{E} [(y_i - x_i^T \theta^*) x_i^T (\theta_1 - \theta_2)] \right| \leq \gamma
\]

for some constant \( \gamma > 0 \). Then, \( (1.2) \) implies

\[
\sigma_{\min} \| \hat{\theta}_N - \theta^* \|^2 \leq \gamma \| \hat{\theta}_N - \theta^* \|_2 \Rightarrow \| \hat{\theta}_N - \theta^* \|_2 \leq \frac{\gamma}{\sigma_{\min}}.
\]

However, there are only limited scenarios where \( (1.3) \) holds. It is wrong, for example, when \( N < d \) and \( \Theta - \Theta \) spans \( \mathbb{R}^d \). Furthermore, the validity of \( (1.4) \), which essentially requires \( \frac{1}{N} \sum_{i=1}^{N} (y_i - x_i^T \theta^*) x_i^T (\theta_1 - \theta_2) \) to be uniformly concentrated around \( 2\mathbb{E} [(y_i - x_i^T \theta^*) x_i^T (\theta_1 - \theta_2)] \), is also questionable.

On the other hand, it is obvious that \( \frac{1}{N} \sum_{i=1}^{N} x_i x_i^T \) has to satisfy some invertibility conditions in order to estimate \( \theta^* \). For example, when \( \theta^* \) lies in the null space of \( \frac{1}{N} \sum_{i=1}^{N} x_i x_i^T \), asking for a bound on \( \| \hat{\theta}_N - \theta^* \|_2 \) is meaningless. Over the years, people have been trying to identify minimal conditions so that objectives like \( (1.3) \) and \( (1.4) \) holds true probabilistically, and our goal is to further expand the scope of this line of research.

### 1.1.2 Supremum of an empirical process: binary functions

It turns out that proving inequalities \( (1.3) \) and \( (1.4) \) belongs to a more general class of problems, namely, bounding the supremum of an empirical process. Historically, such kind of

\[\text{Throughout the thesis, an absolute constant is a constant that is independent of parameters of the problem.}\]
problems originates from the well-known Glivenko-Cantelli theorem.

**Theorem 1.1.1** (Glivenko-Cantelli). Suppose $X_1, X_2, \ldots, X_N \in \mathbb{R}$ is a sequence of independent and identically distributed (i.i.d.) random variables on the probability space $(\Omega, \Sigma, P)$ with a cumulative distribution function (CDF) $F(t) := P(X \leq t)$. Define the empirical CDF as $F_N(t) := \frac{1}{N} \sum_{i=1}^{N} 1\{X_i \leq t\}$, where $1\{x \leq t\}$ is the indicator function which is 1 if $x \leq t$ and 0 otherwise. Then,

$$\lim_{N \to \infty} \sup_{t \in \mathbb{R}} |F_N(t) - F(t)| = 0,$$

with probability 1.

The class of random variables $\{F_N(t) - F(t)\}_{t \in \mathbb{R}}$ is historically called an empirical process. Of course, one can show that the supremum is measurable (i.e. $\sup_{t \in \mathbb{R}} |F_N(t) - F(t)|$ is a random variable on the space $(\Omega, \Sigma, P)$, see [Dur19]), on which we will not discuss here. We further refer readers to Chapter 1 of [W+13] for a synthetic treatment of the measurability issue of the supremum. In the absence of supremum (i.e. for a fixed $t \in \mathbb{R}$), this is just law of large numbers. However, with the supremum, it is not immediately clear why the convergence is still true. More generally, for any class of (measurable) sets $\mathcal{S}$, one can ask if the following supremum always converges to zero:

$$\lim_{N \to \infty} \sup_{S \in \mathcal{S}} \left| \frac{1}{N} \sum_{i=1}^{N} 1\{X_i \in S\} - \mathbb{E}[1\{X_i \in S\}] \right|,$$

which turns out to be wrong, as is illustrated in the following simple example:

**Remark 1.1.1** (A non-Glivenko-Cantelli class). Consider the following class of indicator functions $\mathcal{F} := \{1_S(x) : |S| < \infty\}$, where $|S|$ denotes the cardinality of the set $S$. Then, it can be easily seen that for any random variable $X_i$ with a continuous distribution function $F$, $\mathbb{E}[1\{X_i \in S\}] = P(X_i \in S) = 0$. However, we have $\sup_{S \in \mathcal{S}} \frac{1}{N} \sum_{i=1}^{N} 1\{X_i \in S\} = 1$. Thus, the supremum does not converge to 0.

This example indicates that there has to be some measure of complexity which indicates that the class of function $\{1_S(x) : \text{Card}(S) < \infty\}$ is “too large” for the supremum to converge, whereas $\{1\{x \leq t\} : t \in \mathbb{R}\}$ is small. This type of complexity, which appears very often in machine learning theory, is call Rademacher complexity.

![This example is from Peter Bartlett’s lecture notes](https://www.stat.berkeley.edu/~bartlett/courses/2013spring-stat210b/notes/8notes.pdf)
**Definition 1.1.1.** Consider a set of samples \( \{X_i\}_{i=1}^N \subseteq \mathcal{X} \) and a function class \( \mathcal{F} \) containing \( f : \mathcal{X} \to \{-1, +1\} \). The empirical Rademacher complexity of the function class \( \mathcal{F} \) given \( \{X_i\}_{i=1}^N \) is defined as
\[
R_N(\mathcal{F}) := \mathbb{E} \left[ \sup_{f \in \mathcal{F}} \frac{2}{N} \sum_{i=1}^N \varepsilon_i f(X_i) \middle| X_1, \ldots, X_N \right],
\]
where \( \varepsilon_i \) being i.i.d. Rademacher random variables (taking +1 and −1 with equal probability) and independent of \( \{X_i\}_{i=1}^N \).

We have the following general theorem from [BM02]:

**Theorem 1.1.2** (Theorem 5 of [BM02]). Let \( P \) be a probability distribution on the product space \( \mathcal{X} \times \{-1, +1\} \), where \( \mathcal{X} \subseteq \mathbb{R}^d \) is a set. Let \( \mathcal{F} \) be a class of functions containing \( f : \mathcal{X} \to \{-1, +1\} \). Let \( \{X_i, Y_i\}_{i=1}^N \) be i.i.d. samples drawn according to \( P \), then, with probability at least \( 1 - \delta \), for every function \( f \in \mathcal{F} \),
\[
\left| P(Y \neq f(X)) - \frac{1}{N} \sum_{i=1}^N 1_{\{Y_i \neq f(X_i)\}} \right| \leq R_N(\mathcal{F}) + \sqrt{\frac{\ln(1/\delta)}{N}},
\]
Intuitively, \( R_N(\mathcal{F}) \) measures the correlations of \( \mathcal{F} \) with random noise, and if \( \mathcal{F} \) can fit noise very well, then, its complexity is high. To use this theorem, one should be able to compute or upper bound \( R_N(\mathcal{F}) \). One way is to apply the following theorem.

**Theorem 1.1.3** (Theorem 6 of [BM02]). Fix any sequence of samples \( X_1, \ldots, X_N \). For a function class \( \mathcal{F} \) containing \( f : \mathcal{X} \to \{-1, +1\} \), define the restriction of \( \mathcal{F} \) to the samples as follows:
\[
\mathcal{F}|_X := \{ (f(X_1), \ldots, f(X_N)) : f \in \mathcal{F} \}.
\]
Then,
\[
R_N(\mathcal{F}) \leq L \sqrt{\frac{\log |\mathcal{F}|_X}{N}},
\]
where \( L \) is an absolute constant and \( |\mathcal{F}|_X \) denotes the cardinality of the set \( \mathcal{F}|_X \).

This theorem can be proved by using the fact that \( \varepsilon_i \) is a sub-Gaussian random variable, together with a union bound. Using this lemma, one can easily prove the Glivenko-Cantelli

---

4In general this set does not have to be in \( \mathbb{R}^d \). We state this way mainly because we only care about finite dimensional spaces in this thesis.
theorem. To be more specific, we let $F = \{1_{\{x \leq t\}} : t \in \mathbb{R}\}$. One can show that $|F|_X = N + 1$, and thus, it follows from Theorem 1.1.2 with probability at least $1 - \delta$,

$$\sup_{t \in \mathbb{R}} |F_N(t) - F(t)| \leq L \sqrt{\frac{\log(N + 1)}{N}} + \sqrt{\frac{\ln(1/\delta)}{N}}.$$ 

By Borel-Cantelli Lemma, we finish the proof. Thus, not only do we prove the Glivenko-Cantelli theorem, we also get the explicit rate of convergence $O(\sqrt{\log(N + 1)}), which is otherwise difficult to obtain from “classical” proof (for example, in [Dur19]). However, as we shall see, this log $N$ is in fact not needed.

It turns out for a class of binary functions $F$, Rademacher complexity can be upper bounded by the well known complexity measure, namely, the Vapnik-Chervonenkis(VC) dimension.

**Definition 1.1.2 (VC dimension of sets).** Consider a class of sets $C$ in $X$. For a sequence of samples $X_1, \ldots, X_N \in X$, we say $C$ shatters $X_1, \ldots, X_N$ if

$$\Delta(C, X_1, \ldots, X_N) := |\{C \cap \{X_1, \ldots, X_N\} : C \in C\}| = 2^N.$$ 

The VC dimension of the class $C$, denoted as $V(C)$, is defined as

$$V(C) = \min\{N \in \mathbb{N} : \max_{X_1, \ldots, X_N \in X} \Delta(C, X_1, \ldots, X_N) < 2^N\}.$$ 

We also have the definition of VC dimension for a class of binary functions $F$:

**Definition 1.1.3 (VC dimension for classification functions).** Consider a function class $F$ containing $f : X \rightarrow \{-1, +1\}$. The VC dimension of the class $F$, denoted as $V(F)$, is defined as

$$V(F) = \min\{N \in \mathbb{N} : \max_{X_1, \ldots, X_N \in X} |F|_X | < 2^N\},$$ 

where $F|_X$ is defined in (1.5).

We have the following theorem:

**Theorem 1.1.4 (Theorem 7 of [BM02]).** Fix any sequence of samples $X_1, \ldots, X_N$. For a function class $F$ containing $f : X \rightarrow \{-1, +1\}$,

$$R_N(F) \leq L \sqrt{\frac{V(F)}{N}}.$$
where $L$ is an absolute constant.

The proof of this theorem is highly non-trivial as it is a delicate combination of Dudley’s entropy bound together with Haussler’s inequality (see Chapter 2.6-2.7 of [W+13]). One can see immediately though by using this theorem instead, we can remove the log factor in the earlier proof of Glivenko-Cantelli theorem.

1.1.3 Supremum of an empirical process: General cases

In this section, we review some key results which bound supremum of a classes of function with range in $\mathbb{R}$ instead of $\{+1, -1\}$. During the last 80’s and 90’s, there has been tremendous progress in empirical process theory, mostly associated with the name of Michel Talagrand, who has made significant contributions on various aspects of concentration of empirical processes including (but not limited to): Talagrand’s isoperimetric inequality [Tal95], Talagrand’s concentration inequality [M+00], contraction principle [LT13] and generic chaining [Tal14a]. Several of his results will be in use throughout this thesis.

We will take this opportunity trying to explain why Talagrand’s generic chaining is of central importance in modern empirical process theory and how it leads to a tight bound for the supremum of an empirical process. To understand this, we start with the follow basic definition of covering and packing numbers:

Definition 1.1.4 (Covering and packing numbers). Consider a compact metric space consisting of a set $T$ and a metric $d : T \times T \to \mathbb{R}_+$,

- An $\varepsilon$-covering of $T$ under the metric $d$ is a collection of $\{t_1, \cdots, t_N\} \subseteq T$ such that for all $t \in T$, there exists some $i \in \{1, 2, \cdots, N\}$ with $d(t, t_i) \leq \varepsilon$. The $\varepsilon$-covering number $\mathcal{N}(T, d, \varepsilon)$ is the cardinality of the minimal $\varepsilon$-covering.

- An $\varepsilon$-packing of $T$ under the metric $d$ is a collection of $\{t_1, \cdots, t_N\} \subseteq T$ such that for all $i \neq j$, $d(t_i, t_j) \geq \varepsilon$. The $\varepsilon$-packing number $\mathcal{M}(T, d, \varepsilon)$ is the cardinality of the maximal $\varepsilon$-packing.

It can be shown that covering and packing are (up to constant) the same [W+13]:

$$\mathcal{M}(T, d, \varepsilon) \leq \mathcal{N}(T, d, \varepsilon) \leq \mathcal{M}(T, d, \varepsilon/2).$$
The covering number can also be expressed in terms of general sets as opposed to metrics.

**Definition 1.1.5** (Covering net for general sets). Let $A, B$ be two sets in $\mathbb{R}^d$, the covering number $\mathcal{N}(A, B)$ is the minimum number of translates of $B$ in order to cover $A$.

It is obvious that when $A = T \subseteq \mathbb{R}^d$, $B$ is the unit ball under the metric $d$, then, $\mathcal{N}(A, \varepsilon B) = \mathcal{N}(T, d, \varepsilon)$.

The log of the covering number is also commonly referred to as the entropy number. A classical way of estimating the covering number in $\mathbb{R}^d$ is the volume argument: Let $A, B$ be a subset of $\mathbb{R}^d$, then, it is not difficult to see that (Proposition 4.2 of [Ver10b]):

$$\frac{\text{Vol}(A)}{\text{Vol}(\varepsilon B)} \leq \mathcal{N}(A, \varepsilon B) \leq \frac{\text{Vol}(A + \frac{1}{\varepsilon} B)}{\text{Vol}(\frac{1}{\varepsilon} B)},$$ (1.6)

where $\text{Vol}(A)$ is the Euclidean $\mathbb{R}^d$ volume of the set $A$. In particular, this implies for $B$ being the unit ball under the metric $d$ in $\mathbb{R}^d$,

$$\left(\frac{1}{\varepsilon}\right)^d \leq \mathcal{N}(B, d, \varepsilon) \leq \left(2 + \frac{1}{\varepsilon}\right)^d$$

However, in general, the volume argument can be suboptimal and sometimes difficult to compute. A somewhat easier way to bound the covering number is through Sudakov inequality. We need the following definition:

**Definition 1.1.6** (Gaussian mean width). Let $K$ be a set in $\mathbb{R}^d$ and let $g \sim \mathcal{N}(0, I_{d \times d})$. The Gaussian mean width of the set $K$ is $\omega(K) = \mathbb{E}[\sup_{x \in K} \langle g, x \rangle]$.

The quantity $\omega(K)$ is crucial in learning theory. Intuitively, it measures the average width of a set. One can easily check when $K$ being a unit ball in the $k$ dimensional subset of $\mathbb{R}^d$, $\omega(K) = \sqrt{k}$, and when $K$ is the cross-polytope, i.e. $K = \{x \in \mathbb{R}^d, \|x\|_1 = 1\}$, $\omega(K) = C \sqrt{\log d}$ for some absolute constant $C$.

The following is the well-known Sudakov inequality:

**Theorem 1.1.5** (Theorem 2.2 of [Ver10b]). Let $B$ be a unit ball in $\mathbb{R}^d$. For every symmetric convex set $K \subseteq \mathbb{R}^d$, we have $\sqrt{\text{log}(K, B)} \leq C \omega(K)$, where $C$ is an absolute constant.

One might wonder if it is possible to reverse the Sudakov inequality and derive an upper bound on $\mathbb{E}[\sup_{x \in K} \langle g, x \rangle]$, i.e. the supremum of a Gaussian process, in terms of covering numbers. This
turns out to be a highly non-trivial task. The technique bounding the supremum via covering nets is commonly referred to as chaining. Intuitively, chaining is a method of taking fine-grained union bounds on sets of infinite cardinality through progressively finer covering nets. We start by defining the sub-Gaussian process:

**Definition 1.1.7.** A zero mean stochastic process \( \{X_t\}_{t \in T} \) with respect to a metric \( d \) in \( T \) is called sub-Gaussian, if for every \( t_1, t_2 \in T \), and any \( \lambda \geq 0 \),

\[
\mathbb{E} \exp(\lambda (X_{t_1} - X_{t_2})) \leq \exp \left( \frac{\lambda^2 d(t_1, t_2)^2}{2} \right).
\]

For sub-Gaussian processes, we have the following key result due to R. Dudley. The technique proving this theorem is commonly referred to as Dudley’s chaining:

**Theorem 1.1.6 (Dudley’s entropy integral (Corollary 2.2.8 of [W+13])).** Consider a zero mean sub-Gaussian stochastic process \( \{X_t\}_{t \in T} \) with respect to a metric \( d \) in \( T \). Then,

\[
\mathbb{E} \left[ \sup_{t \in T} X_t \right] \leq \int_0^\infty \sqrt{\log \mathcal{N}(T, d, \varepsilon)} d\varepsilon.
\]

One might wonder how tight this bound is. The following (not so trivial) example indicates that this bound is far from being tight.

**Remark 1.1.2 (A difficult set for Dudley’s entropy integral).** This example can be found as an exercise in Chapter 2.2 of [Tal14a]. Consider the Gaussian mean width \( \omega(T) \) of the probability simplex:

\[
T = \{ t \in \mathbb{R}^d : t \geq 0, \| t \|_1 = 1 \}, \tag{1.7}
\]

where \( t \geq 0 \) is entrywise. It is easy to check that \( W(K) = C \sqrt{\log d} \) for some absolute constant \( C \). Now, compute the Dudley’s entropy integral with \( d \) being the \( \ell_2 \)-norm. One can show that (somewhat surprisingly)

\[
\int_0^\infty \sqrt{\log \mathcal{N}(T, \ell_2, \varepsilon)} d\varepsilon \geq c (\log d)^{3/2},
\]

where \( c > 0 \) is some absolute constant. Thus, Dudley’s integral is off by a factor of \( \log d \).

One way to prove the previous remark is to rewrite the Dudley integral in another form. We
consider a sequence of subsets \( T_n \subseteq T, \ n = 0, 1, 2, \cdots \) with the condition that \( |T_n| \leq N_n \) where

\[
N_0 = 1, \ N_n = 2^{2^n}, \ n \geq 1.
\]

For any \( t \in T \), define \( d(t, T_n) = \inf_{t_n \in T_n} d(t, t_n) \). Note right away we have \( \sqrt{\log N_n} = 2^{n/2} \), \( N_n^2 = N_{n+1} \) and the function \( \sqrt{\log x} \) is related to the fact that in some sense this is the inverse of the function \( \exp(-x^2) \) that governs the size of the tails of a Gaussian random variables. Define the entropy number \( e_n(T) \) as

\[
e_n(T) = \inf \sup_{t \in T} d(t, T_n),
\]

where the infimum is taken over all possible admissible sequences.

**Lemma 1.1.1** (Lemma 2.2.11 of [Tal14a]). Under the aforementioned conditions, there exists an absolute constant \( L \) such that

\[
\frac{1}{L} \sum_{n \geq 0} 2^{n/2} e_n(T) \leq \int_0^\infty \sqrt{\log \mathcal{N}(T, d, \varepsilon)} d\varepsilon \leq L \sum_{n \geq 0} 2^{n/2} e_n(T)
\]

Then, one could lower bound the entropy integral by the left hand side and lower bound the sum by a properly constructed subset of the probability simplex (1.7) (e.g. one can take subsets \( T_n \) of \( T \) consisting of sequences \( t = [t(i)]_{i=1}^d \) for which \( t(i) \in \{0, 1/n\} \).)

Note that combining Lemma 1.1.1 with Theorem 1.1.6 one readily get

\[
\mathbb{E} \left[ \sup_{t \in T} X_t \right] \leq L \sum_{n \geq 0} 2^{n/2} e_n(T) = L \sum_{n \geq 0} 2^{n/2} \inf_{t \in T} \sup_{t \in T} d(t, T_n)
\]

The key contribution of Talagrand is to realize that, surprisingly, if we exchange \( \inf \sup_{t \in T} \) with the sum, then, this bound is tight! To make this rigorous, we need the following definition of admissible sequence:

**Definition 1.1.8** (Admissible sequence). Given a metric space \((T, d)\). We say a sequence of subsets \( \{A_n\}_{n \geq 0} \) of \( T \) is increasing if \( A_n \subseteq A_{n+1}, \ \forall n \). A sequence of subsets \( \{A_n\}_{n \geq 0} \) is admissible if it is increasing and satisfy the condition that \(|A_n| \leq N_n\) where

\[
N_0 = 1, \ N_n = 2^{2^n}, \ n \geq 1.
\]
**Definition 1.1.9** (Talagrand functionals). Given a constant $\alpha > 0$ and a metric space $(T,d).$ The Talagrand $\gamma_\alpha$ functional is defined as

$$
\gamma_\alpha(T) = \inf \sup_{t \in T} \sum_{n \geq 0} 2^{n/\alpha} d(t, A_n),
$$

where the infimum is taken over all possible admissible sequences $\{A_n\}_{n \geq 0}$.

We are now ready to state the main theorem due to Talagrand:

**Theorem 1.1.7** (Talagrand majorizing measure theorem). Consider a centered Gaussian process $\{G_t\}_{t \in T}$ index by the set $T$ and the metric $d$ defined by

$$
d(s,t) = \mathbb{E}[|G_s - G_t|^2]^{1/2}.
$$

There exists some absolute constant $L > 0$ such that

$$
\frac{1}{L} \cdot \gamma_2(T) \leq \mathbb{E}\left[\sup_{t \in T} G_t\right] \leq L \cdot \gamma_2(T).
$$

Throughout the thesis, the $L_p$-norm of a random variable $X$ is defined as $\|X\|_{L_p} := \mathbb{E}[|X|^p]^{1/p}$.

**Definition 1.1.10.** A random variable $X$ is $L$ sub-Gaussian if $p^{-1/2}\|X\|_{L_p} \leq L\|X\|_{L_2}, \forall p \geq 1.$ The corresponding sub-Gaussian norm ($\psi_2$-norm) is defined as $\|X\|_{\psi_2} := \sup_{p \geq 1} p^{-1/2}\|X\|_{L_p}.$

**Definition 1.1.11** (Subgaussian random vector). A random vector $X \in \mathbb{R}^d$ is $L$ sub-Gaussian if the collection random variables $\langle X, z \rangle, z \in S^{d-1}$ are $L$ sub-Gaussian. The corresponding sub-Gaussian norm of the vector $X$ is then given by

$$
\|X\|_{\psi_2} = \sup_{z \in S^{d-1}} \|\langle X, z \rangle\|_{\psi_2}.
$$

For sub-Gaussian processes, we have

**Theorem 1.1.8** (Theorem 2.2.18 of [Tal14a]). Consider a centered sub-Gaussian process $\{X_t\}_{t \in T}$ index by the set $T$ and the metric $d$ defined by

$$
d(s,t) = \mathbb{E}[|X_s - X_t|^2]^{1/2}.
$$
We have
\[ \mathbb{E} \left[ \sup_{t \in T} X_t \right] \leq L \cdot \gamma_2(T). \]
and
\[ P \left( \sup_{t \in T} X_t \geq Lu \cdot \gamma_2(T) \right) \leq 2 \exp(-u^2). \]

Throughout the thesis, we seldom encounter any exact computation and our bounds are always in terms of unspecified absolute constants. Furthermore, the constants (for example, \( L \) and \( C \)) can be different per occurrence.

### 1.1.4 Other key inequalities

Let \((T, d)\) be a semi-metric space, and let \(X_1(t), \ldots, X_m(t)\) be independent stochastic processes indexed by \(T\) such that \(\mathbb{E}|X_j(t)| < \infty\) for all \(t \in T\) and \(1 \leq j \leq m\). We are interested in bounding the supremum of the empirical process
\[ Z_m(t) = \frac{1}{m} \sum_{i=1}^{m} \left[ X_i(t) - \mathbb{E}[X_i(t)] \right]. \]
(1.8)

The following well-known symmetrization inequality reduces the problem to bounds on a (conditionally) Rademacher process
\[ R_m(t) = \frac{1}{m} \sum_{i=1}^{m} \varepsilon_i X_i(t), \quad t \in T, \]
where \(\varepsilon_1, \ldots, \varepsilon_m\) are i.i.d. Rademacher random variables (meaning that they take values \(-1, +1\) with probability \(1/2\) each), independent of \(X_i\)’s.

**Lemma 1.1.2** (Symmetrization inequalities).
\[ \mathbb{E} \sup_{t \in T} |Z_m(t)| \leq 2 \mathbb{E} \sup_{t \in T} |R_m(t)|, \]
and for any \(u > 0\), we have
\[ \mathbb{P} \left( \sup_{t \in T} |Z_m(t)| \geq 2 \mathbb{E} \sup_{t \in T} |Z_m(t)| + u \right) \leq 4 \mathbb{P} \left( \sup_{t \in T} |R_m(t)| \geq u/2 \right). \]

See Lemmas 6.3 and 6.5 in [LT13] for proofs.

**Lemma 1.1.3** (Bernstein’s inequality [W+13]). Let \(X_1, \ldots, X_m\) be a sequence of independent centered random variables. Assume that there exist positive constants \(\sigma\) and \(D\) such that for all
integers $p \geq 2$

$$\frac{1}{m} \sum_{i=1}^{m} \mathbb{E}[|X_i|^p] \leq \frac{p!}{2} \sigma^2 D^p,$$

then

$$P \left( \left| \frac{1}{m} \sum_{i=1}^{m} X_i \right| \geq \frac{\sigma}{\sqrt{m}} \sqrt{2u} + \frac{D}{m} u \right) \leq 2 \exp(-u).$$

In particular, if $X_1, \cdots, X_m$ are all sub-exponential random variables, then $\sigma$ and $D$ can be chosen as $\sigma = \frac{1}{m} \sum_{i=1}^{m} \|X_i\|_{\psi_1}$ and $D = \max_{i=1,\cdots,m} \|X_i\|_{\psi_1}$.

**Lemma 1.1.4** (Contraction principle [LT13]). Let $X_1, \cdots, X_N$ be a sequence of samples in $X$ and let $F$ be a class of functions containing $f : X \to \mathbb{R}$. Let $\Psi_1, \Psi_2, \cdots, \Psi_N : \mathbb{R} \to \mathbb{R}$ be a sequence of $L$-Lipschitz functions for some $L > 0$, then, we have

$$\mathbb{E} \left[ \sup_{f \in F} \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i \Psi(f(X_i)) \right] X_1, \cdots, X_N \leq L \cdot \mathbb{E} \left[ \sup_{f \in F} \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i f(X_i) \right] X_1, \cdots, X_N.$$

**Lemma 1.1.5** (Contraction principle [LT13]). Let $X_1, \cdots, X_N$ be a sequence of samples in $X$, let $F$ be a class of functions containing $f : X \to \mathbb{R}$, and let $\alpha_1, \cdots, \alpha_N$ be a sequence of real numbers (possibly depends on the samples) such that $|\alpha_i| \leq 1$. We have for any $u \geq 0$,

$$P \left( \sup_{f \in F} \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i \alpha_i f(X_i) \geq u \right) X_1, \cdots, X_N \leq 2P \left( \sup_{f \in F} \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i f(X_i) \geq u \right) X_1, \cdots, X_N.$$

**Lemma 1.1.6** (Paley-Zygmund inequality [PZ30]). Suppose $Z \geq 0$ is a random variable with finite variance and $\theta \in (0, 1)$, then,

$$P(Z \geq \theta \mathbb{E}[Z]) \geq (1 - \theta)^2 \frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]}.$$

Finally, the following lemma is crucial in the analysis of heavy-tailed processes which is sometimes referred to as the Montgomery-Smith inequality:

**Lemma 1.1.7** ([MS90a]). Let $X = [X_1, \cdots, X_m]$ be a sequence of scalars. Define the following quantity:

$$K_{1,2}(X, u) := \inf \left\{ \sum_{i \in I} |X_i| + u \left( \sum_{i \notin I} |X_i|^2 \right)^{1/2}, \ I \subseteq \{1, 2, \cdots, m\} \right\}.$$
Then, we have
\[ P \left( \left| \sum_{i=1}^{m} \varepsilon_i X_i \right| \geq K_{1,2}(X, u) \right) \leq 2 \exp(-u^2/2). \quad (1.9) \]

Furthermore, there exists a universal constant \( c > 0 \) such that
\[ c^{-1} K_{1,2}(X, u) \leq \sum_{i=1}^{\lfloor u^2 \rfloor} X_i^* + u \left( \sum_{i=\lfloor u^2 \rfloor+1}^{m} (X_i^*)^2 \right)^{1/2} \leq c K_{1,2}(X, u) \]

where \( \{X_i^*\}_{i=1}^{m} \) is the non-increasing rearrangement of \( \{|X_i|\}_{i=1}^{m} \) and \( \{\varepsilon_i\}_{i=1}^{m} \) is a sequence of i.i.d. Rademacher random variables independent of \( \{X_i\}_{i=1}^{m} \).

### 1.1.5 Gordon’s theorem and bounds on the estimation error

Let’s go back to the least squares ERM discussed at the beginning and see how to perform a rigorous analysis on the estimation error. We start with (1.2) and assume the bias is 0. Further assume that \( \{x_i\}_{i=1}^{N} \) are i.i.d. Gaussian vectors from \( N(0, I_d) \), and the noise \( |y_i - x_i^T \theta^*| \leq b \) for some absolute constant \( b > 0 \). Recall the following Gordon’s “escape through the mesh” theorem:

**Theorem 1.1.9** (Gordon’s theorem (Corollary 1.2 of [Gor88])). Let \( S \) be a closed subset of unit sphere, and let matrix \( G \) be a \( N \times d \) entry-wise i.i.d. random matrix drawn from a standard Gaussian distribution \( N(0, I_N) \). Then, for any \( u \geq 0 \),
\[ P \left( \sup_{x \in S} \|Gx\|_2 - \mathbb{E}\|g_N\|_2 \geq \omega(S) + u \right) \leq \exp(u^2/2) \]

where \( g_N \sim N(0, I_{N \times N}) \).

Note that we have \( \sqrt{N} \geq \mathbb{E}\|g_N\|_2 \geq \frac{N}{\sqrt{N+1}} \). Let \( r > 0 \) and \( S_2(r) \) is the sphere centered at the origin with radius \( r \), i.e. \( S_2(r) = \{x \in \mathbb{R}^d : \|x\|_2 = r\} \). Furthermore, define the descent cone of a set \( T \subseteq \mathbb{R}^d \) at some point \( x \) as
\[ D(T, x) = \{\lambda(t-x), \ \lambda \geq 0, \ t \in T\}. \]

Note that for any vector \( \theta \in \Theta \), \( (\theta - \theta^*)/\|\theta - \theta^*\|_2 \in D(\Theta, \theta^*) \). Thus, we consider the following
infimum:
\[ \inf_{\theta \in D(\Theta, \theta^*) \cap S_2(1)} \frac{1}{N} \sum_{i=1}^{N} (x_i, \theta)^2. \]

Using Gordon’s theorem, we readily have

\[ \inf_{\theta \in D(\Theta, \theta^*) \cap S_2(1)} \frac{1}{N} \sum_{i=1}^{N} (x_i, \theta)^2 \geq \left( \sqrt{\frac{N}{N+1}} - \frac{\omega(D(\Theta, \theta^*) \cap S_2(1)) + u}{\sqrt{N}} \right)^2 \]

with probability at least \(1 - \exp(u^2/2)\). Suppose \(N \geq 4(\omega(D(\Theta, \theta^*) \cap S_2(1)) + u)^2\), then, the above quantity is no less than \(1/2\) and it follows with probability at least \(1 - \exp(u^2/2)\),

\[ \frac{1}{N} \sum_{i=1}^{N} (\hat{\theta}_N - \theta^*)^T x_i x_i^T (\hat{\theta}_N - \theta^*) \geq \frac{1}{2} \|\hat{\theta}_N - \theta^*\|_2^2. \]  

(1.10)

On the other hand, for the right hand side of (1.2), we would like to upper bound

\[ \sup_{\theta \in D(\Theta, \theta^*) \cap S_2(1)} \frac{2}{N} \sum_{i=1}^{N} (y_i - x_i^T \theta^*) x_i^T \theta - 2E [(y_i - x_i^T \theta^*) x_i^T \theta] \]

By symmetrization inequality (Lemma 1.1.2), it is enough to consider

\[ \sup_{\theta \in D(\Theta, \theta^*) \cap S_2(1)} \frac{2}{N} \sum_{i=1}^{N} \varepsilon_i (y_i - x_i^T \theta^*) x_i^T \theta, \]

where \(\varepsilon_i\)’s are i.i.d Rademacher random variable. Since \(|y_i - x_i^T \theta^*| \leq b\), by contraction principle (Lemma 1.1.5), it is enough to consider

\[ b \cdot \sup_{\theta \in D(\Theta, \theta^*) \cap S_2(1)} \frac{2}{N} \sum_{i=1}^{N} \varepsilon_i x_i^T \theta. \]

Using Theorem 1.1.8 we readily get with probability at least \(1 - 2 \exp(-u^2)\),

\[ b \cdot \sup_{\theta \in D(\Theta, \theta^*) \cap S_2(1)} \frac{2}{N} \sum_{i=1}^{N} \varepsilon_i x_i^T \theta \leq \frac{bL u \cdot \omega(D(\Theta, \theta^*) \cap S_2(1))}{\sqrt{N}}, \]

where \(L > 0\) is some absolute constant. Thus, with probability at least \(1 - c \exp(-u^2)\), where
\( c > 0 \) is some absolute constant,

\[
\frac{2}{N} \sum_{i=1}^{N} (y_i - x_i^T \theta^*) x_i^T \theta - 2 \mathbb{E} \left[ (y_i - x_i^T \theta^*) x_i^T (\hat{\theta}_N - \theta^*) \right] \leq \frac{bL u \cdot \omega(D(\Theta, \theta^*) \cap S_2(1))}{\sqrt{N}} \|\hat{\theta}_N - \theta^*\|_2.
\]

Overall, combining this inequality with (1.10), we conclude with the following theorem, which can also be found, for example, in [RV08]:

**Theorem 1.1.10.** Suppose \( \{x_i\}_{i=1}^{N} \) are i.i.d. Gaussian vectors from \( \mathcal{N}(0, I_{d \times d}) \), and the noise \( |y_i - x_i^T \theta^*| \leq b \) for some absolute constant \( b > 0 \). For any \( u \geq 0 \), if \( N \geq 4(\omega(D(\Theta, \theta^*) \cap S_2(1)) + u)^2 \), then with probability at least \( 1 - c \exp(-u^2) \), the solution to minimizing (1.1) satisfies

\[
\|\hat{\theta}_N - \theta^*\|_2 \leq \frac{bL u \cdot \omega(D(\Theta, \theta^*) \cap S_2(1))}{\sqrt{N}}.
\]

Note that such a quantity measures the “true” complexity of estimating \( \theta^* \) in the sense that the Gaussian mean width of a set can be much smaller than the ambient dimension of that set. For example, one can apply this theorem to sparse recovery problems and easily obtain a minimax optimal rate. More specifically, the work [CRPW12] shows that when taking \( \Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq \|\theta^*\|_1\} \), i.e. the ball of \( \|\cdot\|_1 \) with radius \( \|\theta^*\|_1 \), and \( \theta^* \) is \( s \)-sparse, we have \( \omega(D(\Theta, \theta^*) \cap S_2(1)) \) is on the order of \( \sqrt{s \log(d/s)} \). Thus, instead of having number of samples \( N \) scales with the dimension \( d \), we only need the sample to scale with the sparsity level \( s \log(d) \) in order to get an accurate estimation, which is in fact minimax optimal.

### 1.1.6 Theorem 1.1.10 is restrictive

Despite the simplicity of proving Theorem 1.1.10, it is fairly restrictive due to Gaussian measurements and bounded noise assumptions. One might wonder if these two assumptions are really necessary. The short answer is that they cannot be much relaxed if we would like to more or less keep the same idea of analysis. The reason is that proving Gordon’s theorem for general measurements is difficult. It is known that one can significantly relax the Gaussian assumption for special sets (For example, unit ball in \( \mathbb{R}^d \) [MP12]). For general sets, it is recently established in [LMPV17] that one can recover Theorem 1.1.10 using sub-Gaussian measurements, but with inexplicit constants. For measurements that have heavier tails than Gaussian, such a result is not known and likely untrue.
However, a closer look at the proof indicates that only a lower bound of $\frac{1}{N} \sum_{i=1}^{N} \langle x_i, \theta \rangle^2$ is needed whereas Gordon’s theorem provides a double sided bound. As a simple example, we look at bounds like

$$\frac{1}{N} \sum_{i=1}^{N} \langle x_i, \theta \rangle^2 \geq \frac{1}{2} \mathbb{E} \left[ \langle x_i, \theta \rangle^2 \right],$$

as oppose to

$$\left| \frac{1}{N} \sum_{i=1}^{N} \langle x_i, \theta \rangle^2 - \mathbb{E} \left[ \langle x_i, \theta \rangle^2 \right] \right| \leq \frac{1}{2} \mathbb{E} \left[ \langle x_i, \theta \rangle^2 \right].$$

Obviously, there are huge differences between these two inequalities. Intuitively, large values on $\langle x_i, \theta \rangle^2$ might ruin the second inequality, it only helps with the first inequality. An example demonstrating this fact is as follows:

**Remark 1.1.3** (Differences between upper and lower bounds [Men14a]). Fix an integer $N \geq 100$ and consider a sequence of i.i.d. random variables $Z_1, \ldots, Z_N$ such that each $Z_i$ takes $2\sqrt{N}$ with probability $\frac{1}{N^2}$ and takes $1$ with probability $1 - \frac{1}{N^2}$. We have

$$\mathbb{E}[Z_i^2] = 1 - \frac{1}{N^2} + \frac{4}{N}.$$

With probability at least $1/2N$, there exists some $i$ such that $Z_i = 2\sqrt{N}$, which implies $\frac{1}{N} \sum_{i=1}^{N} Z_i^2 \geq 4$. Thus, we have

$$\Pr \left( \left| \frac{1}{N} \sum_{i=1}^{N} Z_i^2 - \mathbb{E}[Z_i^2] \right| \leq \frac{1}{2} \mathbb{E}[Z_i^2] \right) \leq 1 - \frac{1}{2N}.$$  

On the other hand, if we consider the lower bound only, then, using Chernoff’s inequality, we obtain

$$\Pr \left( \frac{1}{N} \sum_{i=1}^{N} Z_i^2 \geq \frac{1}{2} \mathbb{E}[Z_i^2] \right) \geq 1 - \exp(-cN),$$

where $c > 0$ is an absolute constant.

An immediate consequence of these observations is that the standard method of analysis for the estimation problem, which is based on a two-sided concentration argument that holds with exponential probability, can never work in heavy-tailed situations. Thus, one must find a different argument altogether if one wishes to deal with learning problems that include classes of heavy-tailed functions or with a heavy-tailed target.
1.2 Small-ball Method

1.2.1 A general theorem

A key contribution in [Men14a, KM15] is a completely new method bounding the lower tail on the infimum of the quadratic form $\frac{1}{N} \sum_{i=1}^{N} \langle x_i, \theta \rangle^2$ without concentration. As is mentioned in [Men14a], the term “without concentration” should be understood in the sense of “when the concentration is false”, as oppose to “concentration methods are not needed and will not take any part in the analysis of ERM”. To state the main theorem, we need the following definition, so called “small-ball condition”.

Definition 1.2.1. A random vector $x$ is said to satisfy the small-ball condition over a set $H \subseteq \mathbb{R}^d$ if for any $v \in H$, there exist positive constants $\delta$ and $Q$ so that

$$\inf_{v \in H} P(|\langle v, x \rangle| \geq \delta \|v\|_2) \geq Q.$$ 

To see how weak the small-ball condition is, we consider a random vector $x$ satisfying $\|\langle v, x \rangle\|_{L_4} = \|v\|_2$ and the $L_4 - L_2$ equivalence condition, i.e. $\forall v \in H \subseteq \mathbb{R}^d$, $\|\langle v, x \rangle\|_{L_4} \leq L \|\langle v, x \rangle\|_{L_2}$, where $L > 0$ is an absolute constant. By Paley-Zygmund inequality, for any $\eta \in [0, 1]$,

$$P\left(|\langle v, x \rangle|^2 \geq \eta \|v\|_2^2\right) \geq \left(1 - \eta\right)^2 \frac{E[|\langle v, x \rangle|^2]^2}{E[|\langle v, x \rangle|^4]} = \left(1 - \eta\right)^2 \frac{\|\langle v, x \rangle\|_{L_4}^4}{\|\langle v, x \rangle\|_{L_4}^4} \geq \frac{(1 - \eta)^2}{L^4}.$$ 

Thus, small-ball condition does allow heavy-tailed random vectors. The key theorem by Mendelson is as follows:

Lemma 1.2.1 ([Men14a]). Let $H \subseteq S_2(1)$ and define the empirical mean width

$$\omega_N(H) := \mathbb{E}\left[\sup_{h \in H} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i \langle x_i, h \rangle\right].$$

Suppose $P(|\langle x, h \rangle| \geq \delta \|h\|_2) \geq Q$, $\forall h \in H$, then, it follows

$$\inf_{h \in H} \left(\sum_{i=1}^{N} \langle x_i, h \rangle^2\right)^{1/2} \geq \delta Q \sqrt{N} - 2\omega_N(H) - \frac{\delta u}{2}.$$ 

18
with probability at least \(1 - c e^{-u^2}\) for any \(u > 0\).

### 1.2.2 Application to least squares ERM

Lemma 1.2.1 is very powerful and applicable to analysis of many different loss functions. Here, we will show how it helps in the estimation error analysis of minimizing (1.1). We assume that the measurement \(x_i\) satisfies \(\|\langle v, x_i \rangle\|_{L_2} = \|v\|_2, \forall v \in \mathbb{R}^d\) and the \(L_4 - L_2\) equivalence condition, i.e. \(\forall v \in \mathbb{R}^d, \|\langle v, x_i \rangle\|_{L_4} \leq L \|\langle v, x_i \rangle\|_{L_2}\), where \(L > 0\) is an absolute constant.

Again, we consider the following infimum:

\[
\inf_{\theta \in D(\Theta, \theta^*) \cap \mathcal{S}_2(1)} \frac{1}{N} \sum_{i=1}^{N} (x_i, \theta)^2.
\]

By Paley-Zygmund inequality, we have

\[
P\left( \|\langle v, x_i \rangle\|^2 \geq \frac{1}{2} \|v\|_2^2 \right) \geq \frac{1}{4L^4}
\]

Applying Lemma 1.2.1, we readily have

\[
\inf_{\theta \in D(\Theta, \theta^*) \cap \mathcal{S}_2(1)} \frac{1}{N} \sum_{i=1}^{N} (x_i, \theta)^2 \geq \left( \frac{1}{8L^4} - \frac{2\omega_N(D(\Theta, \theta^*) \cap \mathcal{S}_2(1))}{\sqrt{N}} - \frac{u}{4\sqrt{N}} \right)^2
\]

with probability at least \(1 - \exp(u^2/2)\), where

\[
\omega_N(D(\Theta, \theta^*) \cap \mathcal{S}_2(1)) = \mathbb{E}\left[ \sup_{h \in D(\Theta, \theta^*) \cap \mathcal{S}_2(1)} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i \langle x_i, h \rangle \right]
\]

is the empirical mean width. Suppose

\[
N \geq 256L^8 \left( 2\omega(D(\Theta, \theta^*) \cap \mathcal{S}_2(1)) + \frac{u}{4} \right)^2,
\]

then, it follows with probability at least \(1 - \exp(u^2/2)\),

\[
\frac{1}{N} \sum_{i=1}^{N} (\hat{\theta}_N - \theta^*)^T x_i x_i^T (\hat{\theta}_N - \theta^*) \geq \frac{1}{16L^4} \|\hat{\theta}_N - \theta^*\|_2^2.
\]

(1.12)
On the other hand, define \( \xi_i = y_i - x_i^T \theta^* \), and define another empirical width:

\[
\bar{\omega}_N(D(\Theta, \theta^*) \cap S_2(1)) := \sup_{\theta \in D(\Theta, \theta^*) \cap S_2(1)} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \xi_i x_i^T \theta - 2E[\xi_i x_i^T \theta],
\]

from which we have

\[
\frac{2}{N} \sum_{i=1}^{N} (y_i - x_i^T \theta^*) x_i^T \theta - 2E[(y_i - x_i^T \theta^*) x_i^T (\hat{\theta}_N - \theta^*)] \leq \frac{2\bar{\omega}_N(D(\Theta, \theta^*) \cap S_2(1))}{\sqrt{N}} \|\hat{\theta}_N - \theta^*\|_2.
\]

Overall, we obtain the following theorem:

**Theorem 1.2.1.** Suppose \( \{x_i\}_{i=1}^{N} \) are \( L_4-L_2 \) equivalence condition, i.e. \( \forall v \in \mathbb{R}^d, \|v, x_i\|_{L_4} \leq L \|v, x_i\|_{L_2} \), where \( L > 0 \) is an absolute constant. For any \( u \geq 0 \), if

\[
N \geq 256L^8 \left( 2\omega_N(D(\Theta, \theta^*) \cap S_2(1)) + \frac{u^2}{4} \right)^2,
\]

then with probability at least \( 1 - \exp(-u^2) \), the solution to minimizing \( 1.1 \) satisfies

\[
\|\hat{\theta}_N - \theta^*\|_2 \leq \frac{2\bar{\omega}_N(D(\Theta, \theta^*) \cap S_2(1))}{\sqrt{N}}.
\]

Bounds of this flavor via the small-ball method can be found, for example, in [Tro15a]. To apply this theorem to specific problems, we need to compute the two quantities \( 1.11 \) and \( 1.13 \).

One might wonder if anything can be said regarding the general properties of these two empirical quantities. It turns out when both \( \xi_i \) and \( x_i \) are sub-Gaussian, we recover Theorem 1.1.10 up to constant via the following theorem:

**Theorem 1.2.2** (Lemma 3.2 of [GW18]). Suppose \( x_i \) is an isotropic sub-Gaussian random vector and \( \xi_i \) is a sub-Gaussian random variable. Suppose \( N \geq \omega(D(\Theta, \theta^*) \cap S_2(1))^2 \), then, with probability at least \( 1 - e^{-u^2} \),

\[
\sup_{\theta \in D(\Theta, \theta^*) \cap S_2(1)} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \xi_i x_i^T \theta - 2E[\xi_i x_i^T \theta] \leq C(\|\xi\|_{\psi_2}^2 + \|x_i\|_{\psi_2}^2)(\omega(D(\Theta, \theta^*) \cap S_2(1)) + u^2),
\]

where \( C > 0 \) is an absolute constant.

This theorem gives a bound on \( \bar{\omega}_N(D(\Theta, \theta^*) \cap S_2(1)) \). For the term \( \omega_N(D(\Theta, \theta^*) \cap S_2(1)) \),
one can simply invoke Theorem 1.1.8

\[
\mathbb{E} \left[ \sup_{\theta \in D(\Theta, \theta^*) \cap S_2(1)} \left\| \frac{2}{N} \sum_{i=1}^{N} \varepsilon_i x_i^T \theta \right\| \right] \leq \frac{L \omega(D(\Theta, \theta^*) \cap S_2(1))}{\sqrt{N}},
\]

where \( L \) is an absolute constant. Overall, we obtain the following corollary of Theorem 1.2.1.

**Corollary 1.2.1.** Suppose \( x_i \) is an isotropic sub-Gaussian random vector, \( \xi_i \) is a sub-Gaussian random variable, and

\[
N \geq C_1 (\omega_N(D(\Theta, \theta^*) \cap S_2(1)) + u^2),
\]

then, for any \( u \geq 1 \), with probability at least \( 1 - \exp(-u^2) \),

\[
\| \hat{\theta}_N - \theta^* \|_2 \leq C_2 (\| \xi \|_{\psi_2}^2 + \| x_i \|_{\psi_2}^2) \frac{\omega(D(\Theta, \theta^*) \cap S_2(1)) + u^2}{\sqrt{N}},
\]

where \( C_1, C_2 \) are absolute constants.

However, in general, when \( \xi_i \) and \( x_i \) exhibit heavier tails than Gaussian, it is highly non-trivial to bound (1.11) and (1.13) in terms of Gaussian mean width. It is an active research area and we will introduce several methods later to bound them.

### 1.3 Organization of the Thesis

The rest of the thesis is organized as follows. In Chapter 2, we introduce a new adaptively thresholded ERM for generalized linear model with a new analysis framework, which refines the results from an earlier draft [Wei18]. Special attention is devoted to recovering an approximately sparse vector in \( \ell_1 \)-ball as well as bounded sparse vectors with the minimax statistical rates under a rather weak assumption that the design vector has more than 15 moments. This result significantly improves the previously known results which require \( \mathcal{O}(\log d) \) moments (\( d \) being the dimension of the vector). In Chapter 3, we show that if one knows the design vectors are sampled from a specific class of distributions, then, a somewhat simpler analysis with even weaker assumptions is possible [GMW16][GW19]. In particular, we show that when the design vectors are elliptical symmetric with more than 2 moments, then, one can recover a structured signal (up to constant scaling) with minimax rate from measurements with unknown nonlinear transformations. Finally, in Chapter 4, we look at a problem with a somewhat different flavor,
namely, the robust covariance matrix estimation. We show that a Huber-type estimator achieves
the minimax optimal statistical rate with more than 4 moments on the samples [WM17, MW+20].
Chapter 2

Optimal Statistical Rate in Generalized Linear Models under Weak Moment Assumptions

In this Chapter, we consider the scenario of high-dimensional estimation in generalized linear models (GLMs). While high-dimensional recovery problems have been studied extensively under the sub-Gaussian assumption, much less is known in the case of heavy-tailed measurements, such as those with moments of only constant order. In this paper, we propose and analyze new thresholding methods recovering high-dimensional structured vectors from nonlinear measurements under very weak assumptions on the underlying distributions. In particular, we show that, by solving a convex program, the proposed method achieves the minimax statistical rate of estimation in $\ell_1$-ball with only $(15 + \delta)$ moments on the design vectors. Our results improve upon the best known analysis on the convex methods for ordinary linear models, i.e. LASSO type estimators, which require $O(\log d)$ moments to achieve the minimax optimal statistical rate.

2.1 Introduction

We study a general model where the response $y \in \mathbb{R}$ is linked to the covariate $x \in \mathbb{R}^d$ via a generalized linear model through a canonical link function. More specifically, we assume $y$ satisfies the following distribution

$$Pr(y \mid x; \theta^*, \sigma) \propto \exp \left( \frac{y \langle x, \theta^* \rangle - g(\langle x, \theta^* \rangle)}{c(\sigma)} \right),$$

where $\sigma$ is a known scalar parameter and $c$ is a known mapping. The vector $\theta^* \in \mathbb{R}^d$ is unknown to be estimated and $g : \mathbb{R} \to \mathbb{R}$ is the link function. Using the standard properties of an exponential
family \cite{Bro86}, we know that the function $g$ is twice differentiable and $g''$ is strictly positive on the real line. In particular, this implies the function $g$ is a strictly convex function. Some examples of GLMs are as follows:

- The ordinary linear model, i.e. $y = \langle x, \theta^* \rangle + \xi$ with $\xi \sim N(0, 1)$, corresponds to the condition distribution of $y$ being a Gaussian distribution with mean $\langle x, \theta^* \rangle$ and variance $\sigma^2$. More specifically, we have $g(\langle x, \theta^* \rangle) = (\langle x, \theta^* \rangle)^2/2$ and $c(\sigma) = \sigma^2$.

- The logistic regression model corresponds to $y$ being a Bernoulli random variable (taking values in $\{0, 1\}$). More specifically, we have $g(\langle x, \theta^* \rangle) = \log(1 + \exp(\langle x, \theta^* \rangle))$ and $c(\sigma) = 1$. In particular, we have $\Pr(y = 1 \mid x; \theta^*) = \frac{\exp(\langle x, \theta^* \rangle)}{1 + \exp(\langle x, \theta^* \rangle)}$.

- The poisson regression model corresponds to $y$ being a Poisson distribution taking values in $\mathbb{N}$ and $g(\langle x, \theta^* \rangle) = \exp(\langle x, \theta^* \rangle)$ and $c(\sigma) = 1$.

The goal is to estimate the true parameter $\theta^* \in \mathbb{R}^d$ from a sequence of $N$ samples $\{(x_i, y_i)\}_{i=1}^N$. When assuming $\theta^*$ possesses certain structure which tends to make the corresponding norm function $\Psi(\theta^*)$ small, one proposes to estimate $\theta^*$ via the following maximum likelihood (ML) with regularization:

$$
\hat{\theta}_N := \arg\min_{\theta \in \mathbb{R}^d} -\frac{1}{N} \sum_{i=1}^N y_i \langle x_i, \theta \rangle + \frac{1}{N} \sum_{i=1}^N g(\langle x_i, \theta \rangle) + \lambda \Psi(\theta). \tag{2.2}
$$

In particular, if $\theta^*$ is an approximately sparse vector, then, the usual choice for $\Psi$ is $\Psi(\theta) = \|\theta\|_1$.

Note that in general, there is a sharp contrast between ordinary linear model and the GLMs from an analysis perspective. For linear model, the analysis in the previous chapter demonstrates that an important step of controlling the error is to argue that the smallest eigenvalue of the covariance matrix $\frac{1}{N} \sum_{i=1}^N x_i x_i^T$ is away from zero in certain restricted area. However, the same argument does not work here since the quadratic component in least squares ERM is now replaced by $\frac{1}{N} \sum_{i=1}^N g(\langle x_i, \theta \rangle)$, where $g$ is only approximately quadratic on compact sets and it is not always possible to bound $g(\langle x, \theta \rangle)$ by a quadratic form.

This should be distinguished from the more restricted class of strongly convex functions for which there is a positive lower bound $c$ such that $g''(x) \geq c, \forall x \in \mathbb{R}$. On the other hand, for a strictly convex function, there is no such a uniform lower bound.
The difference is even more significant if we further assume that the covariance matrix of \( x_i \) is known, i.e. we know \( \Sigma = \mathbb{E}[x_i x_i^T] \) and it is positive definite. Consider again the ordinary linear problem. Since we know the covariance, instead of (1.1), we consider using the following ERM problem:

\[
\hat{\theta}_N = \arg \min_{\theta \in T} \mathcal{L}_m(\theta) = \theta^T \Sigma \theta - \frac{2}{N} \sum_{i=1}^{N} y_i x_i^T \theta. \tag{2.3}
\]

We then show this objective is much easier to analyze. To start, we have

\[
\hat{\theta}_N^T \Sigma \hat{\theta}_N - \frac{2}{N} \sum_{i=1}^{N} y_i x_i^T \hat{\theta}_N \leq \theta_*^T \Sigma \theta_* - \frac{2}{N} \sum_{i=1}^{N} y_i x_i^T \theta_*.
\]

Rearranging terms gives

\[
(\hat{\theta}_N - \theta_*)^T \Sigma (\hat{\theta}_N - \theta_*) \leq \frac{2}{N} \sum_{i=1}^{N} \left( y_i x_i^T (\hat{\theta}_N - \theta_*) - \mathbb{E}[y_i x_i^T (\hat{\theta}_N - \theta_*)] \right),
\]

where the expectation is taken given the \( N \) samples \( \{(x_i, y_i)\} \) and we use the fact that \( \mathbb{E}[y_i x_i^T \theta] = \mathbb{E}[\theta_*^T x_i x_i^T \theta] = \theta_*^T \Sigma \theta \). Since the covariance matrix is positive definite, we have

\[
\|\hat{\theta}_N - \theta_*\|_2 \leq \frac{1}{\lambda_{\min}(\Sigma)} \sup_{\theta \in T} \frac{2}{N} \sum_{i=1}^{N} \frac{y_i x_i^T (\theta - \theta_*) - \mathbb{E}[y_i x_i^T (\theta - \theta_*)]}{\|\theta - \theta_*\|_2}.
\]

As a consequence, we refrain from bounding the smallest eigenvalue of the empirical covariance matrix completely and small-ball method is never needed. This method was first proposed in the seminal work [KLT11] which deals with a low-rank matrix regression. However, this very method cannot be extended to analyzing objectives with general convex functions such as (2.2).

Of course knowing the covariance matrix and solving problems like (2.3) can be unrealistic depending on the application. For example, in a typical image classification problem [DDS+09], we are given a series of image samples and several class hypotheses. We would like to known which class they belong to. In such a scenario, it is unclear how one is able to obtain the population covariance of the samples and the notion of “population covariance” might not even be well-defined.
2.1.1 Related works

The ordinary linear model with \( \theta^* \) being an \( s \)-sparse vector and \( \Psi(\cdot) = \|\cdot\|_1 \) corresponds to the classical compressed sensing problem. Over the past two decades, compressed sensing has been thoroughly studied under the assumption that the measurement vectors are isotropic subgaussian and the noise is also subgaussian, e.g. [Tib96, CRT06, Can08, BRT09, HTW15]. It is shown that when each row of the measurement matrix \( \Gamma = [x_1, x_2, \ldots, x_N]^T \) is sub-Gaussian, \( N \gtrsim s \log(d/s) \), then, the restricted isometric property (RIP) holds over all \( s \)-sparse vectors \( v \in \mathbb{R}^d \), i.e. there exists a fixed constant \( \delta \in (0, 1) \), \( (1 - \delta)\|v\|_2 \leq \|\Gamma v\|_2/\sqrt{N} \leq (1 + \delta)\|v\|_2 \).

Then, one can show that by solving the LASSO: \( \hat{\theta} := \arg\min_{\theta \in \mathbb{R}^d} \|\Gamma \theta - y\|_2^2 + \lambda \|\theta\|_1 \), one can achieve the following optimal error rate: \( \|\hat{\theta} - \theta^*\|_2 \lesssim \sqrt{s \log d/N} \). Estimation of sparse vectors in generalized linear model via (2.2) with a similar statistical rate is also proved in the work [NRW12+].

As is mentioned in the previous chapter, the sub-Gaussian assumption is restrictive, but RIP does not necessarily hold with the optimal sample rate \( N \gtrsim s \log(d/s) \) when the tail of \( \langle v, x \rangle \) decays slower than sub-Gaussian. The crux lies in the fact that RIP simultaneously requires upper bounds on the quadratic form, which is not needed in the proof of performance in sparse recovery. Extending the small-ball method originally proposed in [KML15], the work [LM17b] shows that by assuming the condition that \( x \) has sub-Gaussian property up to only \( O(\log d) \) moments, i.e. \( \mathbb{E}[|\langle v, x \rangle|^p]^{1/p} \leq C \sqrt{p} \cdot \mathbb{E}[|\langle v, x \rangle|^2]^{1/2} \), \( \forall 2 \leq p \leq c_1 \log d \), where \( c_1 > 0 \) is an absolute constant, one can achieve the same aforementioned sample and error rates with high probability by solving the LASSO. Furthermore, the work [LM17a] shows that the same \( O(\log d) \) moments assumption also leads to minimax optimal estimation of an approximate sparse signal in the \( \ell_1 \)-ball instead of exact sparse signals. Outlier robust methods for sparse recovery based on the median-of-mean (MOM) estimators is also proposed and analyzed in several works (e.g. [LL17, LM16]) but they generally require solving a highly non-convex program with \( O(\log d) \) type moment assumptions on the measurement vectors in order to get the optimal rate.

Our goal in this chapter is to further relax \( O(\log d) \) moment assumption for optimal \( \ell_1 \)-ball recovery to just a constant moment requirement, which we termed “weak moment assumption”, and at the same time allow GLMs instead of just ordinary linear model. Recently, the works [FWZ17] and [SZF17] propose a new class of thresholded estimators for sparse recovery, based
on the earlier work [Cat12] on adaptive shrinkage for heavy-tailed mean estimation. While their methods are quite effective when dealing with the heavy-tailed noise, the sample rate is suboptimal when it comes to heavy-tailed measurement vectors.

2.2 Main Results

2.2.1 Optimal Estimation in $\ell_1$-ball

Throughout the chapter, we adopt the following assumption on the measurements:

**Assumption 2.2.1.** The samples \( \{(x_i, y_i)\}_{i=1}^{N} \) are i.i.d. copies of \((x, y)\) with \(E[x] = 0\), satisfying the model (2.1). For some absolute constants \(q > 15, q' > 5\), there exist corresponding constants \(\nu, \nu_q, \nu_{q'} > 0\) such that

1. Bounded kurtosis: \(\sup_{v \in S_2(1)} E[|\langle x, v \rangle|^4] \leq \nu\).
2. Bounded moments: \(\|x_i\|_{L_q} := E[|x_i|^q]^{1/q} \leq \nu_q\) and \(\|y - g'(\langle x, \theta^* \rangle)\|_{L_{q'}} \leq \nu_{q'} \forall i \in \{1, 2, \cdots, d\}\).
3. Non-degeneracy: \(\inf_{v \in S_2(1)} E[|\langle x, v \rangle|^2] \geq \kappa\).

Our result in this section concerns with the estimation in $\ell_1$-ball:

**Assumption 2.2.2.** The true parameter \(\theta^* \in B_1(R) := \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq R\}\).

Note that the set \(B_1(R)\) includes all bounded vectors that tend to be small in the $\ell_1$-norm ball (but not necessarily exactly sparse). The benchmark we will compare to is the following minimax lower bound on estimation within \(B_1(R)\) via Gaussian measurements:

**Theorem 2.2.1** (Theorem 1 of [RWY11]). Consider the ordinary linear model, i.e. \(y = \langle x, \theta^* \rangle + \xi\) with \(\xi \sim \mathcal{N}(0, 1)\) and \(x \sim \mathcal{N}(0, I_{d \times d})\). Suppose Assumption 2.2.2 holds and \(R \sqrt{\log(ed)/N} < c_1\) for some absolute constant \(c_1 > 0\), then,

\[
\min_{\hat{\theta}} \max_{\theta^* \in B_1(R)} E\left[\left\|\hat{\theta} - \theta^*\right\|_2^2\right] \geq c_2 R \sqrt{\frac{\log ed}{N}},
\]

for some absolute constant \(c_2 > 0\).

Note that an underlying assumption in this theorem (which is not explicit in [RWY11]) is that the the number of of measurements \(N \leq c_3 d^2/R^2\) for some absolute constant \(c_3 > 0\).\(^2\)

\(^2\)It is easy to see when \(N > d^2/R^2\), \(R(\log ed/N)^{1/2} \geq d/N\) and the minimax lower bound in this region should be \(d/N\), which is achieved by the least squares regression.
Our goal would be to design an estimator achieving this rate for GLMs (2.1) under Assumption 2.2.1 and 2.2.2. Our robust estimator involves generating the adapted truncated measurements \([\bar{x}_i, y_i]\)\(_{i=1}^N\) from the samples \((x_i, y_i)\)\(_{i=1}^N\) and solving the following problem:

\[
\hat{\theta}_N := \arg\min_{\theta \in \mathbb{R}^d} -\frac{1}{N} \sum_{i=1}^{N} y_i \langle \bar{x}_i, \theta \rangle + \frac{1}{N} \sum_{i=1}^{N} g(\langle \bar{x}_i, \theta \rangle) + \lambda \Psi(\theta).
\] (2.4)

where \(\lambda\) is a trade-off parameter to be determined later and \(\Psi(\theta) = \|\theta\|_1\) for the \(\ell_1\)-ball recovery problem. We take \(\bar{x}_i\) such that

\[
\bar{x}_{ij} = \text{sign} \left( x_{ij} \right) \left( |x_{ij}| \wedge \tau \right), \quad \forall j \in \{1, 2, \cdots, d\},
\] (2.5)

where \(\tau = (N/\log (ed))^{1/4}\).

Next, we will describe conditions on the link function \(g\) in (2.1), which trivially holds for the ordinary linear models.

**Assumption 2.2.3.** There exists some constant \(M_g > 0\) such that the Hessian of the cumulant function is uniformly bounded, i.e. \(\|g''\|_{\infty} \leq D_{\text{max}}\).

The following is our main result.

**Theorem 2.2.2.** Suppose Assumptions 2.2.1, 2.2.2, 2.2.3 hold. Let

\[
D_{\text{min}} := \min_{z \in [-c_1(\nu, \kappa)\|\theta_*\|_1, c_1(\nu, \kappa)\|\theta_*\|_1]} g''(z).
\]

Suppose \(N \geq C_1(\nu, \nu_q, \nu_{q'}, \kappa)\beta^2(\|\theta_*\|_1^2 + 1) \log(\frac{ed}{\epsilon^4})\), \(\lambda \geq C_2(\nu, \nu_q, \nu_{q'}, \kappa)(wu^2v + w)^{\beta/4}\sqrt{\frac{\log(\frac{ed}{\epsilon^4})}{N}}\).

Then, with probability at least

\[
1 - c' \left( e^{-\beta} + e^{-v^2} + u^{-q}(ed)^{-1}(\frac{\epsilon}{\kappa}) + (u^{-q/4} + u^{-q'})^\epsilon\right)^{-c/2} \\
+ (eN)^{-\frac{1}{2} + \frac{1}{\beta}}(\log(\frac{ed}{\epsilon^4}))^{q/5}w^{-q} + (eN)^{-\frac{1}{2} + \frac{1}{\beta}}(\log(\frac{ed}{\epsilon^4}))^{q/2}w^{-q'},
\]

for some absolute constant \(c, c' > 0\), we have

\[
\|\hat{\theta}_N - \theta_*\|_2^2 \leq \lambda \|\theta_*\|_1
\]
for any $\beta, u, v, w > 7$, where $C_i(\nu, \nu_q, \nu_q', \kappa), \ i = 1, 2, 3$ and $c_1(\nu, \kappa)$ are constants depending polynomially on the parameters $\nu, \nu_q, \nu_q', \kappa$.

**Remark 2.2.1.** Theorem 2.2.2 shows that our proposed method can attain the minimax statistical rate when $N \geq O(\log ed)$, and it does so without knowing how large $R$ is. This result also (up to constants) matches previous bounds on $\ell_1$-ball estimation which in general require stronger moment assumptions. For example, Theorem 4.2 of [LM17a] shows when the model is linear and $N \geq \log ed$, one can attain the minimax rate with $O(\log d)$ moments on the measurement vector $\{x_i\}_{i=1}^N$.

### 2.2.2 Optimal Estimation of Bounded Sparse Vectors

In this section, we show a result regarding optimal estimation of sparse vectors in a bounded range in the presence of heavy-tailed measurements. More specifically, we consider the following set of vectors:

**Assumption 2.2.4.** The true parameter $\theta^* \in \Sigma_s \setminus S_2(0, 1)$, where $\Sigma_s$ denotes the set of $s$-sparse vectors and $S_2(0, 1)$ is the unit $\ell_2$-norm ball.

The benchmark we compare to is the following lower bound:

**Theorem 2.2.3.** Consider the ordinary linear model, i.e., $y = \langle x, \theta^* \rangle + \xi$ with $\xi \sim N(0, 1)$ and $x \sim N(0, I_{d \times d})$. Suppose $\theta^* \in \Sigma_s \cap S_2(0, 1)$, $s \leq d/4$, and $(1 + \sqrt{s \log(d/s)})/\sqrt{N} < c_1$ for some absolute constant $c_1 > 0$, then,

$$
\min_{\hat{\theta}} \max_{\theta^* \in \Sigma_s \cap S_2(0, 1)} \mathbb{E} \left[ \|\hat{\theta} - \theta^*\|_2 \right] \geq c_2 \cdot \sqrt{\frac{s \log(d/s)}{N}},
$$

for some absolute constant $c_2 > 0$.

This lower bound is somewhat different from known lower bounds (e.g. [RWY11]) in the sense that it considers a restricted candidate set of sparse vectors in a bounded set $S_2(0, 1)$ instead of all sparse vectors. Nevertheless, Theorem 2.2.3 shows that imposing such a restriction does not make the problem easier. To show why it is true, we need the following definition:

**Definition 2.2.1** (Local packing number). Given a set $K \subseteq \mathbb{R}^d$, the local packing number $P_t, \ t > 0$ is the packing number of $K \cap B_2(0, t)$ with balls of radius $t/10$.  

29
Theorem 2.2.3 is a corollary of the following theorem:

**Theorem 2.2.4** (Theorem 4.2 of [PVY16]). Assume that \( \theta^* \in K \) where \( K \) is a star-shaped subset of \( \mathbb{R}^d \). Assume that \( y = \langle x, \theta^* \rangle + \xi \) with \( \xi \sim \mathcal{N}(0, \sigma^2) \) and \( x \sim \mathcal{N}(0, I_{d \times d}) \). Let

\[
\delta_* := \inf_{t > 0} \left\{ t + \frac{\sigma}{\sqrt{N}} \left( 1 + \sqrt{\log P_t} \right) \right\}.
\]

Then, there exists an absolute constant \( c > 0 \) such that any estimator \( \hat{\theta} \) which depends only on the observations \( y_i \) and \( x_i \) satisfies

\[
\sup_{x \in K} \mathbb{E} \left[ \| \hat{\theta} - \theta^* \|_2 \right] \geq c \min\{\delta_*, \text{diam}(K)\}.
\]

Now, using this theorem, it is enough to compute \( P_t \) in our problem with \( K = \Sigma_s \cap S_2(0,1) \) and \( \sigma = 1 \), for which one can show the following:

**Lemma 2.2.1.** When \( s \leq d/4 \) and \( t \leq 1 \), \( P_t \geq \exp(cs \log d/s) \), where \( c > 0 \) is an absolute constant.

**Proof of Lemma 2.2.1** The proof of this lemma follows from ideas in Section 4.3 of [PVY16]. To compute \( P_t \) for \( t \leq 1 \), it is enough to consider 1/10 packing of \( \Sigma_s \cap S_2(0,1) \). Consider a set \( \mathcal{N} \subset \Sigma_s \cap S_2(0,1) \), which contains vectors of \( s \) cardinality, where each nonzero entry is equal to \( s^{-1/2} \). Thus, \( |\mathcal{N}| = \binom{d}{s} \). We will show that there exists a subset \( \mathcal{X} \subset \mathcal{N} \) such that \( \forall x, y \in \mathcal{X}, \|x - y\|_2 > 1/10 \). Consider picking vectors \( x, y \in \mathcal{N} \) uniformly at random and compute the probability of the event \( \|x - y\|_2^2 \leq 1/100 \). When the event happens, it requires \( x \) and \( y \) to have at least 0.99s matching non-zero coordinates. Assume without loss of generality that 0.01s is an integer, this event happens with probability

\[
\left( \begin{array}{c} s \\ 0.99s \end{array} \right) \left( \begin{array}{c} d - 0.99s \\ 0.01s \end{array} \right) / \binom{d}{s}.
\]

Using Stirling’s approximation and \( s \leq n/4 \), we have \( Pr(\|x - y\|_2^2 \leq 1/100) \leq \exp(-c' s \log d/s) \), where \( c' > 0 \) is an absolute constant. This implies the claim when choose \( \mathcal{X} \) to have \( cs \log d/s \) uniformly chosen vectors from \( \mathcal{N} \), which satisfies \( \forall x, y \in \mathcal{X}, \|x - y\|_2 > 1/10 \) with a constant probability. □
Thus, by Lemma 2.2.1, it follows that

$$\inf_{t \in (0, 1]} t + \frac{1}{\sqrt{N}} \left(1 + \sqrt{cs \log d/s}\right) = \frac{1}{\sqrt{N}} \left(1 + \sqrt{cs \log d/s}\right).$$

When $N \geq c_1(1 + \log d/s)$, the claim in Theorem 2.2.3 follows.

Our main result in this section is the following theorem:

**Theorem 2.2.5.** Suppose Assumption 2.2.1, 2.2.3, 2.2.4 hold. Let $s_0 = \sqrt{\nu_\delta^2 Q^2 s} \leq d$, where $\delta = \frac{1}{2} \sqrt{\nu \kappa^2}$, $Q = \frac{\kappa^2}{2\nu}$, and $D_{\min} := \min_{z \in [-c_1(\nu, \kappa) \sqrt{\nu}, c_1(\nu, \kappa) \sqrt{\nu}]} g''(z)$. Suppose $N \geq C_1(\nu, \nu_\delta, \nu_\delta', \kappa) \beta(s_0 + 1) \log(\frac{e}{d})$, $\lambda = C_2(\nu, \nu_\delta, \nu_\delta', \kappa)(wu^2v + w^{3/4})\frac{\max_{\beta} + 1}{\min_{\beta}} \sqrt{\log(\frac{ed}{N})}$. Then, with probability at least

$$1 - c'(e^{-\beta} + e^{-v^2} + u^{-q}(\frac{e}{d})^{-(\frac{e}{d} - 1)} + (u^{-q/4} + u^{-q'})(\frac{e}{d})^{-(\frac{e}{d} - 1)}),$$

for some absolute constant $c, c' > 0$, we have

$$||\hat{\theta}_N - \theta_*||_2 \leq D_{\max} + 1 C_3(\nu, \nu_\delta, \nu_\delta', \kappa)(wu^2v + w^{3/4}) \sqrt{s \log(\frac{ed}{N})}\frac{\log(\frac{ed}{N})}{N}$$

for any $\beta, u, v, w > 7$, where $C_i(\nu, \nu_\delta, \nu_\delta', \kappa)$, $i = 1, 2, 3$ and $c_1(\nu, \kappa)$ are constants depending polynomially on the parameters $\nu, \nu_\delta, \nu_\delta', \kappa$.

### 2.3 Proof of Theorems: A Heavy-tailed Framework

In this section, we provide a general analysis on ERM of the form (2.4) which can also be applied to problems beyond $\ell_1$-regularization, and show that to control the estimation error, it is enough to control local complexities around the true vector $\theta_*$. Our procedure here is an extension of the small-ball method proposed in the works [LM17a, LM17b, Men14a], and the difference lies in the treatment of a general function $g(\cdot)$ as well as the bias caused by the thresholding.

For the rest of the paper, the notations $B_\Psi(x, r)$, $B_2(x, r)$ denote the ball of radius $r$ centered at $x$ for $\Psi$-norm, 2-norm respectively, and $S_\Psi(x, r)$, $S_2(x, r)$ denote the sphere of radius $r$ centered at $x$ for $\Psi$-norm, 2-norm respectively. We omit $x$ if they are centered at the origin.

We start with the usual optimality analysis of the ERM. Since $\hat{\theta}_N$ is the solution to (2.4), we
have

\[ \frac{1}{N} \sum_{i=1}^{N} \left( g \left( \langle \bar{x}_i, \bar{\theta}_N \rangle \right) - y_i \langle \bar{x}_i, \bar{\theta}_N \rangle \right) + \lambda \Psi \left( \bar{\theta}_N \right) \leq \frac{1}{N} \sum_{i=1}^{N} \left( g \left( \langle \bar{x}_i, \theta_* \rangle \right) - y_i \langle \bar{x}_i, \theta_* \rangle \right) + \lambda \Psi \left( \theta_* \right) \]

Simple algebraic manipulations give

\[ \frac{1}{N} \sum_{i=1}^{N} \left( g \left( \langle \bar{x}_i, \bar{\theta}_N \rangle \right) - g \left( \langle \bar{x}_i, \theta_* \rangle \right) - g' \left( \langle \bar{x}_i, \theta_* \rangle \right) \langle \bar{x}_i, \bar{\theta}_N - \theta_* \rangle \right) \]

\[ \frac{1}{N} \sum_{i=1}^{N} \langle \bar{x}_i, \bar{\theta}_N - \theta_* \rangle (y_i - g' \left( \langle \bar{x}_i, \theta_* \rangle \right)) + \lambda \left( \Psi \left( \bar{\theta}_N \right) - \Psi \left( \theta_* \right) \right) \leq 0. \] (2.6)

To simplify the notations, for any \( \mathbf{v} \in \mathbb{R}^d \), define

\[ Q_{\mathbf{v}}(x) := g \left( \langle \bar{x}, \theta_* + \mathbf{v} \rangle \right) - g \left( \langle \bar{x}, \theta_* \rangle \right) - g' \left( \langle \bar{x}, \theta_* \rangle \right) \langle \bar{x}, \mathbf{v} \rangle \]

\[ M_{\mathbf{v}}(x) := (y - g' \left( \langle \bar{x}, \theta_* \rangle \right)) \langle \bar{x}, \mathbf{v} \rangle - E[(y - \langle \bar{x}, \theta_* \rangle) \langle \bar{x}, \mathbf{v} \rangle] \]

\[ V_{\mathbf{v}} := E[(y - g' \left( \langle \bar{x}, \theta_* \rangle \right)) \langle \bar{x}, \mathbf{v} \rangle] \]

In addition, for any Borel measurable function \( G : \mathbb{R}^d \to \mathbb{R} \), \( \mathcal{P}_N G := \frac{1}{N} \sum_{i=1}^{N} G(x_i) \). Let

\[ \mathcal{L}^\lambda_{\mathbf{v}}(x) := Q_{\mathbf{v}}(x) - M_{\mathbf{v}}(x) - V_{\mathbf{v}} + \lambda \left( \Psi \left( \theta_* + \mathbf{v} \right) - \Psi \left( \theta_* \right) \right) \] (2.7)

Having defined these notations, the criterion [2.6] simply implies \( \mathcal{P}_N \mathcal{L}^\lambda_{\theta_N, \theta_*} \leq 0 \). Our goal is then to show that for any \( \theta \in \mathbb{R}^d \) such that \( \| \theta - \theta_* \|_2 > r \), where \( r > 0 \) is a certain bounding radius, then,

\[ \mathcal{P}_N \mathcal{L}^\lambda_{\theta - \theta_*} = \mathcal{P}_N Q_{\theta - \theta_*} - \mathcal{P}_N M_{\theta - \theta_*} - V_{\theta - \theta_*} + \lambda \left( \Psi \left( \theta \right) - \Psi \left( \theta_* \right) \right) > 0. \]

The intuition why one would expect this to happen is as follows. Suppose \( \Psi(\cdot) \) is not a smooth function near \( \theta_* \) and the set of sub-differentials of the norm function \( \Psi(\cdot) \) near \( \theta_* \) (which we denote as \( \partial \Psi(\theta_*) \)) is “large”, then, the set of descent directions i.e. \( D_{\Psi}(\theta_*):= \{ \theta \in \mathbb{R}^d : \Psi(\theta) \leq \Psi(\theta_*) \} \) would be relatively small\(^3\). This implies

- For \( \theta \in \mathbb{R}^d \) not in the descent directions, \( \Psi(\theta) > \Psi(\theta_*) \), and for an appropriate choice of \( \lambda \),

\(^3\)The descent cone and the cone of sub-differentials are dual to each other.
the possibly negative linear terms \(-\mathcal{P}_N \mathcal{M}_{\theta - \theta^*} - \mathcal{V}_{\theta - \theta^*}\) would be dominated by \(\Psi(\theta) - \Psi(\theta^*)\).

- For the set of \(\theta \in \mathbb{R}^d\) in the descent directions, we would expect the term \(\mathcal{P}_N \mathcal{Q}_{\theta - \theta^*}\) to dominate the linear terms \(-\mathcal{P}_N \mathcal{M}_{\theta - \theta^*} - \mathcal{V}_{\theta - \theta^*}\). Using the strictly convex property, for a sufficiently small set of descent directions intersecting with a bounded region, \(\mathcal{P}_N \mathcal{Q}_{\theta - \theta^*}\) would be a non-degenerated quadratic form (i.e. \(\mathcal{P}_N \mathcal{Q}_{\theta - \theta^*} \geq c \|\theta - \theta^*\|^2\) for some constant \(c > 0\)), which dominates the linear terms \(\mathcal{P}_N \mathcal{M}_{\theta - \theta^*}\) and \(\mathcal{V}_{\theta - \theta^*}\) for all \(\theta\) sufficiently away from \(\theta^*\) within this bounded region. We then extend this result to any vector sufficiently away from \(\theta^*\) via convexity of \(g(\cdot)\).

To this point, we invoke an idea from [LM17b] and consider the intersection of an \(\ell_2\)-ball \(B_2(\theta^*, r)\) and a \(\Psi\)-ball \(B_\Psi(\theta^*, \rho)\), with a properly chosen \(\rho > 0\), and we aim to show that if \(\theta\) is outside of \(B_2(\theta^*, r) \cap B_\Psi(\theta^*, \rho)\) with appropriate choices of \(r\) and \(\rho\), then \(\mathcal{P}_N \mathcal{L}_{\theta - \theta^*}^\lambda > 0\). As is shown in Fig. 2.1, having this intersection essentially divides the space outside of \(B_2(\theta^*, r) \cap B_\Psi(\theta^*, \rho)\) into two types of regions: 1. The region containing the set of descent directions \(D_\Psi(\theta^*)\), where the term \(\mathcal{P}_N \mathcal{Q}_{\theta - \theta^*}\) is expected to take effect. 2. The region where \(\Psi(\theta) > \Psi(\theta^*)\), and the term \(\lambda(\Psi(\theta) - \Psi(\theta^*))\) is expected to take effect.

Figure 2.1: (\(\eta = 1\)) A geometric interpretation that \(\theta \notin B_2(\theta^*, r) \cap B_\Psi(\theta^*, \rho)\) implies \(\mathcal{P}_N \mathcal{L}_{\theta - \theta^*}^\lambda > 0\): When the set of sub-differentials \(\partial \Psi(\theta^*)\) is large, the set of descent directions \(D_\Psi(\theta^*)\) is small. Then, region I contains \(D_\Psi(\theta^*)\), in which \(\Psi(\theta) \leq \Psi(\theta^*)\), and the quadratic term \(\mathcal{P}_N \mathcal{Q}_{\theta - \theta^*}\) is expected to dominate \(-\mathcal{P}_N \mathcal{M}_{\theta - \theta^*} - \mathcal{V}_{\theta - \theta^*}\). On the other hand, any vector \(\theta\) in region II has \(\Psi(\theta) > \Psi(\theta^*)\), which gives sufficient increase of norm values to dominate \(-\mathcal{P}_N \mathcal{M}_{\theta - \theta^*} - \mathcal{V}_{\theta - \theta^*}\).
Let \( \Lambda_Q, \Lambda_M \) and \( \Lambda_V \) be three positive constants. For chosen \( \rho > 0 \) and \( p_Q, p_M \in (0, 1) \), we define three critical radii:

\[
\begin{align*}
   r_Q & := \inf \left\{ r > 0 : \Pr_{\theta \in S_2(\theta_*, r) \cap B_{\Psi}(\theta_*, \rho)} \left( \mathcal{P}_N \mathcal{Q}_{\theta-\theta_*} \geq \Lambda_Q r^2 \right) \geq 1 - p_Q \right\}, \\
   r_V & := \inf \left\{ r > 0 : \sup_{\theta \in B_2(\theta_*, r) \cap B_{\Psi}(\theta_*, \rho)} |\mathcal{V}_{\theta-\theta_*}| \leq \Lambda_V r^2 \right\}, \\
   r_M & := \inf \left\{ r > 0 : \Pr_{\theta \in B_2(\theta_*, r) \cap B_{\Psi}(\theta_*, \rho)} \left( |\mathcal{P}_N \mathcal{M}_{\theta-\theta_*}| \leq \Lambda_M r^2 \right) \geq 1 - p_M \right\},
\end{align*}
\]

We then set

\[ r(\rho) := \max \{ r_Q, r_M, r_V \}. \]

Define the set of sub-differentials of the norm function \( \Psi(\cdot) \) near \( \theta_* \) (i.e. within \( \Psi \)-radius of \( \rho/4 \)) as

\[ \Gamma_{\Psi}(\theta_*, \rho) := \{ \mathbf{z} \in \mathbb{R}^d : \Psi(\mathbf{u} + \Delta \mathbf{u}) - \Psi(\mathbf{u}) \geq \langle \mathbf{z}, \Delta \mathbf{u} \rangle, \ \exists \mathbf{u} \in B_{\Psi}(\theta_*, \rho), \ \forall \Delta \mathbf{u} \in \mathbb{R}^d \}. \quad (2.8) \]

Then, the set \( \Gamma_{\Psi}(\theta_*, \rho) \) being “large” is characterized by the following quantity:

\[ \Delta(\theta_*, \rho) := \inf_{\theta \in B_2(\theta_*, r) \cap \mathcal{S}_{\Psi}(\theta_*, \rho)} \sup_{\mathbf{z} \in \Gamma_{\Psi}(\theta_*, \rho)} \langle \mathbf{z}, \theta - \theta_* \rangle \]

It characterizes the minimum amount of increase of the norm function \( \Psi(\cdot) \) from \( \Psi(\theta_*) \) on the boundary of region II in Fig. 2.1 and the set of sub-differentials \( \Gamma_{\Psi}(\theta_*, \rho) \) being “large” means for any \( \theta \in B_2(\theta_*, r) \cap \mathcal{S}_{\Psi}(\theta_*, \rho) \), there exists a vector in \( \Gamma_{\Psi}(\theta_*, \rho) \) which is close to the sub-differential of \( \theta - \theta_* \). Our goal is to show that when \( \theta \notin B_2(\theta_*, r(\rho)) \cap B_{\Psi}(\theta_*, \rho) \) and \( \Delta(\theta_*, \rho) \) is comparable to \( \rho \), then, one has \( \mathcal{P}_N \mathcal{C}^0_{\theta - \theta_*} > 0 \), as is shown in the following theorem.

**Theorem 2.3.1.** Suppose there exists \( \rho > 0 \) and \( c_1 \frac{r(\rho)^2}{\rho} \leq \lambda \leq c_2 \frac{r(\rho)^2}{\rho} \) for some constant \( c_1, c_2 \), such that \( \Lambda_Q > \Lambda_M + \Lambda_V + c_2, c_1 \geq 8(\Lambda_M + \Lambda_V) \) and \( \Delta(\theta_*, \rho) \geq \frac{3}{4} \rho \). Then, for any \( \theta \notin B_2(\theta_*, r(\rho)) \cap B_{\Psi}(\theta_*, \rho) \), \( \mathcal{P}_N \mathcal{C}^0_{\theta - \theta_*} > 0 \) with probability at least \( 1 - p_Q - p_M \).

Furthermore, if \( \rho = c\Psi(\theta_*) \) for some absolute constant \( c > 4 \), then, for \( \lambda \geq c_1 \frac{r(\rho)^2}{\rho} \) such that \( c_1 > 8(\Lambda_M + \Lambda_V) \) and \( \Lambda_Q > \Lambda_M + \Lambda_V \). Then, with probability at least \( 1 - p_Q - p_M \),

\[ \|
\hat{\theta}_N - \theta_* \|_2 \leq \max \left\{ r(\rho), \frac{\lambda}{r(\rho)(\Lambda_Q - \Lambda_M - \Lambda_V)} \Psi(\theta_*) \right\} \]

34
Remark 2.3.1. This theorem shows that the desired estimation error follows readily from tight bounds on $r_Q$, $r_M$ and $r_V$. Furthermore, in the second scenario when $\rho = c \Psi(\theta_*)$ for $c > 4$, the set $B_\Psi(\theta_*, \frac{\rho}{Q})$ contains the origin, in which case $\Gamma_\Psi(\theta_*, \rho)$ must contain the unit ball in the dual norm and $\Delta(\theta_*, \rho) \geq \rho$.

To prove this theorem we need the following simple preliminary lemma:

Lemma 2.3.1. For any $v \in \mathbb{R}^d$, $Q_{\gamma v} \geq \gamma \cdot Q_v$.

Proof of Lemma 2.3.1. First of all, by convexity of the function $g(\cdot)$,

$$
\frac{1}{\gamma} \cdot g((\bar{x}, \theta_* + \gamma v)) + \frac{\gamma - 1}{\gamma} \cdot g((\bar{x}, \theta_*)) \geq g((\bar{x}, \theta_* + v)).
$$

Rearranging the terms gives

$$
g((\bar{x}, \theta_* + \gamma v)) - g((\bar{x}, \theta_*)) \geq \gamma \cdot (g((\bar{x}, \theta_* + v)) - g((\bar{x}, \theta_*))).
$$

Substituting this relation into the definition of $Q_{\gamma v}(x)$ gives

$$
Q_{\gamma v}(x) = g((\bar{x}, \theta_* + \gamma v)) - g((\bar{x}, \theta_*)) - g'(\langle \bar{x}, \gamma v \rangle) \langle \bar{x}, \gamma v \rangle
\geq \gamma \cdot (g((\bar{x}, \theta_* + v)) - g((\bar{x}, \theta_*)) - g'(\langle \bar{x}, \theta_* \rangle) \langle \bar{x}, v \rangle)
= \gamma Q_v(x),
$$

finishing the proof.

Proof of Theorem 2.3.1. First of all, we have for any $\theta \in \mathbb{R}^d$

$$
\mathcal{P}_N \mathcal{L}_{\theta - \theta_*} \geq \mathcal{P}_N Q_{\theta - \theta_*} - |\mathcal{P}_N \mathcal{M}_{\theta - \theta_*}| - |\mathcal{V}_{\theta - \theta_*}| + \lambda (\Psi(\theta) - \Psi(\theta_*))
$$

We now prove the first part of the lemma, which is divided into the following three steps.

1. Consider first that $\|\theta - \theta_*\|_2 > r(\rho)$ and $\Psi(\theta - \theta_*) \leq \rho$. By Lemma 2.3.1 and then the definition of $r(\rho)$, we have

$$
\mathcal{P}_N Q_{\theta - \theta_*} = \frac{\|\theta - \theta_*\|_2}{r(\rho)} \cdot \mathcal{P}_N Q_{\frac{\theta - \theta_*}{\|\theta - \theta_*\|_2} r(\rho)} \geq \Lambda_Q \|\theta - \theta_*\|_2 r(\rho),
$$

35
with probability at least \(1 - p_{\mathcal{Q}}\), and
\[
|\mathcal{P}_{N}\mathcal{M}_{\theta - \theta_*}| = \left|\mathcal{P}_{N}\mathcal{M}_{\frac{\theta - \theta_*}{\|\theta - \theta_*\|_2}}(r(\rho))\right| \cdot \frac{\|\theta - \theta_*\|_2}{r(\rho)} \leq \Lambda_M \|\theta - \theta_*\|_2 r(\rho),
\]
with probability at least \(1 - p_{\mathcal{M}}\). Also,
\[
|\mathcal{V}_{\theta - \theta_*}| = \left|\mathcal{V}_{\frac{\theta - \theta_*}{\|\theta - \theta_*\|_2}}(r(\rho))\right| \cdot \frac{\|\theta - \theta_*\|_2}{r(\rho)} \leq \Lambda_V \|\theta - \theta_*\|_2 r(\rho).
\]
Thus,
\[
\mathcal{P}_{N}\mathcal{L}_{\theta - \theta_*}^\lambda \geq (\Lambda_Q - \Lambda_M - \Lambda_V)\|\theta - \theta_*\|_2 r(\rho) + \lambda (\Psi(\theta) - \Psi(\theta_*)) \tag{2.9}
\]
For \(\lambda \leq c_2 \frac{r(\rho)^2}{\rho}\), we have
\[
\lambda (\Psi(\theta) - \Psi(\theta_*)) \geq -c_2 - \frac{r(\rho)^2}{\rho} \cdot \Psi(\theta - \theta_*) \geq -c_2 r(\rho)^2 \geq -c_2 \|\theta - \theta_*\|_2 r(\rho). \tag{2.10}
\]
By the assumption that \(\Lambda_Q > \Lambda_M + \Lambda_V + c_2\), we know that \(\mathcal{P}_{N}\mathcal{L}_{\theta - \theta_*}^\lambda > 0\) with probability at least \(1 - p_{\mathcal{Q}} - p_{\mathcal{M}}\).

2. Consider the case \(\|\theta - \theta_*\|_2 \leq r(\rho)\) and \(\Psi(\theta - \theta_*) > \rho\), then, for any specific \(\theta\) satisfying the aforementioned conditions,
\[
\mathcal{P}_{N}\mathcal{L}_{\theta - \theta_*}^\lambda \geq -|\mathcal{P}_{N}\mathcal{M}_{\theta - \theta_*}| - |\mathcal{V}_{\theta - \theta_*}| + \lambda (\Psi(\theta) - \Psi(\theta_*))
\]
\[
= \left(-\left|\mathcal{P}_{N}\mathcal{M}_{\frac{\theta - \theta_*}{\|\theta - \theta_*\|_2}}(r(\rho))\right| - \left|\mathcal{V}_{\frac{\theta - \theta_*}{\|\theta - \theta_*\|_2}}(r(\rho))\right|\right) \cdot \frac{\Psi(\theta - \theta_*)}{\rho} + \lambda (\Psi(\theta) - \Psi(\theta_*))
\]
\[
\geq - (\Lambda_M + \Lambda_V) r(\rho)^2 \cdot \frac{\Psi(\theta - \theta_*)}{\rho} + \lambda (\Psi(\theta) - \Psi(\theta_*)).
\]
Let \(u \in B_{\Psi}(\theta_*, \rho/4)\) be the vector containing a sub-differential \(z \in \partial \Psi(u)\) such that \(\langle z, \theta - \theta_* \rangle \geq \frac{3}{4} \Psi(\theta - \theta_*)\). Note that this is possible because by the assumption that \(\Delta(\theta_*, \rho) \geq \frac{3}{4} \rho\), we have there exists \(u \in B_{\Psi}(\theta_*, \rho/4)\) with a sub-differential \(z \in \partial \Psi(u)\) such that \(\langle z, \frac{\theta - \theta_*}{\Psi(\theta - \theta_*)} \rangle \geq \frac{3}{4} \rho\). Thus, for the same choice of \(u\) and \(z\), \(\Psi(\theta - \theta_*) > \rho\) implies
\[
\langle z, \theta - \theta_* \rangle = \left\langle z, \frac{\theta - \theta_*}{\Psi(\theta - \theta_*)} \right\rangle \cdot \frac{\Psi(\theta - \theta_*)}{\rho} \geq \frac{3}{4} \Psi(\theta - \theta_*). \tag{2.11}
\]
This implies

\[
P_N \mathcal{L}_{\theta - \theta_*}^\lambda \geq - (A_M + A_N) r(\rho)^2 \cdot \frac{\Psi(\theta - \theta_*)}{\rho} + \lambda (\Psi(\theta) - \Psi(\theta_* + u - u))
\]

\[
\geq - (A_M + A_N) r(\rho)^2 \cdot \frac{\Psi(\theta - \theta_*)}{\rho} + \lambda \left( \langle z, \theta - u \rangle - \frac{\rho}{4} \right)
\]

\[
\geq - (A_M + A_N) r(\rho)^2 \cdot \frac{\Psi(\theta - \theta_*)}{\rho} + \lambda \left( \langle z, \theta - \theta_* \rangle - \frac{\rho}{2} \right)
\]

\[
\geq \left( -(A_M + A_N) r(\rho)^2 + \lambda \cdot \frac{\rho}{4} \right) \cdot \frac{\Psi(\theta - \theta_*)}{\rho},
\]

where the second inequality follows from \( u \in B_\psi(\theta_* \rho/4) \), the third inequality follows from the definition of sub-differential, the fourth inequality follows from Holder’s inequality \( \langle z, \theta_* - u \rangle \leq \Psi^*(z) \Psi(\theta_* - u) \leq \frac{\rho}{4} \) and the final inequality follows from the preceding argument \((2.11)\). Now, we use the assumption that \( \lambda \geq c_1 \frac{r(\rho)^2}{\rho} \) and \( c_1 \geq 8(A_M + A_N) \) to conclude that \( P_N \mathcal{L}_{\theta - \eta \theta_*} > 0 \).

3. The case \( \| \theta - \theta_* \|_2 > r(\rho) \) and \( \Psi(\theta - \theta_*) > \rho \). If \( \| \theta - \theta_* \|_2 > \frac{r(\rho)}{\rho} \), then, let \( \alpha = \frac{\Psi(\theta - \theta_*)}{\rho} \). We have by Lemma \((2.3.1)\) and then \((2.9), (2.10)\) in step 1,

\[
P_N \mathcal{L}_{\theta - \theta_*}^\lambda \geq \alpha P_N \mathcal{Q}_{\frac{\theta - \theta_*}{\Psi(\theta - \theta_*) \rho}} - \alpha \left( \left| P_N \mathcal{M}_{\frac{\theta - \theta_*}{\Psi(\theta - \theta_*) \rho}} \right| + \left| V_{\frac{\theta - \theta_*}{\Psi(\theta - \theta_*) \rho}} \right| \right) - \lambda (\Psi(\theta) - \Psi(\theta_*)) > 0.
\]

On the other hand, if \( \| \theta - \theta_* \|_2 \leq \frac{r(\rho)}{\rho} \), then, let \( \alpha = \frac{\| \theta - \theta_* \|_2}{\frac{r(\rho)}{\rho}} \) and we have

\[
P_N \mathcal{L}_{\theta - \theta_*}^\lambda \geq -2\alpha \left( \left| P_N \mathcal{M}_{\frac{\theta - \theta_*}{\Psi(\theta - \theta_*) \rho}} r(\rho) \right| + \left| V_{\frac{\theta - \theta_*}{\Psi(\theta - \theta_*) \rho}} r(\rho) \right| \right) + \lambda (\Psi(\theta) - \Psi(\theta_*)) > 0,
\]

by step 2.

This finishes the proof of the first part.

For the second part of the claim, one first considers the case \( \| \theta - \theta_* \|_2 > r(\rho) \) and \( \Psi(\theta - \theta_*) \leq \rho \). Using the fact that \( \theta_N \) is a minimizer of \( P_N \mathcal{L}_{\theta - \theta_*}^\lambda \), we get \( P_N \mathcal{L}_{\theta_N - \theta_*}^\lambda \leq 0 \). By \((2.9)\) in step 1 of the proof,

\[
(L_Q - A_M - A_N) \| \theta_N - \theta_* \|_2 r(\rho) \leq \lambda \Psi(\theta_*).
\]

37
This implies
\[ \|\hat{\theta}_N - \theta_*\|_2 \leq \frac{\lambda \Psi(\theta_\ast)}{r(\rho)(\Lambda_Q - \Lambda_M - \Lambda_V)}. \]

For the case \( \|\theta - \theta_*\|_2 \leq r(\rho) \) and \( \Psi(\theta - \theta_\ast) > \rho \), one can invoke step 2 of the above proof. Instead of using the assumption \( \Delta(\theta_\ast, \rho) \geq \frac{3\rho}{4} \), we consider the following argument: Since \( \rho/4 > \Psi(\theta_\ast) \), the set \( B_{\Psi}(\theta_\ast, \rho/4) \) must contain the origin. Thus, one can take \( u \) in (2.8) to be 0 and by Hahn-Banach theorem, the set \( \Gamma_{\Psi}(\theta_\ast, \rho) \) must contain the unit ball of the dual norm, i.e. for any \( v \in \mathbb{R}^d \), there exists a vector \( z \in \mathbb{R}^d \) such that \( \Psi^*(z) = 1 \) and \( \langle z, v \rangle = \Psi(v) \). As a consequence, \( \Delta(\theta_\ast, \rho) \geq \rho \) and we have for any \( \theta \), there exists a \( z \in \Gamma_{\Psi}(\theta_\ast, \rho) \), such that \( \langle z, \theta - \theta_\ast \rangle = \Psi(\theta - \theta_\ast) > \rho \). The rest of step 2 and step 3 carry through. Overall, we finish the proof.

\[
2.4 \ \text{Proof of Theorem 2.2.2: Computing Local Complexities}
\]

\[
2.4.1 \ \text{Bounding } r_Q : \text{ Preliminary estimates}
\]

In this section, we bound the local complexity \( r_Q \). We let \( \Psi(\cdot) \) to be the \( \ell_1 \)-norm. Note first that
\[
Q_{\theta - \theta_*}(x) = g(\langle \bar{x}, \theta \rangle) - g(\langle \bar{x}, \theta_* \rangle) - g'(\langle \bar{x}, \theta_* \rangle) \langle \bar{x}, \theta - \theta_* \rangle
\]
\[
= g''(\langle \bar{x}, \theta_* \rangle + \alpha \langle \bar{x}, \theta - \theta_* \rangle) \langle \bar{x}, \theta - \theta_* \rangle, \tag{2.12}
\]

where \( \alpha \in [0, 1] \). Define the constants \( \delta = \frac{1}{4} \sqrt{\frac{2}{\lambda}} \) and \( Q = \frac{Q^2}{\nu^2} \), where \( \kappa, \nu \) are defined in Assumption 2.2.1. Let \( s_0 \) be a constant less than \( d \) (to be defined later) and define \( G_{s_0} \) to be the set of vectors with \( s_0 \) cardinality.

Our goal is to show that the intersection of the following three sets, when taking infimum over \( v_1 \in G_{s_0} \cap S_2(1) \) and \( v_2 \in S_2(0, r) \cap B_{\Psi}(0, \rho) \) is sufficiently large:

\[
\{ i : |\langle \bar{x}, v_1 \rangle | \geq \delta \} \cap \{ i : |\langle \bar{x}, v_2 \rangle | \leq 32(\nu^2 + \nu + 1)\rho/Q \} \cap \{ i : |\langle \bar{x}, \theta_* \rangle | \leq 32\nu \|\theta_*\|_1/Q \},
\]

where \( c \) is an absolute constant.
Lemma 2.4.1. Let \( u \geq 1 \) and \( N \geq 1024u/Q^2 + c \log e d \) for some absolute constant \( c > 0 \). With probability at least \( 1 - e^{-u} \),

\[
\sup_{v \in B_{\psi}(0, \rho)} \frac{1}{N} \sum_{i=1}^{N} 1 \{ |\langle v, \bar{x}_i \rangle | \geq 32(\nu_q^2 + \nu_q + 1)\rho/Q \} \leq \frac{Q}{16}.
\]

Proof of Lemma 2.4.1. Let \( \tau_1 = 32(\nu_q^2 + \nu_q + 1)/Q \) and define \( \psi(t) = t/\tau_1 \rho \). First, by finite difference inequality, we have with probability \( 1 - e^{-u} \)

\[
\sup_{v \in B_{\psi}(0, \rho)} \frac{1}{N} \sum_{i=1}^{N} 1 \{ |\langle v, \bar{x}_i \rangle | \geq \tau_1 \rho \} \leq \frac{Q}{16}.
\]

Thus, it is enough to bound the expected supremum. We have

\[
\mathbb{E}\left[ \sup_{v \in B_{\psi}(0, \rho)} \frac{1}{N} \sum_{i=1}^{N} 1 \{ |\langle v, \bar{x}_i \rangle | \geq \tau_1 \rho \} \right] \leq \mathbb{E}\left[ \sup_{v \in B_{\psi}(0, \rho)} \frac{1}{N} \sum_{i=1}^{N} \psi(|\langle v, \bar{x}_i \rangle |) \right]
\]

\[
= \mathbb{E}\left[ \sup_{v \in B_{\psi}(0, \rho)} \frac{1}{N} \sum_{i=1}^{N} \psi(|\langle v, \bar{x}_i \rangle |) - \mathbb{E}[\psi(|\langle v, \bar{x}_i \rangle |)] + \mathbb{E}[\psi(|\langle v, \bar{x}_i \rangle |)] \right]
\]

\[
\leq 2 \mathbb{E}\left[ \sup_{v \in B_{\psi}(0, \rho)} \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i \psi(|\langle v, \bar{x}_i \rangle |) \right] + \sup_{v \in B_{\psi}(0, \rho)} \mathbb{E}[\psi(|\langle v, \bar{x}_i \rangle |)]
\]

\[
\leq 2 \mathbb{E}\left[ \sup_{v \in B_{\psi}(0, \rho)} \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i \langle v, \bar{x}_i \rangle \right] + \sup_{v \in B_{\psi}(0, \rho)} \mathbb{E}[|\langle v, \bar{x}_i \rangle |] \quad (\text{II})
\]

where the first inequality follows from \( \psi(t) \geq 1_{\{t \geq \tau_1 \rho \}} \), the second inequality follows from symmetrization inequality and the last inequality follows from Talagrand contraction principle. Now, we bound the two terms respectively.

- **Bounding (I):** First, by Bernstein’s inequality,

\[
\Pr \left( \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i \bar{x}_i \geq \sqrt{\frac{2\nu_q^2 u}{N}} + \frac{u}{(\log ed)^{1/4} N^{3/4}} \right) \leq 2e^{-u/2}.
\]
Taking a union bound over \( j \in \{1, 2, \ldots, d\} \),

\[
Pr \left( \max_j \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i x_{ij} \geq \sqrt{\frac{2q v^2 u \log ed}{N}} + \frac{2u (\log ed)^{3/4}}{N^{3/4}} \right) \leq 2e^{-u/2}.
\]

Since \( N \geq \log ed \) and \( u \geq 1 \), this implies

\[
Pr \left( \max_j \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i x_{ij} \geq (\sqrt{2q v} + 2)u \sqrt{\frac{\log ed}{N}} \right) \leq 2e^{-u/2}.
\]

Thus,

\[
E \left[ \max_j \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i x_{ij} \right] \leq (1 + 4\sqrt{e}) (\sqrt{2q v} + 2)u \sqrt{\frac{\log ed}{N}},
\]

and

\[
\frac{2}{\tau_1 \rho} E \left[ \sup_{v \in B_q(0, \rho)} \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i \langle v, \bar{x}_i \rangle \right] \leq \frac{2}{\tau_1} E \left[ \max_j \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i x_{ij} \right] \leq \frac{2(1 + 4\sqrt{e}) (\sqrt{2q v} + 2)}{\tau_1} \sqrt{\frac{\log ed}{N}}.
\]

**Bounding (II):**

\[
\sup_{v \in B_q(0, \rho)} \frac{E[\| \langle v, \bar{x}_i \rangle \|]}{\tau_1 \rho} \leq \frac{1}{\tau_1 \rho} \sup_{v \in B_q(0, \rho)} E[\| \langle v, x_i \rangle \|] + E[\| \langle v, x_i - \bar{x}_i \rangle \|] \leq \frac{1}{\tau_1} \left( \nu_q + \nu_q^2 \left( \frac{\log ed}{N} \right)^{1/4} \right),
\]

where the last inequality follows from:

\[
\sup_{v \in B_q(0, \rho)} E[\| \langle v, x_i \rangle \|] \leq \sup_{v \in B_q(0, \rho)} E[\langle v, x_i \rangle^2]^{1/2} \leq \rho \max_{j,k} E[|x_{ijk}|]^{1/2} \leq \rho \nu_q,
\]

and the following derivation:

\[
E[\| \langle v, x_i - \bar{x}_i \rangle \|] = E \left[ \sum_{j=1}^{d} v_j (x_{ij} - \bar{x}_{ij}) \right] \leq \sum_{j=1}^{d} |v_j| E[|x_{ij} - \bar{x}_{ij}|] \leq \rho \max_j E[|x_{ij} - \bar{x}_{ij}|]
\]

and

\[
E[|x_{ij} - \bar{x}_{ij}|] \leq E[|x_{ij}| 1_{|x_{ij}| > (N/\log ed)^{1/4}}] \leq E[|x_{ij}|^2]^{1/2} Pr(|x_{ij}| > (N/\log ed)^{1/4})^{1/2} \leq E[|x_{ij}|^2]^{1/2} E[|x_{ij}|^2]^{1/2} \left( \frac{\log ed}{N} \right)^{1/4} \leq \nu_q \left( \frac{\log ed}{N} \right)^{1/4},
\]

\[40\]
where the first inequality follows from the definition that $$\tilde{x}_{ij} = \text{sign}(x_{ij})|x_{ij}|^{\wedge}(N/\log ed)^{1/4}$$, the second inequality follows from Holder’s inequality and the third inequality follows from Markov inequality.

Overall, we obtain with probability $$1 - e^{-u}$$,

$$\sup_{v \in B_{\phi}(0, \rho)} \frac{1}{N} \sum_{i=1}^{N} 1\{ |\langle \theta^*, \tilde{x}_i \rangle | \geq \tau_1 \rho \} \leq \frac{2(1 + 4\sqrt{e})(\sqrt{2\nu_q} + 2)}{\tau_1} \sqrt{\frac{\log ed}{N}} + \frac{1}{\tau_1} \left( \nu_q + \nu_q^2 \left( \frac{\log ed}{N} \right)^{1/4} \right) + \sqrt{\frac{u}{N}}.$$  

Since $$N \geq 1024u/Q^2 + e \log ed$$ and $$\tau_1 = 32(\nu_q^2 + \nu_q + 1)/Q$$, it follows for $$c$$ large enough, we have

$$\sup_{v \in B_{\phi}(0, \rho)} \frac{1}{N} \sum_{i=1}^{N} 1\{ |\langle \theta^*, \tilde{x}_i \rangle | \geq \tau_1 \rho \} \leq \frac{Q}{16},$$

finishing the proof.

**Lemma 2.4.2.** Let $$u \geq 1$$ and $$N \geq 1024u/Q^2$$. With probability at least $$1 - e^{-u}$$,

$$\frac{1}{N} \sum_{i=1}^{N} 1\{ |\langle \theta^*, \tilde{x}_i \rangle | \geq 32\nu_q \| \theta^* \|_1/Q \} \leq \frac{Q}{16}.$$  

**Proof of Lemma 2.4.2.** First of all, note that

$$\mathbb{E}[|\langle \theta^*, \tilde{x}_i \rangle |] \leq \mathbb{E}[|\langle \theta^*, \tilde{x}_i \rangle |^2]^{1/2} \leq \sum_{i=1}^{d} |\theta^*| \cdot \mathbb{E}[|\tilde{x}_{ij}|^2]^{1/2} \leq \| \theta^* \|_1 \nu_q.$$  

By Markov inequality,

$$\mathbb{E}[1\{ |\langle \theta^*, \tilde{x}_i \rangle | \geq 32\nu_q \| \theta^* \|_1/Q \}] = \Pr \left( |\langle \theta^*, \tilde{x}_i \rangle | \geq \frac{32\nu_q \| \theta^* \|_1}{Q} \right) \leq \frac{Q}{32}.$$  

By bounded difference inequality,

$$\Pr \left( \frac{1}{N} \sum_{i=1}^{N} 1\{ |\langle \theta^*, \tilde{x}_i \rangle | \geq 32\nu_q \| \theta^* \|_1/Q \} \right) \geq \mathbb{E}[1\{ |\langle \theta^*, \tilde{x}_i \rangle | \geq 32\nu_q \| \theta^* \|_1/Q \}] + \sqrt{\frac{u}{N}} \leq e^{-u}.$$  

Thus, it follows when $$N \geq 1024u/Q^2$$, the desired inequality holds.
2.4.2 Weak small-ball estimates for small \( N \)

In this section, we consider lower bounding the cardinality of the set \( \{ i : |\langle \tilde{x}, v_1 \rangle | \geq \delta \} \), \( v_1 \in G_{s_0} \cap S_2(1) \) when \( s_0 \leq d \). This holds when \( N \geq \frac{N}{\log ed} \cdot \frac{Q}{\delta} \leq d \) which is \( N \leq \frac{\nu}{Q}d \log ed \).

We start with the following small-ball estimate via Paley-Zygmund inequality:

**Lemma 2.4.3.** Under Assumption 2.2.1, let \( \delta = \frac{1}{2} \sqrt{\frac{2}{42}} \) and \( Q = \frac{\nu^2}{8} \), then, we have

\[
\inf_{v \in \mathbb{R}^d} \Pr \left( \left| \langle x_i, v \rangle \right| \geq 2\delta \|v\|_2 \right) \geq 2Q.
\]

**Proof.** By Paley-Zygmund inequality, we know for any nonnegative real valued random variable \( Z \),

\[
Pr(Z > tE[Z]) \geq \left(1 - \frac{t}{E[Z]}\right)^2 E[Z]^2,
\]

for any \( t \geq 0 \). Now, fix any \( v \in \mathbb{R}^d \), we take \( Z = |\langle x_i, v \rangle|^2, t = 1/2 \), and obtain

\[
Pr \left( \left| \langle x_i, v \rangle \right|^2 \geq \frac{1}{4} \frac{1}{\kappa} \\frac{\|v\|_2^2}{E[|\langle x_i, v \rangle|^4]} \right) \geq \frac{1}{4} \frac{1}{\kappa} \\frac{\|v\|_2^2}{E[|\langle x_i, v \rangle|^4]} \geq \frac{\kappa^2}{4\nu},
\]

where the last inequality follows from Assumption 2.2.1. Taking \( \delta = \frac{1}{2} \sqrt{\frac{2}{42}} \) and \( Q = \frac{\nu^2}{8} \) finishes the proof.

We see from Lemma 2.4.3 that indeed such a small-ball condition is easily satisfied merely under a bounded moment assumption. The following lemma is the key to our analysis in this step. It says a somewhat “weak” small-ball condition is preserved under adaptive truncation.

**Lemma 2.4.4.** Let \( s_0 \) be a positive integer such that \( 1 \leq s_0 \leq d \). Let \( G_{s_0} \) be the set of all vectors in \( \mathbb{R}^d \) with \( s_0 \) cardinality of the support set. Suppose Assumption 2.2.1 holds and \( N \geq \frac{\nu}{Q}s_0 \log(ed) \),
then, for any $v \in G_{s_0}$,

$$Pr \left( \left| \langle \bar{x}_i, v \rangle \right| \geq \delta \|v\|_2 \right) \geq Q.$$ 

**Proof.** First, note that for any vector $v \in G_{s_0}$,

$$|\langle \bar{x}_i, v \rangle| = |\langle \bar{x}_i - x_i, v \rangle + \langle x_i, v \rangle| \geq |\langle x_i, v \rangle| - |\langle \bar{x}_i - x_i, v \rangle|.$$ 

Thus, it follows

$$Pr \left( |\langle \bar{x}_i, v \rangle| \geq \delta \|v\|_2 \right) \geq Pr \left( |\langle x_i, v \rangle| \geq 2\delta \|v\|_2 \right)$$

$$\geq Pr \left( |\langle x_i, v \rangle| \geq 2\delta \|v\|_2 \cap \{|\langle \bar{x}_i - x_i, v \rangle| \leq \delta \|v\|_2\} \right)$$

$$\geq Pr \left( |\langle x_i, v \rangle| \geq 2\delta \|v\|_2 \right) - Pr \left( |\langle \bar{x}_i - x_i, v \rangle| \geq \delta \|v\|_2 \right), \quad (2.13)$$

where the last inequality follows from the fact that for any two measurable set $A, B$ in a probability space $(\Omega, \mathcal{E}, \mathbb{P})$, $Pr(A \cap B) = Pr(A \setminus (B^c \cap A)) \geq Pr(A) - Pr(B^c \cap A) \geq Pr(A) - Pr(B^c)$. By Lemma 2.4.3, $Pr \left( |\langle x_i, v \rangle| \geq 2\delta \|v\|_2 \right) \geq 2Q$. It remains to bound $Pr \left( |\langle \bar{x}_i - x_i, v \rangle| \geq \delta \|v\|_2 \right)$ from above. To this point, let $P_v x_i$ be the orthogonal projection of a vector $x \in \mathbb{R}^d$ onto the non-zero coordinates of $v$. Then, by Holder’s inequality, we have

$$Pr \left( |\langle \bar{x}_i - x_i, v \rangle| \geq \delta \|v\|_2 \right) \leq Pr \left( \|P_v(\bar{x}_i - x_i)\|_\infty \|v\|_1 \geq \delta \|v\|_2 \right)$$

$$= Pr \left( \|P_v(\bar{x}_i - x_i)\|_\infty \geq \delta \frac{\|v\|_2}{\|v\|_1} \right)$$

$$\leq Pr \left( \|P_v x_i\|_\infty > \tau \right),$$

where the last inequality follows from the definition of $\bar{x}_i$ in (2.5) that if every entry of $P_v x_i$ is bounded by $\tau$, then $P_v x_i = P_v \bar{x}_i$. Furthermore,

$$Pr \left( \|P_v x_i\|_\infty > \tau \right) \leq Pr \left( \left( \sum_{j \in \mathcal{G}_v} x_{ij}^4 \right)^{\frac{1}{4}} \right)$$

$$= Pr \left( \sum_{j \in \mathcal{G}_v} x_{ij}^4 > \tau^4 \right)$$

$$\leq \frac{\mathbb{E} \left[ \sum_{j \in \mathcal{G}_v} x_{ij}^4 \right]}{\tau^4} \leq \frac{s_0 \nu \log(ed)}{N},$$

where the second from the last inequality follows from Markov inequality and the last inequality

43
follows from the definition of $\tau = (N/ \log(ed))^{1/4}$ and the assumption that $\mathbb{E}[x_{ij}^4] \leq \nu$. Since $N \geq \frac{\nu}{Q} s_0 \log(ed)$ by assumption, we have $Pr(\|P_v x_i\|_\infty > \tau) \geq Q$ and the proof is finished. \(\blacksquare\)

Using the previous lemma one can show the following via a book-keeping VC dimension argument.

**Lemma 2.4.5.** Consider any integer $s_0$ such that $1 \leq s_0 \leq d$. Suppose $N \geq \frac{\nu}{Q} s_0 \log(ed)$, then, with probability at least $1 - \epsilon \exp(-u)$,

$$\inf_{v \in \mathcal{G}_{s_0} \cap S_2(1)} \frac{1}{N} \sum_{i=1}^{N} 1\{|(\bar{x}_i, v)| \geq \delta/2\} \geq Q - L \sqrt{s_0 \log(ed)/N} - \sqrt{u/N},$$

where $L, c \geq 1$ are absolute constants.

**Proof of Lemma 2.4.5.** First of all, by Lemma 2.4.4, for any $i \in \{1, 2, \cdots, N\}$ and $v \in \mathcal{G}_{s_0} \cap S_2(1)$, we have

$$\mathbb{E} \left[ 1\{|(\bar{x}_i, v)| \geq \delta\} \right] = Pr \left( |(\bar{x}_i, v)| \geq \delta \|v\|_2 \right) \geq Q.$$

Let $\bar{x}_N := [\bar{x}_1, \cdots, \bar{x}_N]$, and define the following process parametrized by $v \in \mathcal{G}_{s_0} \cap S_2(1)$:

$$R(\bar{x}_N, v) = \frac{1}{N} \sum_{i=1}^{N} 1\{|(\bar{x}_i, v)| \geq \delta/2\} - \mathbb{E} \left[ 1\{|(\bar{x}_i, v)| \geq \delta/2\} \right],$$

and we aim to bound the following supremum

$$\sup_{v \in \mathcal{G}_{s_0} \cap S_2(1)} |R(\bar{x}_N, v)|.$$

Define the following class of indicator functions:

$$\mathcal{F} := \left\{ 1\{|(\cdot, v)| \geq \delta/2\}, \; v \in \mathcal{G}_{s_0} \cap S_2(1) \right\},$$

By the standard symmetrization argument and then Dudley's entropy estimate (see, for example, [VDVW96a] for details of VC theory), we have

$$\mathbb{E} \left[ \sup_{v \in \mathcal{G}_{s_0} \cap S_2(1)} |R(\bar{x}_N, v)| \right] \leq C_0 \sqrt{\frac{s_0 \log N}{N}} \sqrt{\log \mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_{L_2(\mu_N)})} d\varepsilon, \quad (2.14)$$

where $C_0$ is a constant and $\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_{L_2(\mu_N)})$ is the $\varepsilon$-covering number of $\mathcal{F}$ under the norm.
\[ \| f - g \|_{L^2(\mu_N)} := \sqrt{\frac{1}{N} \sum_{i=1}^{N} (f(x_i) - g(x_i))^2}. \]

Consider, without loss of generality, a particular subspace \( K_{s_0} \) of \( \mathbb{R}^d \) consisting of all vectors whose first \( s_0 \) coordinates are non-zero. Note that for any fixed number \( c \in \mathbb{R} \), the VC dimension of the set of halfspaces \( \mathcal{H} := \{ (\cdot, v) \geq c, \ v \in K_{s_0} \cap S^{s_0-1} \} \) is \( \text{VC}(\mathcal{H}) = s_0 \). Thus, by classical VC theorem, for any distinctive \( p \) points in \( \mathbb{R}^d \), the number distinctive projections from \( \mathcal{H} \) to these \( p \) points is \( \sum_{i=0}^{s_0} \binom{p}{i} \leq (p + 1)^{s_0} \). Furthermore, any set in \( \mathcal{H}' := \{ |(\cdot, v)| \geq c, \ v \in K_{s_0} \cap S^{s_0-1} \} \) is the intersection of two sets in \( \mathcal{H} \), thus, the number of distinctive projections from \( \mathcal{H}' \) to those \( p \) points is at most
\[
\binom{(p + 1)^{s_0}}{2} \leq \frac{e^2(p + 1)^{2s_0}}{4} \leq 2(p + 1)^{2s_0}.
\]
This implies \( \text{VC}(\mathcal{H}') \leq cs_0 \log(s_0) \) for some absolute constant \( c > 0 \).

Thus, the following class of indicator functions
\[
\mathcal{F}_{\delta, K_{s_0}} := \{ 1_{\{ |(\cdot, v)| \geq \delta \}}, \ v \in K_{s_0} \cap S^{s_0-1} \}
\]
has VC dimension \( \text{VC}(\mathcal{F}_{\delta, K_{s_0}}) \leq cs_0 \log(s_0) \). By Haussler’s inequality, we have the \( \varepsilon \) covering number of \( \mathcal{F}_{\delta, K_{s_0}} \) can be bounded as
\[
\mathcal{N} (\varepsilon, \mathcal{F}_{\delta, K_{s_0}}, \| \cdot \|_{L^2(\mu_N)}) \leq C_{s_0}(16e)^{cs_0 \log(s_0)} \varepsilon^{-2cs_0 \log(s_0)},
\]
where \( C > 0 \) is an absolute constant. Furthermore, \( \mathcal{F} \) is the union of \( \binom{d}{s_0} \) different subspaces \( K_{s_0} \). Thus, the \( \varepsilon \) covering number of \( \mathcal{F} \) can be bounded as
\[
\mathcal{N} (\varepsilon, \mathcal{F}, \| \cdot \|_{L^2(\mu_N)}) \leq \binom{d}{s_0} C_{s_0}(16e)^{cs_0 \log(s_0)} \varepsilon^{-2cs_0 \log(s_0)} \\
\leq (ed/s_0)^{s_0} C_{s_0}(16e)^{cs_0 \log(s_0)} \varepsilon^{-2cs_0 \log(s_0)}.
\]
Substituting this bound into (2.14) gives
\[
\mathbb{E} \left[ \sup_{v \in \mathcal{G}_{s_0} \cap S_2(1)} |R(\mathbb{R}^N, v)| \right] \leq L \sqrt{s_0 \log(ed)/N},
\]

45
for some absolute constant \( L > 0 \). By bounded difference inequality, we have

\[
\sup_{v \in \mathcal{G}_{s_0} \cap S(1)} |R(\bar{x}_1^N, v)| \leq \mathbb{E} \left[ \sup_{v \in \mathcal{G}_{s_0} \cap S(1)} |R(\bar{x}_1^N, v)| \right] + \sqrt{u/N},
\]

with probability at least \( 1 - ce^{-u} \) for some constant \( c > 0 \) any \( u \geq 0 \), which implies

\[
\inf_{v \in \mathcal{G}_{s_0} \cap S(1)} \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\{|\langle \bar{x}_i, v \rangle| \geq \delta/2\}} \geq Q - L \sqrt{s_0 \log(\text{ed})/N} - \sqrt{u/N},
\]

with probability at least \( 1 - ce^{-u} \). This implies the claim of the lemma.

Combining Lemma 2.4.5 with Lemma 2.4.1 and 2.4.2 we obtain the following lemma:

**Lemma 2.4.6.** Let \( u \geq 1 \), \( N \geq 1024u/Q^2 + c' \log \text{ed} \), where \( c' > 0 \) is an absolute constant and \( N \leq \max\{Q^{1/2}, 64L^2\}\log \text{ed} \), where \( L \) is the constant defined in Lemma 2.4.5. Let \( s_0 = \frac{N}{\log \text{ed}} \min\{\frac{Q}{\nu}, \frac{Q^2}{64L^2}\} \). then, with probability at least \( 1 - ce^{-u} \) for some absolute constant \( c > 0 \), there exists a set of indices \( I \subseteq \{1, 2, \cdots, N\} \) such that \( |I| \geq Q^4 \) and for any \( i \in I \), \( \forall v_1 \in \mathcal{G}_{s_0} \cap S(1), \forall v_2 \in S(\rho) \),

\[
|\langle \bar{x}_i, v_1 \rangle| \geq \delta/2, \quad |\langle \bar{x}_i, v_2 \rangle| \leq 32(\nu_2^2 + \nu_2 + 1)\rho/Q, \quad |\langle \bar{x}_i, \theta_* \rangle| \leq 32\nu_2\|\theta_*\|_1/Q.
\]

**Proof of Lemma 2.4.6** First of all, by Lemma 2.4.5 \( s_0 = \frac{N}{\log \text{ed}} \min\{\frac{Q}{\nu}, \frac{Q^2}{64L^2}\} \) and \( N \geq 1024u/Q^2 + c' \log \text{ed} \), we have with probability at least \( 1 - e^{-u} \),

\[
\inf_{v \in \mathcal{G}_{s_0} \cap S(1)} \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\{|\langle \bar{x}_i, v \rangle| \geq \delta/2\}} \geq \frac{Q}{2},
\]

On the other hand, by Lemma 2.4.1 and 2.4.2 we have

\[
\inf_{v \in B_{\nu}(0, \rho)} \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\{|\langle \bar{x}_i, v \rangle| \geq 32(\nu_2^2 + \nu_2 + 1)\rho/Q\}} \geq 1 - \frac{Q}{16},
\]

and

\[
\frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\{|\langle \theta_*, \bar{x}_i \rangle| \leq 32\nu_2\|\theta_*\|_1/Q\}} \geq 1 - \frac{Q}{16},
\]

with probability at least \( 1 - 2e^{-u} \). Combining the above three bounds, we have there exists a set of indices \( I \subseteq \{1, 2, \cdots, N\} \) of cardinality at least \( \frac{Q}{4} - \frac{Q}{16} - \frac{Q}{16} > \frac{Q}{4} \) such that the claim in
the lemma holds.

The following theorem bounds $r_Q$:

**Theorem 2.4.1.** Let $u \geq 1$, $D_{\min} := \min_{z \in [-c_1(\nu, \kappa)R, c_1(\nu, \kappa)R]} g''(z)$, $N \geq 1024u/Q^2 + c\log ed$ and $N \leq \max\{\nu^2Q, \frac{64L^2}{Q^2}\}d\log ed$, where $L$ is the constant defined in Lemma 2.4.5. Let $s_0 = \frac{N}{\log ed} \min\{\frac{Q}{\nu}, \frac{Q^2}{\log ed}\}$, $\rho = c\|\theta\|_1$, and $\Lambda_Q = D_{\min}\delta^2Q^2/32$, then,

$$r_Q^2 \leq \frac{C\nu}{\delta^2Q^2} \left(\sqrt{\nu} + (\sqrt{\nu} + 1)\beta \sqrt{\frac{8\log ed}{QN}}\right) \cdot \max \left\{\frac{\nu}{Q}, \frac{64L^2}{Q^2}\right\} \cdot \frac{\|\theta\|_1^2\log ed}{N},$$

with $p_Q = c_1e^{-\beta}$, where $c, c_1, C$ are absolute constants.

To prove Theorem 2.4.1, we need the following useful lower bound on the random quadratic form, which comes from [LM17b]. Lower bounds of this sort via Maurey’s empirical method originate from [Oli13].

**Lemma 2.4.7** (Lemma 2.7 of [LM17b]). Let $\Gamma : \mathbb{R}^d \rightarrow \mathbb{R}^m$. Let $s_0$ be a positive integer such that $1 < s_0 \leq d$. Assume for any $v \in \mathcal{G}_{s_0}$, $\|\Gamma v\|_2 \geq \xi\|v\|_2$ for some absolute constant $\xi > 0$. If $x \in \mathbb{R}^d$ is a non-zero vector and $\mu_j = |x_j|/\|x\|_1$, then,

$$\|\Gamma x\|_2^2 \geq \xi^2\|x\|_2^2 - \|x\|_1^2 \left(\sum_{j=1}^d \|\Gamma e_j\|_2^2 \mu_j - \xi^2\right),$$

where $\{e_j\}_{j=1}^d$ is the standard basis in $\mathbb{R}^d$.

Denote $I$ in Lemma 2.4.6 to be $I = \{i_1, \cdots, i_{|I|}\}$ and let $\bar{\Gamma} := [\bar{x}_{i_1}, \bar{x}_{i_2}, \cdots, \bar{x}_{i_{|I|}}]^T/\sqrt{N}$. We then deduce a lower bound for $I$ in view of the previous lemma, we also need an upper bound for $\max_{1 \leq j \leq d} \|\Gamma e_j\|_2^2$:

**Lemma 2.4.8.** For any $u \geq 1$ chosen by the thresholding parameter $\tau$, we have with probability at least $1 - e^{-\beta}$,

$$\max_{1 \leq j \leq d} \|\Gamma e_j\|_2^2 \leq \sqrt{\nu} + C(\sqrt{\nu} + 1)\beta \sqrt{\frac{8\log(ed)}{QN}},$$

where $C > 0$ is an absolute constant.
Proof of Lemma 2.4.8. By Bernstein’s inequality, we have for any \( t \geq 0 \),
\[
\Pr \left( \left| \frac{1}{|I|} \sum_{i \in I} \bar{x}_{ij}^2 - \mathbb{E}[\bar{x}_{ij}^2] \right| \geq C \left( \frac{2\sigma_j^2 t}{|I|} + \frac{bt}{|I|} \right) \right) \leq \exp(-t),
\]
where
\[
\sigma_j^2 = \mathbb{E} \left[ (\bar{x}_{ij}^2 - \mathbb{E}[\bar{x}_{ij}^2])^2 \right] \leq \mathbb{E}[|x_{ij}|^4] \leq \sup_{v \in \mathcal{B}} \mathbb{E}[|\langle v, x_i \rangle|^4] \leq \nu,
\]
\( |I| \geq \frac{Q}{8} N \), \( b = \tau^2 = \sqrt{\frac{N}{\log(ed)}} \), and \( \mathbb{E}[\bar{x}_{ij}^2] \leq \mathbb{E}[|\bar{x}_{ij}|^4]^{1/2} \leq \sqrt{\nu} \). Thus, it follows for any \( j \in \{1, 2, \ldots, d\} \),
\[
\frac{1}{|I|} \sum_{i \in I} \bar{x}_{ij}^2 \leq \sqrt{\nu} + C \left( \sqrt{\frac{8\nu t}{QN}} + \frac{2t}{QN \log(ed)} \right),
\]
with probability at least \( 1 - \exp(-t) \). Take a union bound over \( j \in \{1, 2, \ldots, d\} \) and let \( t = \beta \log(ed) \) give
\[
\max_{1 \leq j \leq d} \frac{1}{|I|} \sum_{i \in I} \bar{x}_{ij}^2 \leq \sqrt{\nu} + C(\sqrt{\nu} + 1)\beta \sqrt{\frac{8\log(ed)}{QN}},
\]
with probability at least \( 1 - e^{-\beta} \), for some absolute constant \( C > 0 \). This finishes the proof. \( \square \)

Proof of Theorem 2.4.1. First of all, by (2.12) and Lemma 2.4.6, we have with probability at least \( 1 - ce^{-u} \),
\[
\inf_{\theta \in B_1(\theta^*, \rho) \cap S_2(\theta^*, r)} P_N Q_{\theta - \theta^*} \geq D_{\min} \inf_{\theta \in B_1(\theta^*, \rho) \cap S_2(\theta^*, r)} \frac{1}{N} \sum_{i \in I} |\langle \bar{x}_i, \theta - \theta^* \rangle|^2.
\]
Since \( |\langle \bar{x}_i, v_1 \rangle| \geq \delta/2 \), we have
\[
\inf_{v \in \Omega_v / S_2(1)} \frac{1}{|I|} \sum_{i \in I} |\langle \bar{x}_i, v \rangle|^2 \geq \frac{\delta^2 Q}{4}.
\]
By Lemma 2.4.7 and 2.4.8, we have
\[
\inf_{\theta \in B_1(\theta^*, \rho) \cap S_2(\theta^*, r)} P_N Q_{\theta - \theta^*} \geq D_{\min} \left( \frac{\delta^2 Q^2}{8} r^2 - \frac{\rho^2}{s_0 - 1} \left( \sqrt{\nu} + C(\sqrt{\nu} + 1) \beta \sqrt{\frac{4\log(ed)}{QN}} \right) \right).
\]
Note that \( s_0 = \frac{N}{\log(ed)} \min \{ \frac{Q}{\nu}, \frac{Q^2}{16L^2} \} \), \( \rho = c \| \theta^* \|_1 \), and \( \Lambda_Q = D_{\min} \delta^2 Q^2 / 32 \). The infimum of \( r > 0 \) such that the right hand side is greater than \( \Lambda_Q r^2 = \frac{\delta^2 Q^2}{32} D_{\min} r^2 \) can be obtained by letting the
right hand side equal to \( \frac{d^2Q^2}{32}D_{\min}r^2 \) and solve for \( r \), which gives

\[
r^2 = \frac{C\|\theta_\ast\|_1^2 \log ed}{\delta^2Q^2N} \left( \sqrt{\nu} + (\sqrt{\nu} + 1)\beta \sqrt{\frac{\log(ed)}{QN}} \right) \max \left\{ \frac{\nu}{Q}, \frac{64L^2}{Q^2} \right\},
\]

for some absolute constant \( C \). It then follows from the definition of \( r_Q \) that \( r_Q \) must be bounded above by this value.

2.4.3 Applying Mendelson’s small-ball method for large \( N \)

In this section, we consider lower bounding the cardinality of the set \( \{i : |\langle \tilde{x}_i, v_1 \rangle| \geq \delta\} \), \( v_1 \in S_2(1) \) when \( N > \frac{\nu}{Q}d \log ed \). In this case, suppose Assumption 2.2.1 holds, by Lemma 2.4.4 we have for any \( v_1 \in \mathbb{R}^d \),

\[
Pr(\{ |\langle \tilde{x}_i, v_1 \rangle| \geq \|v_1\|_2\delta \} \geq Q. \tag{2.15}
\]

We have the following lemma:

Lemma 2.4.9. Let \( u \geq 1 \), \( \rho = c\|\theta_\ast\|_1 \) for some absolute constant \( c > 0 \),

\[
N \geq \frac{4c^2(2 + \sqrt{2\nu})^2(1 + 4\sqrt{\nu})^2\|\theta_\ast\|_1^2 \log ed/r^2 + 4u}{Q}
\]

and \( N > \frac{\nu}{Q}d \log ed \), then, with probability at least \( 1 - c_1e^{-u} \) for some absolute constant \( c_1 > 0 \), there exists a set of indices \( \mathcal{I} \in \{1, 2, \cdots, N\} \) such that \( |\mathcal{I}| \geq \frac{Q}{4}N \) and for any \( i \in \mathcal{I} \), \( \forall v_1 \in B_1(0, \rho/r) \cap S_2(0, 1) \), \( \forall v_2 \in S_1(\rho) \),

\[
|\langle \tilde{x}_i, v_1 \rangle| \geq \delta/2, \quad |\langle \tilde{x}_i, v_2 \rangle| \leq 32(\nu^2 + \nu + 1)\rho/Q, \quad |\langle \tilde{x}_i, \theta_\ast \rangle| \leq 32\nu\|\theta_\ast\|_1/Q.
\]

Proof of Lemma 2.5.1 The proof of this lemma almost follows from that of Lemma 1.2.1 from [Men14a], the only difference is that we need to take care of indices \( i \) such that \( |\langle \tilde{x}_i, v_2 \rangle| \leq 32(\nu^2 + \nu + 1)\rho/Q, |\langle \tilde{x}_i, \theta_\ast \rangle| \leq 32\nu\|\theta_\ast\|_1/Q \), which are Lemma 2.4.1 and 2.4.2. We consider the quantity

\[
\inf_{v \in B_1(0, \rho/r) \cap S_2(0, 1)} \frac{\delta}{N} \sum_{i=1}^{N} 1\{ |\langle \tilde{x}_i, v \rangle| \geq \delta \}.
\]

By the same argument as that of Theorem 5.4 in [Men14a] (using (2.15)), one obtains with
probability at least $1 - e^{-u/2}$,

$$\inf_{v \in B_1(0, r)} \frac{\delta}{\sqrt{N}} \sum_{i=1}^{N} 1\{\langle \tilde{x}_i, v \rangle \geq \delta \} \geq Q - \frac{2}{\sqrt{N}} \omega_N(B_1(0, r) \cap S_2(0, 1)) - \sqrt{\frac{u}{N}}.$$  

where for any $\mathcal{H} \subseteq S_2(0, 1)$,

$$\omega_N(\mathcal{H}) := \mathbb{E} \left[ \sup_{h \in \mathcal{H}} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \epsilon_i \langle \tilde{x}_i, h \rangle \right].$$

Similar to bounding term (I) is Lemma 2.4.1, one obtains

$$\frac{1}{\sqrt{N}} \omega_N(B_1(0, r) \cap S_2(0, 1)) \leq \frac{1}{\sqrt{N}} \rho \mathbb{E} \left[ \max_j \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \epsilon_i \tilde{x}_{ij} \right] \leq (2 + \sqrt{2} \nu)(1 + 4 \sqrt{e}) \sqrt{\frac{\log ed}{N}} \cdot \frac{\rho}{r} \leq c(2 + \sqrt{2} \nu)(1 + 4 \sqrt{e}) \sqrt{\frac{\log ed}{N}} \cdot \frac{\|\theta_*\|_1}{r},$$

where the last inequality follows from the fact that $\rho = c\|\theta_*\|_1$. When

$$N \geq \frac{4c^2(2 + \sqrt{2} \nu)^2(1 + 4 \sqrt{e})^2 \|\theta_*\|_1^2 \log ed/\nu^2 + 4u}{Q},$$

we have

$$\frac{1}{N} \sum_{i=1}^{N} 1\{\langle \tilde{x}_i, v \rangle \geq \delta \} \geq \frac{Q}{2},$$

with probability $1 - e^{-u/2}$. Combining this result with Lemma 2.4.1 and 2.4.2 finishes the proof.

**Theorem 2.4.2.** Let $u \geq 1$, $D_{\min} := \min_{z \in [-c_2(\nu, \kappa) R, c_2(\nu, \kappa) R]} g''(z)$, $\rho = c\|\theta_*\|_1$ for some absolute constant $c > 0$, $N \geq \frac{8u}{Q}$, and $N > \frac{\nu d \log ed}{Q}$. Suppose $\Lambda_Q = D_{\min} \delta^2 Q^2 / 4$, then,

$$r_Q^2 \leq \frac{8c^2(2 + \sqrt{2} \nu)^2(1 + 4 \sqrt{e})^2 \|\theta_*\|_1^2 \log ed}{N},$$

with $p_Q = c_1 e^{-u}$, where $c_1$ is absolute constant.

**Proof of Theorem 2.4.3** First, note that when $N \geq \frac{8u}{Q}$ and $r = r_Q$ satisfying the condition
asserted in the theorem, then,
\[ N \geq \frac{4c^2(2 + \sqrt{2})^2(1 + 4\sqrt{e})2\|\theta\|_2^2 \log ed/r^2 + 4u}{Q}. \]

For any \( \theta \in B_1(\theta_*, \rho) \cap B_2(0, r) \), let \( v = (\theta - \theta^*)/\|\theta - \theta^*\|_2 \in B_1(0, \rho/r) \cap B_2(0, 1) \) and with probability at least \( 1 - c_1e^{-u} \),
\[
\mathcal{P}_N Q_{\theta - \theta_*} \geq \frac{D_{\min}r^2}{N} \sum_{i \in I} |\langle \bar{x}_i, \theta - \theta_* \rangle|^2 \\
= \frac{D_{\min}r^2}{N} \sum_{i \in I} |\langle \bar{x}_i, v \rangle|^2 \geq \frac{D_{\min}r^2\delta^2}{N|I|} \left( \sum_{i \in I} 1\{|\langle \bar{x}_i, v \rangle| \geq \delta \} \right)^2 \geq \frac{Q^2\delta^2}{4} D_{\min}r^2,
\]
where the first inequality follows from Lemma 2.5.1 by taking the corresponding \( I \), the second from the last inequality follows from \( (1/2) \), and the last inequality follows from Lemma 2.5.1 again.

### 2.4.4 Bounding \( r_M \) via Montgomery-Smith inequality

The main objective is the following bound on \( |\mathcal{P}_N M_{\theta - \theta_*}| \):

**Lemma 2.4.10.** Suppose \( N \geq \|\theta_*\|_2^2 \log(ed) + \log(ed) \) and Assumption 2.2.1 2.2.3 hold. For any \( \beta, u, v, w > 7 \), we have with probability at least
\[ 1 - 2e^{-\beta} - 2e^{-\beta^2} - c'(u^{-q}(ed)^{-q/4} + w^{-q/4} + (ed)^{-q/4} + (eN)^{-q/5}w^{-q/5} + (eN)^{-q/4}w^{-q/4}) \]
where \( c, c' > 2 \) are absolute constants,
\[
\sup_{\theta \in B_1(\theta_*, \rho) \cap B_2(\theta_*, r)} |\mathcal{P}_N M_{\theta - \theta_*}| \leq C(\nu, \nu')(D_{\max} + 1) \left( wu^2v + w\beta^3/4 + \beta \right) \rho \sqrt{\frac{\log(ed)}{N}},
\]
where \( C(\nu, \nu') \) depends polynomially on \( \nu \) and \( \nu' \).
Proof of Lemma 2.4.10. First of all, by symmetrization inequality, it is enough to bound

\[
\sup_{\theta \in B_1(\theta^*, r) \cap B_2(\theta^*, r)} \left| \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i \left( y_i - g'( \langle \bar{x}_i, \theta_s \rangle ) \right) \langle \bar{x}_i, \theta - \theta_s \rangle \right|
\]

\[
= \sup_{\nu \in B_1(0, \rho) \cap B_2(0, r)} \left| \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i \left( y_i - g'( \langle \bar{x}_i, \theta_s \rangle ) \right) \langle \bar{x}_i, \nu \rangle \right|
\]

We define \( z := \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i \left( y_i - g'( \langle \bar{x}_i, \theta_s \rangle ) \right) \bar{x}_i \) and note that

\[
\sup_{\nu \in B_1(0, \rho) \cap B_2(0, r)} \left| \varepsilon_i \left( y_i - g'( \langle \bar{x}_i, \theta_s \rangle ) \right) \langle \bar{x}_i, \nu \rangle \right| \leq \rho \cdot \max_{j \in \{1, 2, \ldots, d \}} |z_j|.
\] (2.16)

Now for each \( |z_j| \),

\[
N |z_j| = \left| \sum_{i=1}^{N} \varepsilon_i \left( y_i - g'( \langle \bar{x}_i, \theta_s \rangle ) \right) \bar{x}_{ij} \right|
\]

\[
\leq \left| \sum_{i=1}^{N} \varepsilon_i \left( y_i - g'( \langle \bar{x}_i, \theta_s \rangle ) \right) \bar{x}_{ij} \right| + \left| \sum_{i=1}^{N} \varepsilon_i \left( g'( \langle \bar{x}_i, \theta_s \rangle ) - g'( \langle \bar{x}_i, \theta_s \rangle ) \right) \bar{x}_{ij} \right|
\]

Thus, it follows

\[
N \cdot \max_{j \in \{1, 2, \ldots, d \}} |z_j| \leq \max_{j \in \{1, 2, \ldots, d \}} \left| \sum_{i=1}^{N} \varepsilon_i \left( y_i - g'( \langle \bar{x}_i, \theta_s \rangle ) \right) \bar{x}_{ij} \right|
\]

\[
+ \max_{j \in \{1, 2, \ldots, d \}} \left| \sum_{i=1}^{N} \varepsilon_i \left( g'( \langle \bar{x}_i, \theta_s \rangle ) - g'( \langle \bar{x}_i, \theta_s \rangle ) \right) \bar{x}_{ij} \right|
\] (2.17)

Then, we need to bound the three terms on the right hand side of (2.28) separately.

1. **Bounding the terms** \( \max_{j \in \{1, 2, \ldots, d \}} \left| \sum_{i=1}^{N} \varepsilon_i \left( g'( \langle \bar{x}_i, \theta_s \rangle ) - g'( \langle \bar{x}_i, \theta_s \rangle ) \right) \bar{x}_{ij} \right| \):

Let \( \tilde{\phi}_i = g'( \langle \bar{x}_i, \theta_s \rangle ) - g'( \langle \bar{x}_i, \theta_s \rangle ) \). A crucial first step analyzing such a Rademacher sum (see, for example, [Men16, GMW16]) is to apply Montgomery-Smith inequality from, i.e. Lemma 1.1.7 conditioned on \( \bar{x}_i \), which results in

\[
\sum_{i=1}^{N} \left| e_i \tilde{\phi}_i \bar{x}_{ij} \right| \leq \sum_{i=1}^{k} \left| \tilde{\phi}_i \bar{x}_{ij} \right|^2 + \varepsilon \left( \sum_{i=k+1}^{N} \left| \tilde{\phi}_i \bar{x}_{ij} \right|^2 \right)^{1/2},
\]

with probability at least \( 1 - e^{-\varepsilon^2} \), where \( k \) is any chosen integer within \( \{0, 1, 2, \ldots, N\} \) and

52
\((\tilde{\phi}_i)^N_{i=1}, (\tilde{x}^i_{ij})^N_{i=1}\) are non-increasing rearrangements of \((\phi_i)^N_{i=1}, (x^i_{ij})^N_{i=1}\). We define the former sum to be 0 when \(k = 0\).

By Holder’s inequality, we have

\[
\left| \sum_{i=1}^{N} \xi_i \tilde{\phi}_i \tilde{x}_{ij} \right| \leq \left( \sum_{i=1}^{k} |\tilde{\phi}_i|^2 \right)^{1/2} \left( \sum_{i=1}^{k} |\tilde{x}^i_{ij}|^2 \right)^{1/2} + \nu \left( \sum_{i=k}^{N} |\tilde{\phi}_i|^{2r} \right)^{1/(2r)} \left( \sum_{i=k}^{N} |\tilde{x}^i_{ij}|^{2r'} \right)^{1/(2r')},
\]

for some positive constants \(r, r'\) such that \(\frac{1}{r} + \frac{1}{r'} = 1\). Take a union bound for all \(j \in \{1, 2, \ldots, d\}\), gives with probability at least \(1 - e^{-v^2}\),

\[
\max_{j \in \{1, 2, \ldots, d\}} \left| \sum_{i=1}^{N} \xi_i \tilde{\phi}_i \tilde{x}_{ij} \right| \leq \left( \sum_{i=1}^{k} |\tilde{\phi}_i|^2 \right)^{1/2} \max_{j \in \{1, 2, \ldots, d\}} \left( \sum_{i=1}^{k} |\tilde{x}^i_{ij}|^2 \right)^{1/2} + \nu \sqrt{\log d} \left( \sum_{i=k}^{N} |\tilde{\phi}_i|^{2r} \right)^{1/(2r)} \max_{j \in \{1, 2, \ldots, d\}} \left( \sum_{i=k}^{N} |\tilde{x}^i_{ij}|^{2r'} \right)^{1/(2r')}, \tag{2.18}
\]

where \(k\) is to be chosen.

Now we bound the four terms in (2.18) respectively.

**Lemma 2.4.11.** Let \(k = \left\lfloor \frac{c \log(\text{ed})}{\log(cN/c \log(\text{ed}))} \right\rfloor\) for some absolute constant \(c > 2\), and suppose \(N \geq \|\theta_0\|_2^2 \log(\text{ed})\), then, we have

\[
\left( \sum_{i=1}^{k} |\tilde{\phi}_i|^2 \right)^{1/2} \leq CD \max_{i} \nu_q^2 \sqrt{e \log(\text{ed})},
\]

with probability at least \(1 - c'(eN)^{-\frac{1}{2}} \left( \log(eN) \right)^2 w^{-\frac{6}{5}}\) for any \(w > 6\) and some absolute constant \(C, c' > 1\).

**Proof of Lemma 2.4.11.** First of all, using Binomial estimates, we have for any \(i\), and any positive constant \(c_i\),

\[
Pr \left( |\tilde{\phi}_i|^2 \geq c_i \|\tilde{\phi}_i\|_{L_p} \right) \leq \left( \frac{N}{i} \right)^i Pr(\|\tilde{\phi}_i\| \geq c_i \|\tilde{\phi}_i\|_{L_p})^i \leq \left( \frac{eN}{i} \right)^i Pr(\|\tilde{\phi}_i\| \geq c_i \|\tilde{\phi}_i\|_{L_p})^i \leq \left( \frac{eN}{i} \right)^i \frac{\mathbb{E} \left[ |\tilde{\phi}_i|^{p_i} \right]^i}{c_i^{p_i} \|\tilde{\phi}_i\|_{L_p}^{p_i}} = \left( \frac{eN}{i} \right)^i c_i^{-p_i},
\]

53
where we define \( \| \tilde{\phi}_i \|_{L^p} := E \left[ \left| \tilde{\phi}_i \right|^p \right]^{1/p} \) and \( p > 2 \) is a chosen positive constant. Then, we choose \( c_i := w \log\left( eN/i \right) \left( \frac{eN}{i} \right)^{-i / 2} \), which implies

\[
Pr \left( \| \tilde{\phi}_i \|_{L^p} \geq w \log\left( eN/i \right) \left( \frac{eN}{i} \right)^{-i / 2} \right) \leq \left( \frac{i}{eN} \right)^{i(\frac{p}{2} - 1)} w^{-pi} (\log(eN/i))^{pi}.
\]

Thus, it follows,

\[
\sum_{i=1}^{N} \left| \tilde{\phi}_i^2 \right| \leq \sum_{i=1}^{N} \left| \bar{\phi}_i \right| \leq \sum_{i=1}^{N} \frac{w^2}{(\log(eN/i))^2} \left( \frac{eN}{i} \right)^{-i \frac{p}{2} - 1} \| \bar{\phi}_i \|_{L^p}^2
\]

\[
\leq w^2 \| \bar{\phi}_i \|_{L^p}^2 eN \int_0^N \frac{1}{x(\log(eN) - \log x)^2} dx \leq Cw^2 \| \bar{\phi}_i \|_{L^p}^2 eN \tag{2.19}
\]

with probability at least

\[
1 - \sum_{i=1}^{N} \left( \frac{i}{eN} \right)^{i(\frac{p}{2} - 1)} w^{-pi} (\log(eN/i))^{pi}.
\]

Note that for \( w > 7 \) and \( p \) chosen to be \( p := q/5 > 3 \), the above sum is a geometrically decreasing sequence, specifically, it is easy to verify that \( \left( \frac{i}{eN} \right)^{i(\frac{p}{2} - 1)} w^{-pi} (\log(eN/i))^{pi} \leq \left( \frac{7}{6} \right)^{-p}, \forall i \in \{1, 2, 3, 4, \cdots, N\} \). Thus, it follows the above probability is at least

\[
1 - c' \left( eN \right)^{-i\left(\frac{p}{2} - 1\right)} \left( \log(eN) \right)^{pi} w^{-p},
\]

for some absolute constant \( c' > 1 \). Now, we bound the term \( \| \bar{\phi}_i \|_{L^p} \). We choose \( p = \frac{q}{5} \). Then, under the condition that \( q > 15, p = \frac{q}{5} > 3, \) and \( E[|x_{ij}|^p] < \infty, \forall i \in \{1, 2, \cdots, N\}, j \in \{1, 2, \cdots, d\} \). Furthermore, we have by Assumption \( 2.2.3 \)

\[
\left\| \bar{x}_i \right\|_{L^p} = \left\| g'(\langle \bar{x}_i, \theta_* \rangle) - g'(\langle \tilde{x}_i, \theta_* \rangle) \right\|_{L^p} \leq D_{\max} \cdot \left\| \tilde{x}_i - \bar{x}_i, \theta_* \right\|_{L^p}
\]

Note that

\[
\left\| \langle x_i - \bar{x}_i, \theta_* \rangle \right\|_{L^p} = E \left[ \left( \sum_{n=1}^{d} |x_{in} - \bar{x}_{in}| \theta_{i,j} \right)^p \right]^{1/p}
\]

\[
\leq \sum_{n=1}^{d} E \left[ |x_{in} - \bar{x}_{in}|^p \right]^{1/p} \theta_{i,j} \leq \max_n E \left[ |x_{in} - \bar{x}_{in}|^p \right]^{1/p} \theta_{i,j} 1
\]
where the first inequality follows from Minkowski’s inequality. Now, for each \( n \), we have

\[
\|x_{in} - \bar{x}_{in}\|_p \leq \|x_{in} \cdot 1_{\{\|x_{in}\| > \tau\}}\|_p \leq \mathbb{E}[\|x_{in}\|^p \cdot 1_{\{\|x_{in}\| > \tau\}}]^{1/p} \leq \mathbb{E}[\|x_{in}\|^p]^{1/5p} \Pr(|x_{in}| > \tau)^{4/5p} \leq \mathbb{E}[\|x_{in}\|^p]^{1/5p} \left(\frac{\mathbb{E}[\|x_{in}\|^p]}{\tau^{5p}}\right)^{4/5p},
\]

where the second from the last inequality follows from Holder’s inequality and the last inequality follows from Markov inequality. Thus, we obtain,

\[
\|\tilde{\sigma}_i\|_p \leq D_{\max} \|\theta_*\|_1 \max_{k} \mathbb{E}[\|x_{in}\|^p]^{1/p} \tau^{-4} \leq D_{\max} \|\theta_*\|_1 \nu_q \frac{\log(ed)}{N} \leq D_{\max} \nu_q \sqrt{\frac{\log(ed)}{N}},
\]

for some constant \( C \) and \( \tau = \left(\frac{N}{\log(ed)}\right)^{1/4} \geq \|\theta_*\|_1^{1/2} \). Overall, substituting the above bound into (2.19), we have with probability at least \( 1 - c' (eN)^{-\left(\frac{N}{2} - 1\right)} (\log(eN))^p w^{-p} \), where \( p = q/5 \),

\[
\sum_{i=1}^{k} \|\tilde{\phi}_i\|^2 \leq CD_{\max}^2 \nu_q^{10} w^2 eN \cdot \frac{\log(ed)}{N} = CD_{\max}^2 \nu_q^{10} w^2 e \log(ed),
\]

for some constant \( C > 1 \).

**Lemma 2.4.12.** Let \( k = \left[\frac{c \log(ed)}{\log(eN)}\right] \) for some absolute constant \( c > 2 \), and suppose \( N \geq \|\theta_*\|_1^2 \log(ed) \), then, we have

\[
\max_{j \in \{1, 2, \ldots, d\}} \left(\sum_{i=1}^{k} \|\tilde{x}_{ij}\|^2\right)^{1/2} \leq C \left(\nu_q^2 \log(ed) + \nu_q^2 \sqrt{\beta \log(ed)} + \sqrt{\frac{N}{\log(ed)} (\beta + \log(ed))}\right)^{1/2},
\]

with probability at least \( 1 - e^{-\beta} \) for any \( \beta > 1 \) and some constant \( C > 1 \).

**Proof of Lemma 2.4.12** First, for any set of \( k \) random variables \( x_{1j}, x_{2j}, \ldots, x_{kj} \) we have by Bernstein’s inequality,

\[
\Pr\left(\sum_{i=1}^{k} |\tilde{x}_{ij}|^2 \geq k\mathbb{E}[\tilde{x}_{ij}^2] + C \left(\sqrt{2\sigma_j^2 kt} + b_j t\right)\right) \leq e^{-t},
\]

for some constant \( C \), where \( \sigma_j^2 := \mathbb{E}[|\tilde{x}_{ij} - \mathbb{E}[\tilde{x}_{ij}]|^2] \leq \mathbb{E}[x_{ij}^4] \leq \nu_q^4 \), and \( \mathbb{E}[\tilde{x}_{ij}] \leq \mathbb{E}[x_{ij}] \) \leq \nu_q^2 \). Take a union bound over all \((N^k)\) different combinations from
Taking a union bound over all $j \in \{1, 2, \cdots, d\}$, we get

$$
\Pr \left( \max_{j \in \{1, 2, \cdots, d\}} \left| \sum_{i=1}^{k} \tilde{x}_{ij}^2 \right| \geq \kappa E \left[ \sum_{i=1}^{k} \phi_i^2 \right] + C \left( \sqrt{2 \sigma_k^2} t + b_2 t \right) \right) \leq \left( \frac{eN}{k} \right)^k e^{-t}
$$

Substituting the definition of $k = \lfloor \frac{c \log(\epsilon d)}{\log(\epsilon N/\log(\epsilon d))} \rfloor$ for some absolute constant $c > 2$, and suppose $N \geq \|\theta^*\|_2^2 \log(\epsilon d)$, then, we have with probability at least $1 - c'u^{-q/3(\epsilon d) - c/2}$, for some absolute constant $c' > 0$,

$$
\left( \sum_{i > k} \left| \tilde{\phi}_i \right|^{2r} \right)^{1/2r} \leq C D_{\max} u \epsilon N^{1/2r},
$$

for $5/4 \leq r < q/12$, any $u > 2$, and some absolute constant $C > 0$.

**Proof of Lemma 2.4.13.** Let $p = q/4$, then, $p > 3r$. Using Binomial estimates, we have for any $i > k$, and any $\alpha > 0$,

$$
\Pr \left( \left| \tilde{\phi}_i \right| > \alpha \right) \leq \binom{N}{i} \Pr (\tilde{\phi}_i > \alpha)^i \leq \binom{N}{i} \left( \frac{\mathbb{E} \left[ |\tilde{\phi}_i|^p \right]}{\alpha^p} \right)^i \leq \left( \frac{eN}{k} \mathbb{E} |\tilde{\phi}_i|^p \right)^i \alpha^p \frac{\alpha^p}{\alpha^p},
$$

where the second inequality follows from Markov inequality. We choose $\alpha = \|\tilde{\phi}\|_{L_p} u \left( \frac{eN}{k} \right)^{3/2p}$ and get

$$
\Pr \left( \left| \tilde{\phi}_i \right| > \|\tilde{\phi}\|_{L_p} u \left( \frac{eN}{k} \right)^{3/2p} \right) \leq u^{-p} \left( \frac{eN}{k} \right)^{-i/2}.
$$
Thus, it follows

\[
\Pr \left( \exists i > k, \ s.t. \ |\phi_i| > \|\phi\|_{L^p} u \left( \frac{eN}{i} \right)^{2/p} \right) \leq \sum_{i > k} u^{-p} \left( \frac{eN}{i} \right)^{-i/2} \\
\leq c' u^{-(k+1)p} \left( \frac{eN}{k+1} \right)^{-(k+1)/2} \leq c' u^{-p} \left( \frac{eN}{k+1} \right)^{-(k+1)/2},
\]

for some absolute constant \( c' > 0 \), where the second from the last inequality follows from the fact that for any \( u > 2 \), the summand is a geometrically decreasing sequence since \( N \geq i \). Plugging in \( k + 1 \geq \frac{c \log(eN)}{\log(eN/\log(ed))} \) and using the fact that \( N \geq k + 1 \) give

\[
\left( \frac{eN}{k+1} \right)^{-(k+1)/2} \leq \exp \left( -\frac{c \log(ed)}{2 \log(eN/\log(ed))} \log \left( \frac{eN}{c \log(ed)} \right) \right) \leq \exp(-c \log(ed)/2) = (ed)^{-c/2},
\]

Thus, it follows with probability at least \( 1 - c_0 u^{-p}(ed)^{-c} \), we have

\[
\left( \sum_{i > k} |\phi_i|^2 r^{1/2r} \right)^{1/2r} \leq \|\bar{\phi}\|_{L^p} u \left( \sum_{i > k} \left( \frac{eN}{i} \right)^{3r/p} \right)^{1/2r}.
\]

(2.20)

Since \( p = q/4 > 3r \), it follows

\[
\sum_{i > k} \left( \frac{1}{i} \right)^{3r/p} \leq \int_0^N \left( \frac{1}{x} \right)^{3r/p} dx = \frac{1}{1 - 3r/p} N^{1 - \frac{3r}{p}}.
\]

Thus, with probability at least \( 1 - c_0 u^{-q/3}(ed)^{-c/2} \),

\[
\left( \sum_{i > k} |\phi_i|^2 r^{2r} \right)^{1/2r} \leq C \|\bar{\phi}\|_{L^p} u N^{1/2r},
\]

(2.21)

for some constant \( C \). It remains to bound \( \|\bar{\phi}\|_{L^p} \). By Assumption 2.2.3

\[
\|\bar{\phi}\|_{L^p} = \|g'((x_i, \theta_*)) - g'(\langle \bar{x}_i, \theta_* \rangle)\|_{L^p} \leq D_{\max} \cdot \|\langle x_i - \bar{x}_i, \theta_* \rangle\|_{L^p}
\]
Note that
\[
\|\langle x_i - \bar{x}, \theta_\ast \rangle \|_{L_p} = \left[ \frac{1}{p} \sum_{n=1}^{d} (x_{in} - \bar{x}_{in})^p \right]^{1/p} \leq \max_n \frac{1}{p} \sum_{n=1}^{d} (x_{in} - \bar{x}_{in})^p |\theta_\ast, j| \leq \max_n \left( \frac{1}{p} \sum_{n=1}^{d} (x_{in} - \bar{x}_{in})^p \right)^{1/p}.
\]
where the first inequality follows from Minkowski’s inequality. Now, for each \(n\), we have
\[
\|x_{in} - \bar{x}_{in}\|_{L_p} \leq \|x_{in} \cdot 1_{\{|x_{in}| > \tau\}}\|_{L_p} \leq \mathbb{E} [\left| x_{in} \right|^{3p} \cdot 1_{\{|x_{in}| > \tau\}}^{2/3p}] \leq \mathbb{E} [\left| x_{in} \right|^{3p}]^{1/3p} \left( \frac{\mathbb{E} [\left| x_{in} \right|^{3p}]}{\tau^{3p}} \right)^{2/3p},
\]
where the second from the last inequality follows from Holder’s inequality and the last inequality follows from Markov inequality. Thus, we obtain,
\[
\|\phi_i\|_{L_p} \leq \max_n \left( \frac{1}{p} \sum_{n=1}^{d} (x_{in} - \bar{x}_{in})^p \right)^{1/p} \leq C \max_n \nu_q^{3/2} \sqrt{\log(\epsilon d^{1/2})},
\]
where \(C\) and \(\tau = \left( \frac{N}{\log(\epsilon d)} \right)^{1/4} \geq \|\theta_\ast\|^{1/2} \). Combining this bound with (2.21) finishes the proof.

**Lemma 2.4.14.** Let \(k = \left\lfloor \frac{c \log(\epsilon d)}{\log(eN/c \log(\epsilon d))} \right\rfloor\) for some absolute constant \(c > 2\), and suppose \(N \geq c s \log(\epsilon d)\), then, we have with probability at least \(1 - c' \epsilon^{-q} (\epsilon d)^{-1}\), for some absolute constant \(c' > 0\).

\[
\max_{j \in \{1, 2, \ldots, d\}} \left( \sum_{i > k} |\bar{x}_{ij}^{r'}|^{2r'} \right)^{1/2r'} \leq C \nu_q N^{1/2r'},
\]
for some constant absolute constant \(C > 0\) and \(r' \in (\frac{q}{q - 12}, 5]\).

**Proof.** First, following the same procedure as that of Lemma 2.4.13 up to (2.20), with \(p = q\), we have with probability at least \(1 - c' \epsilon^{-q} (\epsilon d)^{-c/2}\),
\[
\left( \sum_{i > k} |\bar{x}_{ij}|^{r'} \right)^{1/2r'} \leq \|\bar{x}_{ij}\|_{L_q} u \left( \sum_{i > k} \left( \frac{e N}{i} \right)^{3r'/q} \right)^{1/2r'}.
\]
Note that \(\|\bar{x}_{ij}\|_{L_q} \leq \|x_{ij}\|_{L_q} \leq \nu_q\) by the assumption and \(r' \in (\frac{q}{q - 12}, 5]\), thus, \(3r'/q < 1\) and we
Finally, taking a union bound over all $j \in \{1, 2, \ldots, d\}$ finishes the proof.

Finally, substituting Lemma 2.4.11, 2.4.12, 2.4.13, 2.4.14 into (2.18) with $r = 5/4, r' = 5$ gives with probability at least $1 - e^{-\beta - e^{-v^2} - c' \left( u^{-q}(ed)^{-(q/2)} + u^{-q/4}(ed)^{-c'/2} + e^{-\frac{q}{10} N - \frac{q}{10} + (\log(eN))^{q/5} w^{-q/5}} \right)}$,

\[
\max_{j \in \{1, 2, \ldots, d\}} \left| \sum_{i=1}^{N} \varepsilon_i \tilde{\phi}_j x_{ij} \right| \leq CD_{\max} \left( \nu_0^5 + \nu_0^3 + \nu_0 \right) w \left( \log(ed) \beta^{1/4} + N^{1/4} (\log(ed))^{3/4} \beta^{1/2} \right) + v u^2 \sqrt{N \log d}. \tag{2.22}
\]

2. Bounding the terms $\max_{j \in \{1, 2, \ldots, d\}} \left| \sum_{i=1}^{N} \varepsilon_i (y_i - g'((\bar{x}_i, \theta_*))) \bar{x}_{ij} \right|.$

The proving techniques in this part are essentially the same as those of the last part but with a slight change of exponents when applying Holder’s inequality adapting to the moment condition of the term $y_i - g'((\bar{x}_i, \theta_*))$. For simplicity of notations, let

\[\xi_i := y_i - g'((\bar{x}_i, \theta_*)).\]

Similar as before, one can employ the inequality from [MS90a], conditioned on $\bar{x}_i$, which results in

\[
\left| \sum_{i=1}^{N} \varepsilon_i \xi_i \bar{x}_{ij} \right| \leq \left( \sum_{i=1}^{N} \left| \xi_i \right|^4 \right)^{1/4} \left( \sum_{i=1}^{N} \left| \xi_i \bar{x}_{ij} \right|^{4/3} \right)^{3/4} + v \left( \sum_{i=1}^{N} \left| \xi_i \right|^{2r} \right)^{1/(2r)} \left( \sum_{i>k} \left| \bar{x}_{ij} \right|^{2r} \right)^{1/(2r')},
\]

for some positive exponents $r, r'$ such that $\frac{1}{r} + \frac{1}{r'} = 1$. Take a union bound for all $j \in \{1, 2, \ldots, d\}$,
First of all, by Markov inequality, Proof of Lemma 2.4.15. Thus, it follows

Choosing $c_i = w(eN/i)^{1/4}(log(eN/i))^{1/2}$ gives

Thus, it follows

with probability at least

Again, our goal is to bound the four terms in (2.23) separately.

Lemma 2.4.15. Let $k = \lceil \frac{c \log(ed)}{\log(c \log(ed))} \rceil$ for some absolute constant $c > 2$, and suppose $N \geq \|\theta_*\|_r^2 \log(ed)$, then, we have

with probability at least $1 - c'(eN)^{-1}(log(eN))^\frac{1}{2} w^{-q}$ for any $w > 4$ and some absolute constant $C, c' > 1$, where $\|\xi_i\|_{L_{q'}} \leq \nu_{q'}$ with $q' > 5$ is defined in Assumption 2.2.1.

Proof of Lemma 2.4.15. First of all, by Markov inequality,

Choosing $c_i = w(eN/i)^{1/4}(log(eN/i))^{1/2}$ gives

Thus, it follows

with probability at least

1 - $\sum_{i=1}^{N} \left( \frac{i}{eN} \right)^{i} \left( \frac{1}{2} \right)^{(q'-1)} w^{-q'} \left( \log \left( \frac{eN}{i} \right) \right)^{q'-i}$. 

60
Since for any $w > 4$ and $q' > 5$, the above summand is a geometrically decreasing sequence. Specifically, it is easy to show that $(\frac{q'}{eN})^{(\frac{q-1}{w})} w^{-q'} \log (\frac{eN}{i})^{\frac{q'}{2}} < (4/\sqrt{10})^{-q'}$, \forall i \in \{1,2,\cdots, N\}. Thus, it follows the probability is at least

$$1 - c' (eN)^{-(\frac{q'}{w})} w^{-q'} (\log (eN))^{\frac{q'}{2}}$$

for some absolute constant $c' > 0$.

**Lemma 2.4.16.** Let $k = \lfloor \frac{c \log (ed)}{\log (\log (ed))} \rfloor$ for some absolute constant $c > 2$, then, we have

$$\max_{j \in \{1,2,\cdots,d\}} \left( \frac{k}{3/4} \sum_{i=1}^{k} |x_{ij}|^{4/3} \right)^{3/4} \leq C \left( \nu_q^{4/3} \log (ed) + \nu_q^{4/3} \sqrt{\beta} \log (ed) + \left( \frac{N}{\log (ed)} \right)^{1/3} (\beta + \log (ed)) \right)^{3/4},$$

with probability at least $1 - e^{-\beta}$ for any $\beta > 1$ and some constant $C > 1$.

**Proof of Lemma 2.4.16.** First, for any set of $k$ random variables $x_{1j}, x_{2j}, \cdots, x_{kj}$ we have by Bernstein’s inequality,

$$Pr \left( \sum_{i=1}^{k} |x_{ij}|^{4/3} \geq kE [|x_{ij}|^{4/3}] + C \left( \sqrt{2\sigma_2^2 kt + b_2 t} \right) \right) \leq e^{-t},$$

for some constant $C$, where $\sigma_2^2 := E \left( |x_{ij}|^{4/3} - E[|x_{ij}|^{4/3}] \right)^2 \leq E[|x_{ij}|^{8/3}] \leq \nu_q^{8/3}$, $b_2 := (N/\log (ed))^{1/3}$ and $E[|x_{ij}|^{4/3}] \leq E[|x_{ij}|^{4/3}] \leq \nu_q^{4/3}$. Take a union bound over all $(N \choose k)$ different combinations from $x_{1j}, x_{2j}, \cdots, x_{Nj}$, we obtain,

$$Pr \left( \sum_{i=1}^{k} |x_{ij}|^{4/3} \geq kE [|x_{ij}|^{4/3}] + C \left( \sqrt{2\sigma_2^2 kt + b_2 t} \right) \right) \leq (N \choose k) e^{-t} \leq \left( \frac{eN}{k} \right)^{k} e^{-t}.$$ 

Taking a union bound over all $j \in \{1,2,\cdots,d\}$, we get

$$Pr \left( \max_{j \in \{1,2,\cdots,d\}} \sum_{i=1}^{k} |x_{ij}|^{4/3} \geq kE [|x_{ij}|^{4/3}] + C \left( \sqrt{2\sigma_2^2 kt + b_2 t} \right) \right) \leq d \left( \frac{eN}{k} \right)^{k} e^{-t}$$

61
Substituting the definition of \( k = \left\lfloor \frac{c \log(eN)}{\log(eN/\log(ed))} \right\rfloor \leq \frac{c \log(eN)}{\log(eN/\log(ed))} \), we get
\[
d \left( \frac{eN}{k} \right)^k e^{-t} = \exp \left( -t + k \log(eN/k) + \log d \right)
\]
\[
\leq \exp \left( -t + \frac{c \log(ed)}{\log(eN/c \log(ed))} \log \left( \frac{eN}{c \log(ed)} \right) + \log \left( \frac{eN}{c \log(ed)} \right) \right) + \log d
\]
\[
\leq \exp(-t + (2c + 1) \log(ed)).
\]

Setting \( \beta = t - (2c + 1) \log(ed) \) and rearranging the terms gives the claim.

**Lemma 2.4.17.** Let \( k = \left\lfloor \frac{c \log(ed)}{\log(eN/\log(ed))} \right\rfloor \) for some absolute constant \( c > 2 \), then, we have with probability at least \( 1 - c' u^{-q'} (ed)^{-c'/2} \), for some absolute constant \( c' > 0 \),
\[
\left( \sum_{i > k} \left| \xi_i \right|^{2r} \right)^{1/2r} \leq C u q' N^{1/2r},
\]
for \( r \leq 5/4 \), any \( u > 2 \), and some absolute constant \( C > 0 \).

**Proof.** Following from the same proof as that of Lemma 2.4.13 up to (2.20) with \( p = q' \), we have with probability at least \( 1 - c_0 u^{-q'} (ed)^{-c'/2} \),
\[
\left( \sum_{i > k} \left| \xi_i \right|^{2r} \right)^{1/2r} \leq \left\| \xi \right\|_{L_q u} \left( \sum_{i > k} \left( \frac{eN}{i} \right)^{3r/q'} \right)^{1/2r}.
\]
(2.24)

Since \( q' > 5 \geq 4r \) by assumption, it follows,
\[
\sum_{i > k} \left( \frac{1}{i} \right)^{3r/q'} \leq \int_0^N \left( \frac{1}{x} \right)^{3r/q'} dx = \frac{1}{1 - 3r/q'} N^{1 - 3r/q'},
\]
which implies the claim.

Also, by Lemma 2.4.14, we have with probability at least \( 1 - c' u^{-q} (ed)^{-\left( \frac{q'}{q} - 1 \right)} \), for some absolute constant \( c' > 0 \),
\[
\max_{j \in \{1,2,\ldots,d\}} \left( \sum_{i > k} \left| \xi_{ij} \right|^{2r'} \right)^{1/2r'} \leq C u q' N^{1/2r'},
\]
(2.25)
for some constant absolute constant \( C > 0 \) and \( r' \in \left( \frac{q}{q - 12}, 5 \right] \).
Overall, substituting Lemma 2.4.15, 2.4.16, 2.4.17, and (2.25) into (2.23) with $r = 5/4, r' = 5$ gives with probability at least

$$1 - e^{-\beta} - e^{-v^2} - c' \left( (eN)^{-\left(\frac{q}{4}\right)}(\log(eN))^{q'/2}w^{-q'} + u^{-q}(ed)^{-\left(\frac{q}{4}\right)} + u^{-q'}(ed)^{-\left(\frac{q'}{4}\right)} \right),$$

the following holds

$$\max_{j \in \{1, 2, \ldots, d\}} \left| \sum_{i=1}^{N} \xi_{ij}x_{ij} \right| \leq C\nu \left( vu^2\nu q N^{1/2}(log(ed))^{1/2} + w\nu q (log(ed))^{3/4}N^{1/4} \right.$$

$$\left. + w\nu q \beta^{3/8} N^{1/4}(log(ed))^{3/4} + w\beta^{3/4} N^{1/2}(log(ed))^{1/2} \right) \quad (2.26)$$

Overall, substituting the bounds (2.22) and (2.26) into (2.28) gives

$$N \cdot \max_{j \in \{1, 2, \ldots, d\}} |z_j| \leq CD_{\max}(\nu q^5 + \nu q^3 + \nu q) + vu^2 \sqrt{N \log d}$$

$$+ C\nu q (\nu q + 1) (vu^2 + w + w\beta^{3/8} + w\beta^{3/4}) \left( \sqrt{\beta N \log(ed)} + \beta N^{1/4} (log(ed))^{3/4} \right),$$

with probability at least

$$1 - 2e^{-\beta} - 2e^{-v^2} - c' \left( u^{-q}(ed)^{-\left(\frac{q}{4}\right)} + (u^{-q/4} + u^{-q'}) (ed)^{-c/2} \right.$$

$$\left. + (eN)^{-\left(\frac{q}{4}\right)}(\log(eN))^{q'/2}w^{-q'} + (eN)^{-\left(\frac{q}{4}\right)}(\log(eN))^{q'/2}w^{-q'} \right).$$

This implies the claim when combining (2.16) and the fact that $N \geq \log(ed)$. \qed

The following lemma gives a bound on $r_M$ in terms of $\rho$.

**Lemma 2.4.18.** Suppose $N \geq \|\theta_*\|_1^2 \log(ed) + \log(ed)$, $\Lambda_M = \frac{\delta^2Q^2}{128}D_{\min}$ and Assumption 2.2.1 and Assumption 2.2.3 hold. For any $\beta, u, v, w > 7$, we have

$$r_M^2 \leq \frac{C(\nu q, \nu q', \kappa, \nu)}{D_{\min}} \left( vu^2v + w\beta^{3/4} \right) \rho \sqrt{\frac{\log(ed)}{N}},$$

for any $m \in \{1, 2, \ldots, d\}$, where $C(\nu q, \nu q', \kappa, \nu)$ depends polynomially on $\nu q, \nu q', \kappa$ and $\nu$, when
\[ p_M = 2e^{-\beta} + 2e^{-v^2} + c'(u-q(ed)^{-\frac{\beta}{2}}) + (u-q/4 + u)^{-\frac{\beta}{2}} \]
\[ + (eN)^{-\frac{\beta}{4}} (\log(eN))^{\frac{\beta}{4}} w^{-\frac{\beta}{2}} + (eN)^{-\frac{\beta}{4}} (\log(eN))^{\frac{\beta}{4}} w^{-\frac{\beta}{2}} \],

where \( c, c' > 2 \) are absolute constants.

**Proof of Lemma 2.4.18** Since \( \Lambda M = \delta^2 Q^2 / 128 D_{\min} \), the infimum of \( r > 0 \) such that the right hand side of Lemma 2.4.10 is less than \( \Lambda M \) can be achieved by setting the right hand side equal to \( \Lambda M = \delta^2 Q^2 / 128 D_{\min} r^2 \), which gives,
\[ \Lambda M r^2 = \delta^2 Q^2 / 128 D_{\min} r^2 = C(\nu_q, \nu_q')(D_{\max} + 1) \left( w u^2 v + w^3 \right) \rho \sqrt{\log(eN)} / N, \]

which implies the claim.

\[ \square \]

### 2.4.5 Bounding the radius \( r_V \)

**Lemma 2.4.19.** Suppose \( N \geq \|\theta_*\|^2_1 \log(ed) \) and \( \Lambda V = D_{\min} \delta^2 Q^2 / 128 \), then,
\[ r_V^2 \leq \frac{128 D_{\max} \nu_q^6}{D_{\min} \delta^2 Q^2} \rho \sqrt{\log(ed)} / N, \]

**Proof of Lemma 2.4.19** First of all,
\[ \sup_{\theta \in B_2(\theta_*, r) \cap B_1(\theta_*, \rho)} |V_{\theta \cdot \theta_*}| := \sup_{\theta \in B_2(\theta_*, r) \cap B_1(\theta_*, \rho)} \mathbb{E}[|y - g'(\langle x, \theta_* \rangle)| \langle x, v \rangle]. \]

For each \( v \), we have
\[ \mathbb{E}[(y - g'(\langle x, \theta_* \rangle)) \langle x, v \rangle] = \mathbb{E}[(y - g'(\langle x, \theta_* \rangle)) \langle x, v \rangle] + |\mathbb{E}[(g'(\langle x, \theta_* \rangle) - g'(\langle x, \theta_* \rangle)) \langle x, v \rangle]| \]
\[ \leq \rho \mathbb{E}[(g'(\langle x, \theta_* \rangle) - g'(\langle x, \theta_* \rangle)) \langle x, v \rangle]|_{\infty}, \]

where we use the fact that the conditional expectation
\[ \mathbb{E}[y - g'(\langle x, \theta_* \rangle) \mid x] = 0. \]
Note that for any $j \in \{1, 2, \ldots, d\}$, by Cauchy-Schwarz inequality,

$$|\mathbb{E}[(g'((\mathbf{x}, \theta_*)) - g'((\mathbf{x}, \theta_*)))\bar{x}_j]| \leq \mathbb{E}[(g'((\mathbf{x}, \theta_*)) - g'((\mathbf{x}, \theta_*)))^2]^{1/2} \mathbb{E}[ar{x}_j^2]^{1/2} \leq D_{\max} \mathbb{E}[(\mathbf{x} - \langle \mathbf{x}, \theta_* \rangle)^2]^{1/2} \mathbb{E}[ar{x}_j^2]^{1/2} \leq D_{\max} \sum_{i=1}^d \mathbb{E}[(x_i - \bar{x}_i)^2]^{1/2} \mathbb{E}[ar{x}_j^2]^{1/2} \leq D_{\max} \|\theta_*\|_1 \max_i \mathbb{E}[(x_i - \bar{x}_i)^2]^{1/2} \mathbb{E}[ar{x}_j^2]^{1/2},$$

where the second inequality follows from Assumption 2.2.3 and the third inequality follows from Minkowski’s inequality. Now, for each $i$, we have

$$\mathbb{E}[(x_i - \bar{x}_i)^2]^{1/2} \leq \mathbb{E}[x_i^2 1_{|x_i| \geq \tau}]^{1/2} \leq \mathbb{E}[x_i^{10}]^{1/10} \mathbb{P}(x_i \geq \tau)^{2/5},$$

where

$$\mathbb{P}(x_i \geq \tau) \leq \frac{\mathbb{E}[x_i^{10}]}{\tau^{10}} \leq \frac{\mathbb{E}[|x_i|^{10}]}{N} \left( \frac{\log ed}{N} \right)^{5/2}.$$

Thus,

$$\mathbb{E}[(x_i - \bar{x}_i)^2]^{1/2} \leq \mathbb{E}[x_i^{10}]^{1/2} \frac{\log ed}{N} \leq \nu_q^5 \frac{\log ed}{N},$$

and we have

$$|\mathbb{E}[(g'((\mathbf{x}, \theta_*)) - g'((\mathbf{x}, \theta_*)))\bar{x}_j]| \leq D_{\max} \|\theta_*\|_1 \nu_q^6 \frac{\log ed}{N} \leq D_{\max} \nu_q^6 \sqrt{\frac{\log ed}{N}},$$

where we use the assumption that $N \geq \|\theta_*\|_1^2 \log ed$. Overall, we get

$$\sup_{\theta \in B_2(\theta_*, r) \cap \mathcal{B}_\varepsilon(\theta_*, \rho)} |\mathcal{V}_d - \mathcal{V}_d| \leq \nu_q^6 D_{\max} \rho \sqrt{\frac{\log(ed)}{N}}.$$  

Since $\Lambda_V = \frac{\delta^2 Q^2}{128} D_{\min}$, let

$$\frac{\delta^2 Q^2}{128} D_{\min} r^2 = \nu_q^6 D_{\max} \rho \sqrt{\frac{\log(ed)}{N}},$$

65
which results in
\[ r^2 = \frac{128D_{\text{max}}v_q^6}{D_{\text{min}}\delta^2 Q^2} \rho \sqrt{\frac{\log(ed)}{N}}, \]
and \( r^2_V \) must be bounded above by this value. 

### 2.4.6 Putting everything together

**Proof of Theorem 2.2.3** We choose \( \Lambda_Q = \frac{\delta^2 Q^2}{128} D_{\text{min}} \), \( \Lambda_M = \frac{\delta^2 Q^2}{128} D_{\text{min}} \) and \( \Lambda_V = \frac{\delta^2 Q^2}{128} D_{\text{min}} \).

Then, \( \Lambda_Q > \Lambda_M + \Lambda_V \). By Theorem 2.4.1 and 2.4.2

\[ r^2_Q \leq C_1(\nu_q, \nu_q', \nu, \kappa) \beta \frac{\| \theta_* \|_2^2 \log ed}{N}, \]

with \( p_Q = c_1 e^{-\beta} \), when \( N \geq 1024 \frac{\beta}{Q^2} + \beta^2 \frac{\log d}{Q^2} \). By Lemma 2.4.18 we have

\[ r^2_M \leq C_2(\nu_q, \nu_q', \kappa, \nu) \frac{(D_{\text{max}} + 1) (wu^2v + w\beta^{3/4}) \rho}{D_{\text{min}}} \sqrt{\frac{\log(ed)}{N}}, \]

with

\[
 p_M = 2e^{-\beta} + 2e^{-v^2} + c' \left( u^{-q}(ed)^{-\frac{q}{2}} + (u^{-q/4} + u^{-q})(ed)^{-c'/2} 
 + (eN)^{-\frac{q}{2}} \log(ed) \right)^{q/5} w^{-q/5} + (eN)^{-\frac{q'}{4}} \log(ed)^{q'/2} w^{-q'}/2, 
\]

when \( N \geq \| \theta_* \|_2^2 \log(ed) + \log(ed) \). Finally, by Lemma 2.4.19

\[ r^2_V \leq \frac{128D_{\text{max}}v_q^6}{D_{\text{min}}\delta^2 Q^2} \rho \sqrt{\frac{\log(ed)}{N}}, \]

when \( N \geq \| \theta_* \|_2^2 \log(ed) \). Thus, when \( N \geq c_0(\| \theta_* \|_2^2 \log(ed) + \log(ed) + \frac{\beta}{Q^2}) \) for some absolute constant \( c_0 \),

\[ r(\rho)^2 \leq C_3(\nu_q, \nu_q', \kappa, \nu) \frac{(D_{\text{max}} + 1) (wu^2v + w\beta^{3/4} + \beta) \rho}{D_{\text{min}}} \sqrt{\frac{\log(ed)}{N}}, \]

Now, we choose \( \rho = c \| \theta_* \|_1 \) for some \( c > 4 \) and

\[ \lambda \geq \frac{C_3(\nu_q, \nu_q', \kappa, \nu) (D_{\text{max}} + 1) (wu^2v + w\beta^{3/4} + \beta) \sqrt{\frac{\log(ed)}{N}}}{D_{\text{min}}}, \]

66
By Theorem 2.3.1 we have the estimator satisfies
\[ \|\widehat{\theta}_N - \theta_*\|_2^2 \leq \frac{C_3(\nu_q, \nu_{q'q}, \kappa, \nu)(D_{\max} + 1) \left(wu^2v + w\beta^3/4 + \beta\right)}{D_{\min}} \|\theta_*\|_1 \sqrt{\frac{\log(ed)}{N}}, \]
and we finish the proof. \[ \square \]

2.5 Proof of Theorem 2.2.5: Computing Local Complexities

In this section, we prove Theorem 2.2.5 in a similar manner as that of Theorem 2.2.2. Building upon previous intermediate results, the proof will be relatively simpler.

2.5.1 Bounding radius \( r_Q \)

**Lemma 2.5.1.** Let \( u \geq 1, N \geq 1024u/Q^2 + (\frac{\nu}{Q} + 64L^2)s_0 \log ed \) where \( L > 0 \) is the absolute constant defined in Lemma 2.4.3 then, with probability at least \( 1 - ce^{-u} \) for some absolute constant \( c > 0 \), there exists a set of indices \( I \in \{1, 2, \cdots, N\} \) such that \( |I| \geq \frac{Q}{4}N \) and for any \( i \in I \), \( \forall v_1 \in G_{s_0} \cap S_2(1) \), \( \forall v_2 \in S_1(\rho) \),
\[ |\langle \bar{x}_i, v_1 \rangle| \geq \delta/2, \quad |\langle \bar{x}_i, v_2 \rangle| \leq 32(\nu_q^2 + \nu_q + 1)\rho/Q, \quad |\langle \theta^*, \bar{x}_i \rangle| \leq 32\nu_q\|\theta_*\|_1/Q. \]

**Proof of Lemma 2.5.1** First of all, by Lemma 2.4.5 \( N \geq 1024u/Q^2 + (\frac{\nu}{Q} + 64L^2)s_0 \log ed \), we have with probability at least \( 1 - e^{-u} \),
\[ \inf_{v \in G_{s_0} \cap S_2(1)} \frac{1}{N} \sum_{i=1}^{N} 1\{(|\langle \bar{x}_i, v \rangle|) \geq \frac{Q}{4}\} \geq \frac{Q}{4}. \]
On the other hand, by Lemma 2.4.1 and 2.4.2 we have
\[ \inf_{v \in B_{\Psi}(0, \rho)} \frac{1}{N} \sum_{i=1}^{N} 1\{|\langle v, \bar{x}_i \rangle| \geq 32(\nu_q^2 + \nu_q + 1)\rho/Q\} \geq 1 - \frac{Q}{16}, \]
and
\[ \frac{1}{N} \sum_{i=1}^{N} 1\{|\langle \theta^*, \bar{x}_i \rangle| \leq 32\nu_q\|\theta_*\|_1/Q\} \geq 1 - \frac{Q}{16}. \]
with probability at least $1 - 2e^{-u}$. Combining the above three bounds, we have there exists a set of indices $I \subseteq \{1, 2, \cdots, N\}$ of cardinality at least $\frac{Q}{2} - \frac{Q}{16} - \frac{Q}{16} > \frac{Q}{4}$ such that the claim in the lemma holds.

**Theorem 2.5.1.** Suppose $N \geq C_0 \left( \frac{s_0}{\sqrt{Q}} + \frac{\nu + 1}{\nu} \right) \beta^2 \log(ed) + \frac{\nu}{Q} s_0 \log(ed)$ for some absolute constant $C_0 > 0$, and $s_0 = \frac{\sqrt{Q}}{\delta \sqrt{Q^2}} s \leq d$ and $\Lambda_Q = D_{\min} \delta^2 Q^2 / 16$, then,

$$r_Q \leq \frac{8}{\sqrt{s}} \rho$$

when taking $p_Q = ce^{-\beta}$ in the definition of $r_Q$ for $\beta \geq 1$.

**Proof of Theorem 2.5.1.** First of all, recall that $\tilde{\Gamma} := [\tilde{x}_{i_1}, \tilde{x}_{i_2}, \cdots, \tilde{x}_{i_{\|I\|}}]^T / \sqrt{N}$. By Lemma 2.5.1 and the assumption $N \geq C_0 \frac{s_0}{\sqrt{Q}} \beta^2 \log(ed) + \frac{\nu}{Q} s_0 \log(ed)$ for some large enough absolute constant $C_0$, we have

$$\inf_{v \in \mathcal{V} \cap S^d} \left\| \tilde{\Gamma} v \right\|_2 \geq \frac{\delta^2 Q^2}{8},$$

with probability at least $1 - e^{-\beta}$. Thus, it follows from Lemma 2.4.7 and 2.4.8 that

$$\inf_{\theta \in B_1(\theta, \rho) \cap S_2(\theta, r)} P_N Q_{\theta - \theta_*} \geq D_{\min} \left( \frac{\delta^2 Q^2}{8} r^2 - \frac{\rho^2}{s_0 - 1} \left( \sqrt{\nu} + C \left( \sqrt{\nu} + 1 \right) \beta \sqrt{\frac{\log(ed)}{N}} \right) \right),$$

where $C > 0$ is an absolute constant. By assumption that $N \geq C_0 \frac{\nu + 1}{\nu} \beta^2 \log(ed)$ for some $C_0$ large enough, then,

$$\inf_{\theta \in B_1(\theta, \rho) \cap S_2(\theta, r)} P_N Q_{\theta - \theta_*} \geq D_{\min} \left( \frac{\delta^2 Q^2}{8} r^2 - \frac{2 \sqrt{\nu}}{s_0 - 1} \rho^2 \right) \geq D_{\min} \left( \frac{\delta^2 Q^2}{8} r^2 - \frac{4 \sqrt{\nu}}{s_0} \rho^2 \right).$$

Using the assumption that $s_0 = \frac{\sqrt{s}}{\delta \sqrt{Q^2}} s$, we obtain

$$\inf_{\theta \in B_1(\theta, \rho) \cap S_2(\theta, r)} P_N Q_{\theta - \theta_*} \geq \frac{\delta^2 Q^2 D_{\min}}{8} \left( r^2 - \frac{32 \rho^2}{s} \right).$$

The infimum of $r > 0$ such that the right hand side is greater than $\frac{\delta^2 Q^2 D_{\min}}{16} r^2$ can be obtained by letting the right hand side equal to $\frac{\delta^2 Q^2 D_{\min}}{16} r^2$ and solve for $r$, which gives $r = \frac{8}{\sqrt{s}} \rho$. It then follows from the definition of $r_Q$ that $r_Q$ must be bounded above by this value. \qed
2.5.2 Bounding $r_M$ and $r_V$

The main objective is the following bound on $|\mathcal{P}_{N\mathcal{M}_{\theta_1}}|$:

**Lemma 2.5.2.** Suppose $N \geq cs \log(ed)$ for some absolute constant $c > 1$ and Assumption 2.2.1 holds. For any $\beta, u, v, w > 7$, we have with probability at least

$$1 - 2e^{-\beta} - 2e^{-v^2} - c'(u^{-q}(ed)^{-1} + (u^{-q/4} + w^{-q'})(ed)^{-\varepsilon/2}$$

$$+(e^N)^{-\gamma+1}(\log(eN))^{q/5}w^{-q/5} + (e^N)^{-(\gamma+1)(\log(eN))^{q'/2}}w^{-q'},$$

where $c, c' > 2$ are absolute constants,

$$\sup_{\theta \in B_1(0,\rho) \cap B_2(\theta_1, r)} |\mathcal{P}_{N\mathcal{M}_{\theta_1}}| \leq C(\nu_q, \nu_{q'}) (D_{\max} + 1) (wu^2 v + w\beta^{3/4} + \beta) (r\sqrt{m} + \rho) \sqrt{\frac{\log(ed)}{N}},$$

for any $m \in \{1, 2, \cdots, d\}$, where $C(\nu_q, \nu_{q'})$ depends polynomially on $\nu_q$ and $\nu_{q'}$.

**Proof of Lemma 2.5.2.** First of all, by symmetrization inequality, it is enough to bound

$$\sup_{\theta \in B_1(0,\rho) \cap B_2(\theta_1, r)} \left| \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i (y_i - g'((\bar{x}_i, \theta_1))) (\bar{x}_i, \theta - \theta_1) \right| = \sup_{\nu \in B_1(0,\rho) \cap B_2(0, r)} \left| \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i (y_i - g'((\bar{x}_i, \theta_1))) (\bar{x}_i, \nu) \right|$$

We define $z := \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i (y_i - g'((\bar{x}_i, \theta_1)))\bar{x}_i$. Let $J$ be any group of coordinates in $\{1, 2, \cdots, d\}$ with $m$ largest coordinates of $\{|z_j|\}_{j=1}^{N}$ for $m \in \{1, 2, \cdots, d\}$. Then, it follows

$$\sup_{\nu \in B_1(0,\rho) \cap B_2(0, r)} (z, \nu) \leq \sup_{\nu \in B_1(0,\rho) \cap B_2(0, r)} \sum_{j \in J} v_j z_j + \sup_{\nu \in B_1(0,\rho) \cap B_2(0, r)} \sum_{j \in J^c} v_j z_j \leq \max_{j} |z_j| \cdot (r\sqrt{m} + \rho) \quad (2.27)$$

for any $m$, where $\left\{\xi_j\right\}_{j=1}^{d}$ denotes the non-increasing ordering of $\{|z_j|\}_{j=1}^{d}$. Now for each $|z_j|$, let $\xi_i = y_i - g'((\bar{x}_i, \theta_1))$,

$$N|z_j| = \left| \sum_{i=1}^{N} \varepsilon_i (y_i - g'((\bar{x}_i, \theta_1)))\bar{x}_{ij} \right| \leq \sum_{i=1}^{N} \varepsilon_i \xi_i \bar{x}_{ij} + \left| \sum_{i=1}^{N} \varepsilon_i g'((\bar{x}_i, \theta_1)) - \sum_{i=1}^{N} \varepsilon_i (g'((\bar{x}_i, \theta_1)) - g'((\bar{x}_i, \theta_1)))\bar{x}_{ij} \right|$$

69
Thus, it follows

\[
N \cdot \max_{j \in \{1, 2, \ldots, d\}} |z_j| \leq \max_{j \in \{1, 2, \ldots, d\}} \left| \sum_{i=1}^{N} \varepsilon_i \xi_{ij} \right| + \max_{j \in \{1, 2, \ldots, d\}} \left| \sum_{i=1}^{N} \varepsilon_i (g'(\langle \xi_i, \theta_i \rangle) - g'(\langle x_i, \theta_i \rangle)) \right|
\]

(2.28)

By the same analysis as that of Lemma 2.4.10 we obtain

\[
N \cdot \max_{j \in \{1, 2, \ldots, d\}} |z_j| \leq C(D_{\text{max}} + 1) \left( \nu_q^5 + \nu_q^3 + \nu_q \right) w \left( \log(ed) \beta^{3/4} + N^{1/4} \log(ed) \beta^{1/2} + \log^2 N \right)
\]

\[
+ C \nu_q \left( \nu_q + 1 \right) \left( \log(ed) \beta \log(ed) + \beta N^{1/4} \log(ed) \beta^{3/4} \right),
\]

with probability at least

\[
1 - 2e^{-\beta} - 2w^{-\beta} - c' \left( u^{-q} (ed)^{-q/4} + u^{-q/4} (ed)^{-q/4} \right) + \log(eN)^{\beta/2} w^{-\beta/2}.
\]

This implies the claim when combining with (2.27). \qed

**Lemma 2.5.3.** Suppose \( N \geq cs \log(ed) \) for some absolute constant \( c > 1 \), Assumption 2.2.3 and 2.2.4 hold and \( \Lambda_M = \delta^2 Q^2 D_{\text{min}} / 128 \). Then, we have

\[
r_M \leq C(\nu_q, \nu_q') \frac{D_{\text{max}} + 1}{D_{\text{min}}} \left( \frac{wu^2 v + \beta^3 N}{\delta^2 Q^2} + \sqrt{\frac{\log(ed)}{N}} \left( \frac{wu^2 v + \beta^3 N}{\delta^2 Q^2} \right)^{1/4} \right),
\]

when taking

\[
p_M = 2e^{-\beta} - 2w^{-\beta} - c' \left( u^{-q} (ed)^{-q/4} + u^{-q/4} (ed)^{-q/4} \right) + \log(eN)^{\beta/2} w^{-\beta/2}.
\]

for some absolute constant \( c' > 1 \) and any \( \beta, u, v, w > 7 \), where \( C(\nu_q, \nu_q') \) depends polynomially on \( \nu_q \) and \( \nu_q' \).

**Proof of Lemma 2.5.3** Since \( \Lambda_M = \delta^2 Q^2 D_{\text{min}} / 128 \), let \( m = s \) in Lemma 2.5.2 and the infimum of the \( r > 0 \) such that the right hand side of Lemma 2.5.2 is less than \( \delta^2 Q^2 D_{\text{min}} r^2 / 128 \) can be
achieved by setting the right hand side equal to \( \delta^2 Q^2 D_{\text{min}} r^2 / 128 \), which gives,

\[
\frac{\delta^2 Q^2}{128} r^2 = C(\nu_q, \nu_{q'}) (D_{\text{max}} + 1) \left( wu^2 v + w^2 \beta^{3/4} + \beta \right) \left( r \sqrt{m} + \rho \right) \sqrt{\frac{\log(ed)}{N}}.
\]

Solving the above quadratic equation gives

\[
\begin{aligned}
   r &= C(\nu_q, \nu_{q'}) \frac{D_{\text{max}} + 1}{D_{\text{min}}} \left( \frac{wu^2 v + w^2 \beta^{3/4} + \beta}{\delta^2 Q^2} \frac{s \log(ed)}{N} + \sqrt{\rho wu^2 v + w^2 \beta^{3/4} + \beta s \log(ed)} \frac{1}{N} \right) .
\end{aligned}
\]

Thus, the defined \( r_M \) must be bounded above by this value and the lemma is proved. \( \square \)

**Lemma 2.5.4.** Suppose \( N \geq s \log(ed) \) and \( \Lambda_V = D_{\text{min}} \delta^2 Q^2 / 28 \), then,

\[
r^2 \leq \frac{128 D_{\text{max}} \nu_q^6}{D_{\text{min}}^2 Q^2 \rho} \sqrt{\frac{\log(ed)}{N}}.
\]

The proof is the same as that of Lemma 2.4.19.

### 2.5.3 Putting everything together

**Lemma 2.5.5.** Suppose \( ||\theta_* - \theta_0||_1 \leq \rho/4 \), where \( \theta_0 \) an \( s \)-sparse vector and \( \rho \geq 8r(\rho) \sqrt{s} \), then,

\[
\Delta(\eta \theta_* \rho) \geq \frac{3}{4} \rho.
\]

**Proof of Lemma 2.5.5.** Let \( G_s \) be the set of nonzero coordinates of \( \theta_0 \), then, for any vector \( v \in B_2(0, r) \cap S_V(0, \rho) \), we have \( v = P_{G_s} v + P_{\bar{G}_s} v \) and since \( ||\theta_* - \theta_0||_1 \leq \rho/4 \), by definition of \( \Gamma(\theta_*, \rho) \) in (2.8), there exists a sub-differential \( z^* \in \Gamma(\theta_*, \rho) \) such that \( \langle z^*, \theta_0 \rangle = ||\theta_0||_1 \) and \( \langle z^*, \rho_{G_s} v \rangle = ||P_{G_s} v||_1 \). Thus, it follows,

\[
\langle z^*, v \rangle = \langle z^*, P_{G_s} v \rangle + \langle z^*, P_{\bar{G}_s} v \rangle \geq ||P_{G_s} v||_1 - ||P_{\bar{G}_s} v||_1
\]

\[
\geq ||v||_1 - 2||P_{G_s} v||_1 \geq \rho - 2s ||P_{G_s} v||_1 \geq \rho - 2s \rho \sqrt{s},
\]

where the second from the last inequality follows from \( v \in B_2(0, r) \cap S_V(0, \rho) \) that \( ||v||_1 = \rho \) and the last inequality follows from \( ||P_{G_s} v||_2 \leq ||v||_2 \leq r(\rho) \). The above bound is greater than \( 3\rho/4 \) when \( \rho \geq 8r(\rho) \sqrt{s} \).

Finally, we are ready to prove the main theorem .
Proof of Theorem 2.2.5. We set $\Lambda_Q = D_{\min}\delta^2Q^2/32$, $\lambda_M = D_{\min}\delta^2Q^2/128$ and $\lambda_V = D_{\min}\delta^2Q^2/128$.

By Theorem 2.5.1, Lemma 2.5.2 and 2.5.4, we have

\[
\begin{align*}
    r(\rho) &\leq \frac{8}{\sqrt{s}}\rho + \sqrt{\frac{128D_{\max}\nu_q^6}{D_{\min}\delta^2Q^2}} \rho^{1/2} \left( \frac{\log(ed)}{N} \right)^{1/4} + \\
    C(\nu_q,\nu_q') \frac{D_{\max} + 1}{D_{\min}} \left( \frac{wu^2v + w\beta^{3/4}}{\delta^2Q^2} \right) \sqrt{\frac{s\log(ed)}{N}} \left( \frac{\rho wu^2v + w\beta^{3/4} + \beta}{\delta^2Q^2} \right) \left( \frac{s \log(ed)}{N} \right)^{1/4}
\end{align*}
\]

By Lemma 2.5.5, the condition $\Delta(\eta\theta,\rho) \geq 3\rho/4$ is satisfied for any $\rho \geq 8r(\rho)\sqrt{s}$. Take equality in the above bound and choose $\rho$ to be

\[
\rho = C(\nu_q,\nu_q') \frac{D_{\max} + 1}{D_{\min}} \frac{wu^2v + w\beta^{3/4}}{\delta^2Q^2} \frac{\log(ed)}{N},
\]

where $C(\nu_q,\nu_q')$ depends polynomially on $\nu_q,\nu_q'$ and $\rho \geq 8r(\rho)\sqrt{s}$ is satisfied. This implies

\[
r(\rho) \leq C'(\nu_q,\nu_q') \frac{D_{\max} + 1}{D_{\min}} \frac{wu^2v + w\beta^{3/4} + \beta}{\delta^2Q^2} \sqrt{\frac{s \log(ed)}{N}}.
\]

Taking

\[
\lambda = C''(\nu_q,\nu_q') \frac{D_{\max} + 1}{D_{\min}} \frac{wu^2v + w\beta^{3/4} + \beta}{\delta^2Q^2} \sqrt{\frac{\log(ed)}{N}}
\]

finishes the proof. \qed
Chapter 3

Structured Recovery from Non-linear and Heavy-tailed Measurements

In this chapter, we show that when the design vectors are selected from a specific class of distributions, then, one can simultaneously relax the moment condition as well as treat more general structured problems. We study high-dimensional signal recovery from non-linear measurements with design vectors having elliptically symmetric distribution. Special attention is devoted to the situation when the unknown signal belongs to a set of low statistical complexity, while both the measurements and the design vectors are heavy-tailed. We propose and analyze a new estimator that adapts to the structure of the problem, while being robust both to the possible model misspecification characterized by arbitrary non-linearity of the measurements as well as to data corruption modeled by the heavy-tailed distributions. Moreover, this estimator has low computational complexity. Theoretically, our results are expressed in the form of exponential concentration inequalities relying on an improved generic chaining method. Numerically, we conduct simulation experiments demonstrating that our estimator outperforms existing alternatives when data is heavy-tailed.

3.1 Introduction

In many practical settings, exact measurements from linear models or GLMs (2.1) are not available. Instead, the data one observes are often subject to unknown distortions such as quantization and hard thresholding. Furthermore, one might not even know the exact model (2.1). Is it possible to perform faithful parameter estimation in these imperfect scenarios? This chapter
treats this problem with a more general setup than that of (2.1). Instead of adopting a specific model, we assume the link function is unknown. More specifically, let \((x, y) \in \mathbb{R}^d \times \mathbb{R}\) be a random couple satisfying the semi-parametric single index model

\[
y = f((x, \theta_*), \delta),
\]

where \(x\) is a measurement vector with marginal distribution \(\Pi\), \(\delta\) is a noise variable that is assumed to be independent of \(x\), \(\theta_* \in \mathbb{R}^d\) is a fixed but otherwise unknown signal (“index vector”), and \(f : \mathbb{R}^2 \mapsto \mathbb{R}\) is an unknown link function; here and in what follows, \(\langle \cdot, \cdot \rangle\) denotes the Euclidean dot product. We impose no explicit conditions on \(f\), and in particular it is not assumed that \(f\) is convex, or even continuous. Our goal is to estimate the signal \(\theta_*\) from a sequence of samples \((x_1, y_1), \ldots, (x_N, y_N)\) which are copies of \((x, y)\). As \(f(a^{-1}\langle x, a\theta_* \rangle, \delta) = f((x, \theta_*), \delta)\) for any \(a > 0\), the best one can hope for is to recover \(\theta_*\) up to a scaling factor. Hence, without loss of generality, we will assume that \(\theta_*\) satisfies \(\|\Sigma^{1/2}\theta_*\|_2^2 := \langle \Sigma^{1/2}\theta_* , \Sigma^{1/2}\theta_* \rangle = 1\), where \(\Sigma = \mathbb{E}(x - \mathbb{E}x)(x - \mathbb{E}x)^T\) is the covariance matrix of \(x\). Instead of being sparse or approximately sparse, in this chapter, we will assume that \(\theta_*\) is an element of a closed set \(\Theta \subseteq \mathbb{R}^d\) of small statistical complexity that is characterized by its Gaussian mean width.

Due to the ambiguity of \(f\), such a task can easily fail regardless of the algorithms [ALPV14]. As an example, consider the model \(y_i = \text{sign}(\langle x_i, \theta_* \rangle)\). Consider two sparse vectors: \(\theta_1 = [1, 0, 0, \cdots, 0], \theta_2 = [1, -0.5, 0, \cdots, 0]\), and i.i.d. Bernoulli design vectors \(x_i\), where each entry takes +1 and -1 with equal probabilities. It is obvious that for \(\theta_* = \theta_1\) and \(\theta_* = \theta_2\), the responses \(y_i\) are identical and the model cannot distinguish between \(\theta_1\) and \(\theta_2\). Thus, one has to pose extra assumptions on the design vector itself so that the problem is well-defined.

Generally, the task of estimating the index vector requires approximating the link function \(f\) or its derivative, assuming that it exists (the so-called Average Derivative Method), see [Sto86, HJS01]. However, when the measurement vector \(x\) is Gaussian, a somewhat surprising result states that one can estimate \(\theta_*\) directly, avoiding preliminary link function estimation step completely. More specifically, [Bri83] proved that \(\hat{\theta}_* = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}(y - \langle \theta, x \rangle)^2\), where \(\eta = \mathbb{E}(y x, \theta_*)\). Later, [LD89] extended this result to the more general case of elliptically symmetric distributions, which includes the Gaussian as a special case; see Lemma 3.5.5.

Our work was partly inspired by the work of Y. Plan, R. Vershynin and E. Yudovina [PVY14].
[PV16], who presented the non-asymptotic study for the case of Gaussian measurements in the context of high-dimensional structured estimation; also, see [Gen16, ALPV14, TAH15, YWCL15] for further details. On a high level, these works show that when $x_j$’s are Gaussian, nonlinearity can be treated as an additional noise term. To give an example, [PV16] and [PVY14] demonstrate that under the same model as (3.1), when $x_j \sim N(0, I_{d \times d})$, $\theta^* \in \Theta$, and $y_j$ is sub-Gaussian for $j = 1, \ldots, N$, solving the constrained problem

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \| y - X\theta \|_2^2,$$

with $y = [y_1 \cdots y_N]^T$ and $X = \frac{1}{\sqrt{N}}[x_1 \cdots x_N]^T$, recovers $\theta^*$ up to a scaling factor $\eta$ with high probability: namely, for all $\beta \geq 2$,

$$P \left[ \left\| \hat{\theta} - \eta \theta^* \right\|_2 \geq C \frac{\omega(D(\Theta, \eta \theta^*) \cap S_2(1)) + \beta}{\sqrt{N}} \right] \leq ce^{-\beta^2/2},$$

(3.2)

where, with formal definitions to follow in Section 3.2, $S_2(1)$ is the unit sphere in $\mathbb{R}^d$, $D(\Theta, \theta)$ is the descent cone of $\Theta$ at point $\theta$ and $\omega(T)$ is the Gaussian mean width of a subset $T \subset \mathbb{R}^d$. A different approach to estimation of the index vector in model (3.1) with similar recovery guarantees has been developed in [YWCL15]. However, the key assumption adopted in all these works that the vectors $x_j$ follow Gaussian distributions preclude situations where the measurements are heavy tailed, and hence might be overly restrictive for some practical applications; for example, noise and outliers observed in high-dimensional image recovery often exhibit heavy-tailed behavior, see [WYG+09]. The works [YBL17] and [YBWL17] later consider using Stein’s identity to perform nonlinear recovery under the assumption that the distribution of the sensing vector is known, both the distribution function and the nonlinear transform must satisfy certain smoothness assumptions,

As we mentioned above, [LD89] have shown that direct consistent estimation of $\theta^*$ is possible when $\Pi$ belongs to a family of elliptically symmetric distributions. Our main contribution is the non-asymptotic analysis for this scenario, with a particular focus on the case when $d > n$ and $\theta^*$ possesses special structure, such as sparsity. Moreover, we make very mild assumptions on the tails of the response variable $y$: for example, when the link function satisfies $f(\langle x, \theta^* \rangle, \delta) = \tilde{f}(\langle x, \theta^* \rangle) + \delta$, it is only assumed that $\delta$ possesses $2 + \varepsilon$ moments, for some $\varepsilon > 0$. [PV16] present analysis for the Gaussian case and ask “Can the same kind of accuracy be expected for random
non-Gaussian matrices?” In this chapter, we give a positive answer to their question. To achieve our goal, we propose a Lasso-type estimator that admits tight probabilistic guarantees in spirit of despite weak tail assumptions (see Theorem below for details).

3.2 Definitions and Background Material.

This section introduces main notation and the key facts related to elliptically symmetric distributions, convex geometry and empirical processes. The results of this section will be used repeatedly throughout the chapter. For the unified treatment of vectors and matrices, it will be convenient to treat a vector \( v \in \mathbb{R}^{d \times 1} \) as a \( d \times 1 \) matrix. Let \( d_1, d_2 \in \mathbb{N} \) be such that \( d_1 d_2 = d \). Given \( v_1, v_2 \in \mathbb{R}^{d_1 \times d_2} \), the Euclidean dot product is then defined as \( \langle v_1, v_2 \rangle = \text{tr}(v_1^T v_2) \), where \( \text{tr}(\cdot) \) stands for the trace of a matrix and \( v^T \) denotes the transpose of \( v \).

The \( \ell_1 \)-norm of \( v \in \mathbb{R}^d \) is defined as \( \|v\|_1 = \sum_{j=1}^d |v_j| \). The nuclear norm of a matrix \( v \in \mathbb{R}^{d_1 \times d_2} \) is \( \|v\|_* = \sum_{j=1}^{\min(d_1,d_2)} \sigma_j(v) \), where \( \sigma_j(v), j = 1, \ldots, \min(d_1,d_2) \) stand for the singular values of \( v \), and the operator norm is defined as \( \|v\| = \max_{j=1,\ldots,\min(d_1,d_2)} \sigma_j(v) \).

3.2.1 Elliptically symmetric distributions.

A centered random vector \( x \in \mathbb{R}^d \) has elliptically symmetric (alternatively, elliptically contoured or just elliptical) distribution with parameters \( \Sigma \) and \( F_{\mu} \), denoted \( x \sim \mathcal{E}(0, \Sigma, F_{\mu}) \), if

\[
x \overset{d}{=} \mu \mathbf{B} \mathbf{U},
\]

where \( \overset{d}{=} \) denotes equality in distribution, \( \mu \) is a scalar random variable with cumulative distribution function \( F_{\mu} \), \( \mathbf{B} \) is a fixed \( d \times d \) matrix such that \( \Sigma = \mathbf{B} \mathbf{B}^T \), and \( \mathbf{U} \) is uniformly distributed over the unit sphere \( S_2(1) \) and independent of \( \mu \). Note that distribution \( \mathcal{E}(0, \Sigma, F_{\mu}) \) is well defined, as if \( \mathbf{B}_1 \mathbf{B}_1^T = \mathbf{B}_2 \mathbf{B}_2^T \), then there exists a unitary matrix \( \mathbf{Q} \) such that \( \mathbf{B}_1 = \mathbf{B}_2 \mathbf{Q} \), and \( \mathbf{Q} \mathbf{U} \overset{d}{=} \mathbf{U} \). Along these same lines, we note that representation (3.3) is not unique, as one may replace the pair \( (\mu, \mathbf{B}) \) with \( (c \mu, \frac{1}{c} \mathbf{B} \mathbf{Q}) \) for any constant \( c > 0 \) and any orthogonal matrix \( \mathbf{Q} \). To avoid such ambiguity, in the following we allow \( \mathbf{B} \) to be any matrix satisfying \( \mathbf{B} \mathbf{B}^T = \Sigma \), and noting that the covariance matrix of \( \mathbf{U} \) is a multiple of the identity, we further impose the condition that the covariance matrix of \( x \) is equal to \( \Sigma \), i.e. \( \mathbb{E}[xx^T] = \Sigma \).
Alternatively, the mean-zero elliptically symmetric distribution can be defined uniquely via its characteristic function
\[ s \rightarrow \psi(s^T \Sigma s), \quad s \in \mathbb{R}^d, \]
where \( \psi : \mathbb{R}^+ \rightarrow \mathbb{R} \) is called the characteristic generator of \( x \). For further details information about elliptically distribution, see [CHS81] for details.

An important special case of the family \( E(0, \Sigma, F\mu) \) of elliptical distributions is the Gaussian distribution \( N(0, \Sigma) \), where \( \mu = \sqrt{z} \) with \( z = \chi^2_d \), and the characteristic generator is \( \psi(x) = e^{-x^2/2} \).

The following elliptical symmetry property, generalizing the well known fact for the conditional distribution of the multivariate Gaussian, plays an important role in our subsequent analysis, see [CHS81]:

**Proposition 3.2.1.** Let \( x = [x_1, x_2] \sim E_d(0, \Sigma, F\mu) \), where are of dimension \( d_1 \) and \( d_2 \) respectively, with \( d_1 + d_2 = d \). Let \( \Sigma \) be partitioning accordingly as
\[
\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.
\]
Then, whenever \( \Sigma_{22} \) has full rank, the conditional distribution of \( x_1 \) given \( x_2 \) is elliptical \( E_{d_1}(0, \Sigma_{1|2}, F\mu_{1|2}) \), where
\[
\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21},
\]
and \( F\mu_{1|2} \) is the cumulative distribution function of \((\mu^2 - x_2^T \Sigma_{22}^{-1} x_2)^{1/2} \) given \( x_2 \).

Note that \( \mu^2 - x_2^T \Sigma_{22}^{-1} x_2 \) is always nonnegative, hence \( F\mu_{1|2} \) is well defined, since by (3.3) we have
\[
x_2^T \Sigma_{22}^{-1} x_2 = \mu^2 (B_2 U)^T (B_2 B_2^T)^{-1} (B_2 U) = \mu^2 U^T B_2^T (B_2 B_2^T)^{-1} B_2 U \leq \mu^2 U^T U = \mu^2,
\]
where \( B_2 \) is the matrix consisting of the last \( d_2 \) rows of \( B \) in (3.3), and where the inequality holds due to the fact that \( B_2^T (B_2 B_2^T)^{-1} B_2 \) is a projection matrix. The following corollary is easily deduced from the theorem above:

**Corollary 3.2.1.** If \( x \sim E_d(0, \Sigma, F\mu) \) with \( \Sigma \) of full rank, then for any two fixed vectors \( y_1, y_2 \in \mathbb{R}^d \),
\[ \mathbb{R}^d \text{ with } \|y_2\|_2 = 1, \]
\[ \mathbb{E}[\langle x, y_1 \rangle \mid \langle x, y_2 \rangle] = \langle y_1, y_2 \rangle \langle x, y_2 \rangle. \]

**Proof.** Let \( \{v_1, \ldots, v_d\} \) be an orthonormal basis in \( \mathbb{R}^d \) such that \( v_d = y_2 \). Let \( V = [v_1 \ v_2 \ \cdots \ v_d] \) and consider the linear transformation

\[ \bar{x} = V^T x. \]

Then, by (3.3), \( \bar{x} = \mu V^T BU \), which is centered elliptical with full rank covariance matrix \( V^T \Sigma V \).

Applications of Theorem 3.2.1 with \( x_1 = [(x, v_1), \ldots, \langle x, v_{d-1} \rangle] \) and \( x_2 = \langle x, v_d \rangle = \langle x, y_2 \rangle \) yields

\[ \mathbb{E}[\langle x, y_1 \rangle \mid \langle x, y_2 \rangle] = \mathbb{E} \left[ \sum_{i=1}^{d} \langle x, v_i \rangle \langle y_1, v_i \rangle \mid \langle x, v_d \rangle \right] = \mathbb{E} \left[ \sum_{i=1}^{d-1} \langle x, v_i \rangle \langle y_1, v_i \rangle \mid \langle x, v_d \rangle \right] + \langle x, v_d \rangle \langle y_1, v_d \rangle = \langle x, v_d \rangle \langle y_1, v_d \rangle = \langle y_1, y_2 \rangle \langle x, y_2 \rangle, \]

where in the second to last equality we have used the fact that the conditional distribution of \( [(v_1, x), \ldots, \langle v_{d-1}, x \rangle] \) given \( \langle x, v_d \rangle \) is elliptical with mean zero. \( \square \)

### 3.2.2 Geometry.

**Definition 3.2.1 (Restricted set).** Given \( c_0 > 1 \), the \( c_0 \)-restricted set of the norm \( \| \cdot \|_K \) at \( \theta \in \mathbb{R}^d \) is defined as

\[ S_{c_0}(\theta) := S_{c_0}(\theta; K) = \left\{ v \in \mathbb{R}^d : \| \theta + v \|_K \leq \| \theta \|_K + \frac{1}{c_0} \| v \|_K \right\}. \]  \hspace{1cm} (3.4)

**Definition 3.2.2 (Restricted compatibility).** The restricted compatibility constant of a set \( A \subseteq \mathbb{R}^d \) with respect to the norm \( \| \cdot \|_K \) is given by

\[ \Psi(A) := \Psi(A; K) = \sup_{v \in A \setminus \{0\}} \frac{\| v \|_K}{\| v \|_2}. \]

**Remark 3.2.1.** The restricted set from the definition 3.2.1 is not necessarily convex. However,
if the norm $\| \cdot \|_K$ is decomposable (see definition 3.7.1), then the restricted set is contained in a convex cone, and the corresponding restricted compatibility constant is easier to estimate. Decomposable norms have been introduced by [NRWY12] and later appeared in a number of works, e.g. [BCFST14] and references therein. For reader’s convenience, we provide a self-contained discussion in Appendix 3.7.

3.3 Main Results

In this section, we define a version of Lasso estimator that is well-suited for heavy-tailed measurements, and state its performance guarantees.

We will assume that $x_1, x_2, \ldots, x_N \in \mathbb{R}^d$ are i.i.d. copies of an isotropic vector $x$ with spherically symmetric distribution $\mathcal{E}_d(0, \mathbf{I}_d \times \mathbf{I}_d, F_\mu)$. If $x \sim \mathcal{E}_d(0, \Sigma, F_\mu)$ for some positive definite matrix $\Sigma$, then by definition $x = \mu \Sigma^{1/2} U$, and $\langle x, \theta \rangle = \langle \Sigma^{-1/2} x, \Sigma^{1/2} \theta \rangle$, where $\Sigma^{-1/2} x = \mu U \sim \mathcal{E}_d(0, \mathbf{I}_d \times \mathbf{I}_d, F_\mu)$. Hence, if we set $\tilde{\theta} := \Sigma^{1/2} \theta$, then all results that we establish for isotropic measurements hold with $\theta$ replaced by $\tilde{\theta}$; remark after Theorem 3.3.1 includes more details.

3.3.1 Description of the proposed estimator.

We first introduce an estimator under the scenario that $\theta_* \in \Theta$, for some known closed set $\Theta \subseteq \mathbb{R}^d$. Define the loss function $L_N^0(\cdot)$ as

$$L_N^0(\theta) := \|\theta\|_2^2 - \frac{2}{N} \sum_{i=1}^N \langle y_i x_i, \theta \rangle,$$

which is the unbiased estimator of

$$L^0(\theta) := \|\theta\|_2^2 - 2\mathbb{E} \langle y x, \theta \rangle = \mathbb{E} (y - \langle x, \theta \rangle)^2 - \mathbb{E} y^2,$$

where the last equality follows since $x$ is isotropic. Clearly, minimizing $L^0(\theta)$ over any set $\Theta \subseteq \mathbb{R}^d$ is equivalent to minimizing the quadratic loss $\mathbb{E} (y - \langle x, \theta \rangle)^2$. If distribution $F_\mu$ has heavy tails, the sample average $\frac{1}{N} \sum_{i=1}^N y_i x_i$ might not concentrate sufficiently well around its mean, hence we replace it by a more “robust” version obtained via truncation. Let $\mu \in \mathbb{R}, U \in S_2(1)$ be such
that $x = \mu U$ (so that $\mu = \|x\|_2$), and set

\[ \tilde{U} = \sqrt{d} U, \quad q = \mu y / \sqrt{d}, \]

so that $q \tilde{U} = yx$ and $\tilde{U}$ is uniformly distributed on the sphere of radius $\sqrt{d}$, implying that its covariance matrix is $I_d$, the identity matrix. Next, define the truncated random variables

\[ \tilde{q}_i = \text{sign}(q_i)(|q_i| \wedge \tau), \quad i = 1, \ldots, m, \quad (3.7) \]

where $\tau = N^{-\frac{1}{4\kappa + 1}}$ for some $\kappa \in (0, 1)$ that is chosen based on the integrability properties of $q$, see (3.16). Finally, set

\[ L^*_N(\theta) = \|\theta\|_2^2 - \frac{2}{N} \sum_{i=1}^{N} \left\langle \tilde{q}_i \tilde{U}_i, \theta \right\rangle, \]

and define the estimator $\hat{\theta}_N$ as the solution to the constrained optimization problem:

\[ \hat{\theta}_N := \arg\min_{\theta \in \Theta} L^*_N(\theta). \quad (3.9) \]

We will also denote

\[ L^*(\theta) := \mathbb{E} L^*_N(\theta) = \|\theta\|_2^2 - 2\mathbb{E} \left\langle \tilde{q} \tilde{U}, \theta \right\rangle. \quad (3.10) \]

For the scenarios where structure on the unknown $\theta_*$ is induced by a norm $\| \cdot \|_\mathcal{K}$ (e.g., if $\theta_*$ is sparse, then $\| \cdot \|_\mathcal{K}$ could be the $\| \cdot \|_1$ norm), we will also consider the estimator $\hat{\theta}_m^\lambda$ defined via

\[ \hat{\theta}_N^\lambda := \arg\min_{\theta \in \mathbb{R}^d} \left[ L^*_N(\theta) + \lambda \|\theta\|_\mathcal{K} \right], \quad (3.11) \]

where $\lambda > 0$ is a regularization parameter to be specified, and $L^*_N(\theta)$ is defined in (3.8).

Let us note that truncation approach has previously been successfully implemented by [FWZ16b] to handle heavy-tailed noise in the context of matrix recovery with sub-Gaussian design. In the present chapter, we show that truncation-based approach is also useful in the situations where the measurements are heavy-tailed.
Remark 3.3.1. Note that our estimator (3.11) is in general much easier to implement than some other popular alternatives, such as the usual Lasso estimator [Tib96]. For example, when the signal $\theta$ is sparse, our estimator takes the form

$$\hat{\theta}_N := \arg\min_{\theta \in \mathbb{R}^d} \left[ \|\theta\|_2^2 - \frac{2}{N} \sum_{i=1}^{N} \langle \bar{q}_i \bar{U}_i, \theta \rangle + \lambda \|\theta\|_1 \right],$$

which yields a closed form solution in the form of “soft-thresholding”. Specifically, let $b = \frac{1}{N} \sum_{i=1}^{N} \bar{q}_i \bar{U}_i$, then, the $k$-th entry of $\hat{\theta}_N$ takes the form:

$$\left(\hat{\theta}_N\right)_k = \begin{cases} 
    b_k - \lambda/2, & \text{if } b_k \geq \lambda/2, \\
    0, & \text{if } -\lambda/2 \leq b_k \leq \lambda/2, \\
    b_k + \lambda/2, & \text{if } b_k \leq -\lambda/2.
\end{cases} \tag{3.12}$$

We should note however that such simplification comes at the cost of knowing the distribution of measurement vector $x$. Despite being of low computational complexity, our estimator can still exploit the structure of the problem, while being robust both to the possible model misspecification as well as to data corruption modeled by the heavy-tailed distributions. We demonstrate this in the following sections.

Remark 3.3.2 (Non-isotropic measurements). When $x \sim \mathcal{E}_d(0, \Sigma, F_{\mu})$ for some $\Sigma \succ 0$, then estimator (3.9) has to be replaced by

$$\tilde{\theta}_N := \arg\min_{\theta \in \mathbb{R}^d} \left[ \|\theta\|_2^2 - \frac{2}{N} \sum_{i=1}^{N} \langle \bar{q}_i \bar{U}_i, \Sigma \theta \rangle \right], \tag{3.13}$$

which is equivalent to

$$\tilde{\theta}_N := \arg\min_{\theta \in \mathbb{R}^d} \left[ \|\theta\|_2^2 - \frac{2}{N} \sum_{i=1}^{N} \langle \bar{q}_i \bar{U}_i, \theta \rangle \right],$$

is a sense that $\tilde{\theta}_m = \Sigma^{1/2} \tilde{\theta}_m$. Hence, results obtained for isotropic measurements easily extend to the more general case. Similarly, estimator (3.11) should be replaced by

$$\hat{\theta}_N^\lambda := \arg\min_{\theta \in \mathbb{R}^d} \left[ \|\Sigma^{1/2} \theta\|_2^2 - \frac{2}{N} \sum_{i=1}^{N} \langle \bar{q}_i \bar{U}_i, \Sigma^{1/2} \theta \rangle + \lambda \|\Sigma^{1/2} \theta\|_K \right], \tag{3.14}$$
which is equivalent to

$$\tilde{\theta}_N^\lambda := \arg\min_{\theta \in \mathbb{R}^d} \left[ \|\theta\|^2_2 - \frac{2}{N} \sum_{i=1}^{N} \langle \tilde{q}_i \tilde{U}_i, \theta \rangle + \lambda \|\theta\|_{\Sigma^{1/2} \Sigma^{1/2}} \right],$$

meaning that $\tilde{\theta}_m^\lambda = \Sigma^{1/2} \tilde{\theta}_N^\lambda$.

### 3.3.2 Estimator performance guarantees.

In this section, we present the probabilistic guarantees for the performance of the estimators $\tilde{\theta}_N$ and $\tilde{\theta}_m^\lambda$ defined by (3.9) and (3.11) respectively.

Everywhere below, $C, c, C_j$ denote numerical constants; when these constants depend on parameters of the problem, we specify this dependency by writing $C_j = C_j(\text{parameters})$. Let

$$\eta = \mathbb{E} \langle yx, \theta_* \rangle,$$

and assume that $\eta \neq 0$ and $\eta \theta_* \in \Theta$.

**Theorem 3.3.1.** Suppose that $x \sim \mathcal{E}(0, \mathbf{I}_{d \times d}, F_\mu)$. Moreover, suppose that for some $\kappa > 0$

$$\phi := \mathbb{E} \|q\|^{2(1+\kappa)} < \infty.$$  \hspace{1cm} (3.16)

Then there exist constants $C_1 = C_1(\kappa, \phi), C_2 = C_2(\kappa, \phi) > 0$ such that $\tilde{\theta}_N$ satisfies

$$\mathbb{P} \left( \|\tilde{\theta}_N - \eta \theta_*\|_2 \geq C_1 \frac{\omega(D(\Theta, \eta \theta_*), S_2(1)) + 1}{\sqrt{N}} \right) \leq C_2 e^{-\beta/2},$$

for any $\beta \geq 8$ and $N \geq \beta^2 (\omega(D(\Theta, \eta \theta_*), S_2(1)) + 1)^2$.

**Remark 3.3.3.**

1. Unknown link function $f$ enters the bound only through the constant $\eta$ defined in (3.15).

2. Aside from independence, conditions on the noise $\delta$ are implicit and follow from assumptions on $y$. In the special case when the error is additive, that is, when $y = f(\langle x, \theta_* \rangle) + \delta$, the moment condition (3.16) becomes $\mathbb{E} \|x\|_2 f(\langle x, \theta_* \rangle) + \|x\|_2 \delta^{2(1+\kappa)} < \infty$, for which it is sufficient to assume that $\mathbb{E} \|x\|_2 f(\langle x, \theta_* \rangle) < \infty$ and $\mathbb{E} \|x\|_2 \delta^{2(1+\kappa)} < \infty$.

3. Theorem 3.3.1 is mainly useful when $\eta \theta_*$ lies on the boundary of the set $\Theta$. Otherwise, if
ηθ∗ belongs to the relative interior of Θ, the descent cone $D(\Theta, \eta\theta_*)$ is the affine hull of Θ (which will often be the whole space $\mathbb{R}^d$). Thus, in such cases the Gaussian mean width $\omega(D(\Theta, \eta\theta_*) \cap S_2(1))$ can be on the order of $\sqrt{d}$, which is prohibitively large when $d \gg m$.

We refer the reader to [PV16, PVY14] for a discussion of related result and possible ways to tighten them.

Next, we present performance guarantees for the unconstrained estimator (3.11).

**Theorem 3.3.2.** Assume that the norm $\|\cdot\|_K$ dominates the 2-norm, i.e. $\|v\|_K \geq \|v\|_2$, $\forall v \in \mathbb{R}^d$.

Let $x \sim \mathcal{E}(0, I_{d \times d}, F_\mu)$, and suppose that for some $\kappa > 0$

$$\phi := \mathbb{E}[q](1 + \kappa) < \infty.$$  

Then there exist constants $C_3 = C_3(\kappa, \phi), C_4 = C_4(\kappa, \phi) > 0$ such that for all $\lambda \geq \frac{C_3 \beta}{\sqrt{N}} (1 + \omega(\mathcal{G}))$

$$\mathbb{P} \left( \left\| \frac{\hat{\theta}_N^- - \eta\theta_*}{\lambda} \right\|_2 \geq \frac{3}{2} \lambda \cdot \Psi(\mathcal{S}_2(\eta\theta_*)) \right) \leq C_4 e^{-\beta^2/2},$$  

for any $\beta \geq 8$ and $N \geq (\omega(\mathcal{G}) + 1)^{1/2} \beta^2$, where $\mathcal{G} := \{ x \in \mathbb{R}^d : \|x\|_K \leq 1 \}$ is the unit ball of $\|\cdot\|_K$ norm, and $\mathcal{S}_2(\cdot)$ and $\Psi(\cdot)$ are given in Definitions 3.2.1 and 3.2.2 respectively.

**Remark 3.3.4 (Non-isotropic measurements).** It follows from remark 3.3.3 and (3.13) that, whenever $x \sim \mathcal{E}_d(0, \Sigma, F_\mu)$, inequality of Theorem 3.3.1 has the form

$$\mathbb{P} \left( \left\| \Sigma^{1/2} \left( \frac{\hat{\theta}_N^- - \eta\theta_*}{\lambda} \right) \right\|_2 \geq C_1 \left( \frac{\omega(D(\Theta, \eta\theta_*) \cap S_2(1)) + 1}{\sqrt{N}} \right) \frac{\beta}{\lambda} \right) \leq C_2 e^{-\beta^2/2},$$

which can be further combined with the bound

$$\omega(D(\Theta, \eta\theta_*) \cap S_2(1)) \leq \|\Sigma^{1/2}\| \cdot \|\Sigma^{-1/2}\| \omega(D(\Theta, \eta\theta_*) \cap S_2(1)),$$

that follows from remark 1.7 in [PV16]. Similarly, the inequality of Theorem 3.3.2 holds with

$$\mathcal{G}_{\Sigma^{1/2}} := \{ x \in \mathbb{R}^d : \|x\|_K \leq 1 \};$$  

83
the unit ball of \( \| \cdot \|_{\Sigma^{1/2} K} \) norm, in place of \( G \). Namely, for all \( \lambda \geq \frac{C_3 \beta}{\sqrt{N}} (1 + \omega(G_{\Sigma^{1/2}})) \),

\[
P \left( \| \Sigma^{1/2} (\hat{\theta}_N - \eta \theta_*) \|_2 \geq \frac{3}{2} \lambda \cdot \Psi \left( S_2 \left( \eta \Sigma^{1/2} \theta_* \right) ; \Sigma^{1/2} K \right) \right) \leq C_4 e^{-\beta/2}
\]

Note that \( \omega(G_{\Sigma^{1/2}}) \leq \| \Sigma^{1/2} \|_\omega(G) \). Moreover, we show in Appendix 3.7 that for a class of decomposable norms (which includes \( \| \cdot \|_1 \) and nuclear norm), the upper bounds for \( \Psi \left( S_2 \left( \eta \Sigma^{1/2} \theta_* \right) ; \Sigma^{1/2} K \right) \) and \( \Psi \left( S_2(\eta \theta_*) \right) \) differ by the factor of \( \| \Sigma^{-1/2} \| \).

3.3.3 Examples.

We discuss two popular scenarios: estimation of the sparse vector and estimation of the low-rank matrix.

**Estimation of the sparse signal.** Assume that there exists \( J \subseteq \{1, \ldots, d\} \) of cardinality \( s \leq d \) such that \( \theta_{*,j} = 0 \) for \( j \notin J \). Let \( \Theta = \{ \theta \in \mathbb{R}^d : \| \theta \|_1 \leq \| \eta \theta_* \|_1 \} \), with \( \eta \) defined in (3.15). In this case, it is well-known that \( \omega^2 (D(\Theta, \eta \theta_*) \cap S_2(1)) \leq 2s \log(d/s) + \frac{5}{4}s \), see proposition 3.10 in [CRPW12], hence Theorem 3.3.1 implies that, with high probability,

\[
\| \hat{\theta}_N - \eta \theta_* \|_2 \lesssim \sqrt{\frac{s \log(d/s)}{N}} \tag{3.17}
\]
as long as \( m \gtrsim s \log(d/s) \).

We compare this bound to result of Theorem 3.3.2 for constrained estimator. Let \( \| \cdot \|_K \) be the \( \ell_1 \) norm. It is well-known that \( \omega(G) = \mathbb{E} \max_{j=1,\ldots,d} |g_j| \leq \sqrt{2 \log(2d)} \), where \( g \sim N(0, I_{d \times d}) \).

Moreover, we show in Appendix 3.7 that \( \Psi (S_2(\eta \theta_*)) \leq 4\sqrt{s} \). Hence, for \( \lambda \approx \sqrt{\frac{\log(2d)}{N}} \), Theorem 3.3.2 implies that

\[
\| \hat{\theta}_N - \eta \theta_* \|_2 \lesssim \sqrt{\frac{s \log(d)}{N}}
\]

with high probability whenever \( m \gtrsim \log(2d) \). This bound is only marginally weaker than (3.17) due to the logarithmic factor, however, definition of \( \hat{\theta}_N \) does not require the knowledge of \( \| \eta \theta_* \|_1 \), as we have already mentioned before.

**Estimation of a low-rank matrix.** Assume that \( d = d_1 d_2 \) with \( d_1 \leq d_2 \), and \( \theta_* \in \mathbb{R}^{d_1 \times d_2} \) has rank \( r \leq \min(d_1, d_2) \). Let \( \Theta = \{ \theta \in \mathbb{R}^{d_1 \times d_2} : \| \theta \|_* \leq \| \eta \theta_* \|_* \} \). Then the Gaussian mean width of the intersection of a descent cone with a unit ball is bounded as \( \omega^2 (D(\Theta, \eta \theta_*) \cap S_2(1)) \leq 3r(d_1 + d_2 - r) \), see proposition 3.11 in [CRPW12], hence Theorem 3.3.1 yields that, with high
probability,
\[ \| \hat{\theta}_N - \eta \theta^* \|_2 \lesssim \sqrt{\frac{r(d_1 + d_2)}{N}} \]
as long as the number of observations satisfies \( m \gtrsim r(d_1 + d_2) \).

Finally, we derive the corresponding bound from Theorem 3.3.2. The Gaussian mean width of
the unit ball in the nuclear norm is bounded by \( 2(\sqrt{d_1} + \sqrt{d_2}) \), see proposition 10.3 in [Ver15]. It
follows from results in Appendix 3.7 that \( \Psi (S_2 (\eta \theta^*)) \leq 4\sqrt{2r} \). Theorem 3.3.2 now implies that
with high probability
\[ \| \hat{\theta}_N - \eta \theta^* \|_2 \lesssim \sqrt{\frac{r(d_1 + d_2)}{N}}, \]
which matches the bound of Theorem 3.3.1.

### 3.4 Numerical Experiments

In this section, we demonstrate the performance of proposed robust estimator (3.11) for one-bit compressed sensing model. The model takes the following form:

\[ y = \text{sign}(\langle x, \theta^* \rangle) + \delta, \quad (3.18) \]

where \( \delta \) is the additive noise and the parameter \( \theta^* \) is assumed to be \( s \)-sparse. This model is highly non-linear because one can only observe the sign of each measurement.

The 1-bit compressed sensing model was previously discussed extensively in a number of works [PVY14, ALPV14, PV16]. It was shown that when the measurement vectors are either Gaussian or sub-Gaussian, the Lasso estimator recovers the support of \( \theta^* \) with high probability.

Here, we show that under the heavy-tailed elliptically distributed measurements, our estimator numerically outperforms the standard Lasso estimator

\[ \theta_{\text{Lasso}} = \arg \min_{\theta \in \mathbb{R}^d} \| X \theta - y \|_2^2 + \lambda \| \theta \|_1, \]

while taking the form of a simple soft-thresholding as explained in (3.12).

In the first numerical experiment, data are simulated in the following way: \( x_1, x_2, \ldots, x_{128} \in \mathbb{R}^{512} \) are i.i.d. with spherically symmetric distribution \( x_i \overset{d}{=} \mu_i U_i, \quad i = 1, \ldots, N \). The random vectors \( U_i \in \mathbb{R}^{512} \) are i.i.d. with uniform distribution over the sphere of radius \( \sqrt{512} \), and the
random variables $\mu_i \in \mathbb{R}$ are also i.i.d., independent of $U_i$ and such that

$$
\mu_i \overset{d}{=} \frac{1}{\sqrt{2c(q)}} (\xi_{i,1} - \xi_{i,2}),
$$

(3.19)

where $\xi_{i,1}$ and $\xi_{i,2}$, $i = 1, 2, \cdots, 128$ are i.i.d. with Pareto distribution, meaning that their probability density function is given by

$$
p(t; q) = \frac{q}{(1+t)^{1+q}} I_{\{t>0\}},
$$

c(q) \coloneqq \text{Var}(\xi) = \frac{q}{(q-1)^2(q-2)}, \text{ and } q = 2.1. \text{ The true signal } \theta^* \text{ has sparsity level } s = 5, \text{ with index of each non-zero coordinate chosen uniformly at random, and the magnitude having uniform distribution on } [0, 1].

Since we can only recover the original signal $\theta^*$ up to scaling, define the relative error for any estimator $\hat{\theta}$ with respect to $\theta^*$ as follows:

$$
\text{Relative error} = \left| \frac{\hat{\theta}}{\|\hat{\theta}\|_2} - \frac{\theta^*}{\|\theta^*\|_2} \right|.
$$

(3.20)

In each of the following two scenarios, we run the experiment 200 times for both the Lasso estimator and the estimator defined in (3.11) with $\|\cdot\|_K$ being the $\|\cdot\|_1$ norm. We set the truncation level as $\tau = cm^{1/(1+\kappa)}$, and the values of $c$ and regularization parameter $\lambda$ are obtained via the standard 2-fold cross validation for the relative error (3.20). We then plot the histogram of obtained results over 200 runs of the experiment.

In the first scenario, we set the additive error $\delta_i = 0$, $i = 1, 2, \cdots, 128$ in the 1-bit model (3.18) and plot the histogram in Fig. 3.1. We can see from the plot that the robust estimator (3.11) noticeably outperforms the Lasso estimator.

In the second scenario, we set the additive error $\delta_i$, $i = 1, 2, \cdots, 128$ to be i.i.d. heavy tailed noise with signal-to-noise ratio (SNR) equal to 10dB, so that the noise has the distribution

$$
\delta_i \overset{d}{=} \frac{h_i}{\sqrt{10}},
$$

1The signal-to-noise ratio (dB) is defined as $\text{SNR} \coloneqq 10 \log_{10} (\sigma_{\text{signal}}^2/\sigma_{\text{noise}}^2)$. In our case, since $\langle x_i, \theta^* \rangle$ can be positive or negative with equal probability, $\sigma_{\text{signal}}^2 = 1$, and thus, $\sigma_{\text{noise}}^2 = 1/10$. 

86
and \( h_i, \ i = 1, 2, \cdots, 128 \) are i.i.d. random variables with Pareto distribution, see (3.19). The results are plotted in Fig. 3.2. The histogram shows that, while performance of the Lasso estimator becomes worse, results of robust estimator (3.11) are relatively stable.

![Figure 3.1: Lasso vs robust estimator without additive noise.](image1)

![Figure 3.2: Lasso vs robust estimator under heavy-tailed noise with signal-to-noise ratio (SNR) equal to 10dB.](image2)

In the second simulation study, the simulation framework similar to the second scenario above, the only difference being the increased sample size \( N \). The results are plotted in Fig. 3.3-3.5 with sample sizes \( m = 128, 256 \) and 512, respectively.

### 3.5 Proofs.

This section is devoted to the proofs of Theorems 3.3.1 and 3.3.2.

#### 3.5.1 Preliminaries.

We recall several useful facts from probability theory that we rely on in the subsequent analysis.

The following well-known bound shows that the uniform distribution on a high-dimensional sphere enjoys strong concentration properties.

**Lemma 3.5.1** (Lemma 2.2 of [Bal97]). Let \( U \) have the uniform distribution on \( S_2(1) \). Then for any \( \Delta \in (0, 1) \) and any fixed \( v \in S_2(1) \),

\[
\mathbb{P}(\langle U, v \rangle \geq \Delta) \leq e^{-d\Delta^2/2}.
\]
Next, we state several useful results from the theory of empirical processes.

**Definition 3.5.1** ($\psi_q$-norm). For $q \geq 1$, the $\psi_q$-norm of a random variable $\xi \in \mathbb{R}$ is given by

$$\|\xi\|_{\psi_q} = \sup_{p \geq 1} p^{-\frac{1}{q}} (\mathbb{E}|X|^p)^{\frac{1}{p}}.$$  

Specifically, the cases $q = 1$ and $q = 2$ are known as the sub-exponential and sub-Gaussian norms respectively. We will say that $\xi$ is sub-exponential if $\|\xi\|_{\psi_1} < \infty$, and $X$ is sub-Gaussian if $\|\xi\|_{\psi_2} < \infty$.

**Remark 3.5.1.** It is easy to check that $\psi_q$-norm is indeed a norm.

**Remark 3.5.2.** A useful property, equivalent to the previous definition of a sub-Gaussian random variable $\xi$, is that there exists a positive constant $C$ such that

$$\mathbb{P}(|\xi| \geq u) \leq \exp(1 - Cu^2).$$
For the proof, see Lemma 5.5 in [Ver10a].

**Definition 3.5.2** (sub-Gaussian random vector). A random vector \( x \in \mathbb{R}^d \) is called sub-Gaussian if there exists \( C > 0 \) such that \( \|\langle x, v \rangle\|_{\psi_2} \leq C \) for any \( v \in S_2(1) \). The corresponding sub-Gaussian norm is then
\[
\|x\|_{\psi_2} := \sup_{v \in S_2(1)} \|\langle x, v \rangle\|_{\psi_2}.
\]

Next, we recall the notion of the generic chaining complexity. Let \( (T, d) \) be a metric space.

**Definition 3.5.3** (Admissible sequence). An increasing sequence of subsets \( \{A_l\}_{l=0}^\infty \) of \( T \) is admissible if \( |A_l| \leq N_l, \forall l \), where \( N_0 = 1 \) and \( N_l = 2^{2^l}, \forall l \geq 1 \).

For each \( A_l \), define the map \( \pi_l : T \to A_l \) as \( \pi_l(t) = \arg\min_{s \in A_l} d(s, t), \forall t \in T \). Note that, since each \( A_l \) is a finite set, the minimum is always achieved. When the minimum is achieved for multiple elements in \( A_l \), we break the ties arbitrarily. The generic chaining complexity \( \gamma_2 \) is defined as
\[
\gamma_2(T, d) := \inf \sup_{t \in T} \sum_{l=0}^\infty 2^{l/2} d(t, \pi_l(t)), \tag{3.21}
\]
where the infimum is over all admissible sequences. The following theorem tells us that \( \gamma_2 \)-functional controls the “size” of a Gaussian process.

**Lemma 3.5.2** (Theorem 2.4.1 of [Tal14b]). Let \( \{G(t), t \in T\} \) be a centered Gaussian process indexed by the set \( T \), and let
\[
d(s, t) = \mathbb{E}[(G(s) - G(t))^2]^{1/2}, \forall s, t \in T.
\]
Then, there exists a universal constant \( L \) such that
\[
\frac{1}{L} \gamma_2(T, d) \leq \mathbb{E}\left[\sup_{t \in T} G(t)\right] \leq L \gamma_2(T, d).
\]

Let \( (T, d) \) be a semi-metric space, and let \( X_1(t), \cdots, X_m(t) \) be independent stochastic processes indexed by \( T \) such that \( \mathbb{E}|X_j(t)| < \infty \) for all \( t \in T \) and \( 1 \leq j \leq m \). We are interested in
bounding the supremum of the empirical process

\[
Z_N(t) = \frac{1}{N} \sum_{i=1}^{N} \left[ X_i(t) - E[X_i(t)] \right].
\]  
(3.22)

The following well-known symmetrization inequality reduces the problem to bounds on a (conditionally) Rademacher process 

\[
R_N(t) = \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i X_i(t), \quad t \in T,
\]
where \(\varepsilon_1, \ldots, \varepsilon_m\) are i.i.d. Rademacher random variables (meaning that they take values \{-1, +1\} with probability \(1/2\) each), independent of \(X_i\)’s.

**Lemma 3.5.3** (Symmetrization inequalities).

\[
E \sup_{t \in T} |Z_N(t)| \leq 2E \sup_{t \in T} |R_N(t)|,
\]

and for any \(u > 0\), we have

\[
P \left( \sup_{t \in T} |Z_N(t)| \geq 2E \sup_{t \in T} |Z_N(t)| + u \right) \leq 4P \left( \sup_{t \in T} |R_N(t)| \geq u/2 \right).
\]

**Proof.** See Lemmas 6.3 and 6.5 in [LT91] \(\square\)

Finally, we recall Bernstein’s concentration inequality.

**Lemma 3.5.4** (Bernstein’s inequality). Let \(X_1, \ldots, X_m\) be a sequence of independent centered random variables. Assume that there exist positive constants \(\sigma\) and \(D\) such that for all integers \(p \geq 2\)

\[
\frac{1}{N} \sum_{i=1}^{N} E[|X_i|^p] \leq \frac{p!}{2} \sigma^2 D^{p-2},
\]

then

\[
P \left( \left| \frac{1}{N} \sum_{i=1}^{N} X_i \right| \geq \frac{\sigma}{\sqrt{N}} \sqrt{2u} + \frac{D}{N} u \right) \leq 2 \exp(-u).
\]

In particular, if \(X_1, \ldots, X_N\) are all sub-exponential random variables, then \(\sigma\) and \(D\) can be chosen as

\[
\sigma = \frac{1}{N} \sum_{i=1}^{N} \|X_i\|_{\psi_1} \quad \text{and} \quad D = \max_{i=1, \ldots, N} \|X_i\|_{\psi_1}.
\]

### 3.5.2 Roadmap of the proof of Theorem 3.3.1.

We outline the main steps in the proof of Theorem 3.3.1 and postpone some technical details to sections 3.5.4 and 3.5.5.
As it will be shown below in Lemma 3.5.5, \( \arg\min_{\theta} L^0(\theta) = \eta \theta_* \) for \( \eta = \mathbb{E}(\langle yx, \theta_* \rangle) \) and \( L^0(\tilde{\theta}_N) - L^0(\eta \theta_*) = \| \tilde{\theta}_N - \eta \theta_* \|^2_2 \), hence

\[
\| \tilde{\theta}_N - \eta \theta_* \|^2_2 = L^T(\tilde{\theta}_N) - L^T(\eta \theta_*) + \left( L^0(\tilde{\theta}_N) - L^0(\eta \theta_*) + L^T(\eta \theta_*) \right)
\]

\[
= L^T(\tilde{\theta}_N) - L^T(\eta \theta_*) + (L^T_N(\tilde{\theta}_N) - L^T_N(\eta \theta_*))
\]

\[
- (L^T_N(\tilde{\theta}_N) - L^T_N(\eta \theta_*)) - 2\mathbb{E}_N \left( yx - \bar{q} U, \tilde{\theta}_N - \eta \theta_* \right),
\]

(3.23)

where \( \mathbb{E}_N(\cdot) \) stands for the conditional expectation given \( (x_i, y_i)_{i=1}^N \), and where we used the equality \( L^0(\tilde{\theta}_N) - L^T(\tilde{\theta}_N) - L^0(\eta \theta_*) + L^T(\eta \theta_*) = -2\mathbb{E}_N \left( \langle yx - \bar{q} U, \tilde{\theta}_N - \eta \theta_* \rangle \right) \) in the last step. Since \( \tilde{\theta}_N \) minimizes \( L^T_N, L^T_N(\tilde{\theta}_N) - L^T_N(\eta \theta_*) \leq 0 \), and

\[
\| \tilde{\theta}_N - \eta \theta_* \|^2_2 \leq 2 \sum_{i=1}^N \left( \langle \tilde{q}_i U_i, \tilde{\theta}_N - \eta \theta_* \rangle - \mathbb{E}_N \left( \langle \tilde{q} U, \tilde{\theta}_N - \eta \theta_* \rangle \right) \right)
\]

\[
- 2\mathbb{E}_N \left( \langle yx - \bar{q} U, \tilde{\theta}_N - \eta \theta_* \rangle \right).
\]

Note that \( \tilde{\theta}_N - \eta \theta_* \in D(\Theta, \eta \theta_*) \); dividing both sides of the inequality by \( \| \tilde{\theta}_N - \eta \theta_* \|_2 \), we obtain

\[
\| \tilde{\theta}_N - \eta \theta_* \|_2 \leq \sup_{v \in D(\Theta, \eta \theta_*) \cap S_2(1)} \left| \frac{2}{N} \sum_{i=1}^N \langle \tilde{q}_i U_i, v \rangle - \mathbb{E} \left( \langle \tilde{q} U, v \rangle \right) + 2 \sup_{v \in S_2(1)} \mathbb{E} \left( \langle yx - \bar{q} U, v \rangle \right) \right|.
\]

(3.24)

To get the desired bound, it remains to estimate the two terms above. The bound for the first term is implied by Lemma 3.5.8 setting \( T = D(\Theta, \eta \theta_*) \cap S_2(1) \), and observing that the diameter \( \Delta_d(T) := \sup_{t \in T} \| t \|_2 = 1 \), we get that with probability \( \geq 1 - ce^{-\beta / 2} \),

\[
\sup_{v \in D(\Theta, \eta \theta_*) \cap S_2(1)} \left| \frac{2}{N} \sum_{i=1}^N \langle \tilde{q}_i U_i, v \rangle - \mathbb{E} \left( \langle \tilde{q} U, v \rangle \right) \right| \leq C \frac{(\omega(T) + 1)^{\beta}}{\sqrt{N}}.
\]

To estimate the second term, we apply Lemma 3.5.7

\[
2 \sup_{v \in S_2(1)} \mathbb{E} \left( \langle yx - \bar{q} U, v \rangle \right) \leq \frac{\tilde{C}}{\sqrt{N}}.
\]

Result of Theorem 3.3.1 now follows from the combination of these bounds. \( \square \)
3.5.3 Roadmap of the proof of Theorem 3.3.2.

Once again, we will present the main steps while skipping the technical parts. Lemma 3.5.5 implies that argmin \( \theta \in \Theta \) \( L^0(\theta) = \eta \theta_* \) for \( \eta = E \langle yx, \theta_* \rangle \) and

\[
L^0(\widehat{\theta}_N^\lambda) - L^0(\eta \theta_*) = \| \widehat{\theta}_N^\lambda - \eta \theta_* \|^2_2.
\]

Thus, arguing as in (3.23),

\[
\| \widehat{\theta}_N^\lambda - \eta \theta_* \|^2_2 = L^\tau(\widehat{\theta}_N^\lambda) - L^\tau(\eta \theta_*) + (L_N^\tau(\widehat{\theta}_N^\lambda) - L_N^\tau(\eta \theta_*))
- (L_N^\tau(\widehat{\theta}_N^\lambda) - L_N^\tau(\eta \theta_*)) - 2E_N \langle yx - \bar{q}U, \widehat{\theta}_N^\lambda - \eta \theta_* \rangle.
\]

Since \( \widehat{\theta}_N^\lambda \) is a solution of problem (3.11), it follows that

\[
L_N^\tau(\theta_N^\lambda) + \lambda \| \theta_N^\lambda \|_K \leq L_N^\tau(\eta \theta_*) + \lambda \| \eta \theta_* \|_K,
\]

which further implies that

\[
\| \widehat{\theta}_N^\lambda - \eta \theta_* \|^2_2 \leq \frac{2}{N} \sum_{i=1}^N \left( \langle \bar{q}_i \bar{U}_i - \bar{q}U, \widehat{\theta}_N^\lambda - \eta \theta_* \rangle - E_N \langle \bar{q}U, \widehat{\theta}_N^\lambda - \eta \theta_* \rangle \right) - 2E_N \langle yx - \bar{q}U, \widehat{\theta}_N^\lambda - \eta \theta_* \rangle
+ \lambda \left( \| \eta \theta_* \|_K - \| \widehat{\theta}_N^\lambda \|_K \right)
= \left( \frac{2}{N} \sum_{i=1}^N \bar{q}_i \bar{U}_i - E[\bar{q}U], \widehat{\theta}_N^\lambda - \eta \theta_* \right) - 2E_N \langle yx - \bar{q}U, \widehat{\theta}_N^\lambda - \eta \theta_* \rangle
+ \lambda \left( \| \eta \theta_* \|_K - \| \widehat{\theta}_N^\lambda \|_K \right).
\]

Letting \( \| \cdot \|_K^* \) be the dual norm of \( \| \cdot \|_K \) (meaning that \( \| \cdot \|_K^* = \sup \{ \langle x, z \rangle \mid \| z \|_K \leq 1 \} \)), the first term in (3.25) can be estimated as

\[
\left( \frac{1}{N} \sum_{i=1}^N \bar{q}_i \bar{U}_i - E[\bar{q}U], \widehat{\theta}_N^\lambda - \eta \theta_* \right) \leq \left\| \frac{1}{N} \sum_{i=1}^N \bar{q}_i \bar{U}_i - E[\bar{q}U] \right\|_K^* \cdot \| \widehat{\theta}_N^\lambda - \eta \theta_* \|_K.
\]

Since

\[
\left\| \frac{1}{N} \sum_{i=1}^N \bar{q}_i \bar{U}_i - E[\bar{q}U] \right\|_K^* = \sup_{\| t \|_K \leq 1} \left\langle \frac{1}{N} \sum_{i=1}^N \bar{q}_i \bar{U}_i - E[\bar{q}U], t \right\rangle,
\]

92
Lemma 3.5.8 applies with $T = \mathcal{G} := \{ x \in \mathbb{R}^d : \| x \|_K \leq 1 \}$. Together with an observation that $\Delta_d(T) \leq \sup_{t \in T} \| t \|_K = 1$ (due to the assumption $\| v \|_2 \leq \| v \|_K$, $\forall v \in \mathbb{R}^d$), this yields

$$
\mathbb{P} \left( \sup_{\| t \|_K \leq 1} \left\{ \frac{1}{N} \sum_{i=1}^{N} \bar{q}_i \bar{U}_i - \mathbb{E} \left[ \bar{q} \bar{U} \right] , t \right\} \geq C' (\omega(\mathcal{G}) + 1) \beta \sqrt{\frac{1}{N}} \right) \leq c' e^{-\beta/2},
$$

for any $\beta \geq 8$ and some constants $C', c > 0$. For the second term in (3.25), we use Lemma 3.5.7 to obtain

$$
2 \mathbb{E}_N \left( \| x - \bar{q} \bar{U}, \hat{\theta}_N - \eta \theta_* \right) \leq C'' \sqrt{\frac{1}{N}} \| \hat{\theta}_N - \eta \theta_* \|_2 \leq C'' \| \hat{\theta}_N - \eta \theta_* \|_K,
$$

for some constant $C'' > 0$, where we have again applied the inequality $\| v \|_2 \leq \| v \|_K$. Combining the above two estimates gives that with probability at least $1 - ce^{-\beta/2}$,

$$
\| \hat{\theta}_N - \eta \theta_* \|_2^2 \leq C (\omega(\mathcal{G}) + 1) \beta \sqrt{\frac{1}{N}} \| \hat{\theta}_N - \eta \theta_* \|_K + \lambda \left( \| \eta \theta_* \|_K - \| \hat{\theta}_N \|_K \right),
$$

(3.27)

for some constant $C > 0$ and any $\beta \geq 8$. Since $\lambda \geq 2C (\omega(\mathcal{G}) + 1) \beta / \sqrt{N}$ by assumption, and the right hand side of (3.27) is nonnegative, it follows that

$$
\frac{1}{2} \| \hat{\theta}_N - \eta \theta_* \|_K + \| \eta \theta_* \|_K - \| \hat{\theta}_N \|_K \geq 0.
$$

This inequality implies that $\hat{\theta}_N - \eta \theta_* \in S_2(\eta \theta_*)$. Finally, from (3.27) and the triangle inequality,

$$
\| \hat{\theta}_N - \eta \theta_* \|_2^2 \leq \frac{3}{2} \lambda \| \hat{\theta}_N - \eta \theta_* \|_K.
$$

Dividing both sides by $\| \hat{\theta}_N - \eta \theta_* \|_2$ gives

$$
\| \hat{\theta}_N - \eta \theta_* \|_2 \leq \frac{3}{2} \lambda \| \hat{\theta}_N - \eta \theta_* \|_K \leq \frac{3}{2} \lambda \cdot \Psi (S_2(\eta \theta_*)).
$$

This finishes the proof of Theorem 3.3.2.

3.5.4 Bias of the truncated mean.

The following lemma is motivated by and is similar to Theorem 2.1 in [LD89].
Lemma 3.5.5. Let \( \eta = \mathbb{E}\langle yx, \theta_* \rangle \). Then

\[
\eta \theta_* = \arg\min_{\theta \in \Theta} L^0(\theta),
\]

and for any \( \theta \in \Theta \),

\[
L^0(\theta) - L^0(\eta \theta_*) = \| \theta - \eta \theta_* \|_2^2.
\]

Proof. Since \( y = f((x, \theta_*), \delta) \), we have that for any \( \theta \in \mathbb{R}^d \)

\[
\mathbb{E} \langle yx, \theta \rangle = \mathbb{E}(\langle x, \theta \rangle f((x, \theta_*), \delta))
\]

where the third equality follows from the fact that the noise \( \delta \) is independent of the measurement vector \( x \), the second to last equality from the properties of elliptically symmetric distributions (Corollary 3.2.1), and the last equality from the definition of \( \eta \). Thus,

\[
L^0(\theta) = \| \theta \|_2^2 - 2\mathbb{E}[\langle yx, \theta \rangle] = \| \theta \|_2^2 - 2\eta(\theta, \theta_*)^2 = \| \theta - \eta \theta_* \|_2^2 - \| \eta \theta_* \|_2^2,
\]

which is minimized at \( \theta = \eta \theta^* \). Furthermore, \( L^0(\eta \theta^*) = -\| \eta \theta_* \|_2^2 \), hence

\[
L^0(\theta) - L^0(\eta \theta_*) = \| \theta - \eta \theta_* \|_2^2,
\]

finishing the proof. 

Next, we estimate the “bias term” \( \sup_{v \in S_2(1)} \mathbb{E} \langle yx - \tilde{q} \tilde{U}, v \rangle \) in inequality (3.24). In order to do so, we need the following preliminary result.

Lemma 3.5.6. If \( x \sim \mathcal{E}(0, I_{d \times d}, F_\mu) \), then the unit random vector \( x/\|x\|_2 \) is uniformly distributed over the unit sphere \( S_2(1) \). Furthermore, \( \tilde{U} = \sqrt{d} x/\|x\|_2 \) is a sub-Gaussian random vector with sub-Gaussian norm \( \| \tilde{U} \|_{\psi_2} \) independent of the dimension \( d \).
Proof. First, we use decomposition (3.3) for elliptical distribution together with our assumption that \( \Sigma \) is the identity matrix, to write \( x \overset{d}{=} \mu U \), which implies that

\[
x/\|x\|_2 \overset{d}{=} \text{sign}(\mu)U/\|U\|_2 = \text{sign}(\mu)U \overset{d}{=} U,
\]

with the final distributional equality holding as \( S_2(1) \), and hence its uniform distribution, is invariant with respect to reflections across any hyperplane through the origin.

To prove the second claim, it is enough to show that \( \| \langle \mathbf{U}, v \rangle \|_{\psi^2} \leq C, \forall v \in S_2(1) \) with constant \( C \) independent of \( d \). By the first claim and Lemma 3.5.1, we have

\[
P\left( \langle x, v \rangle/\|x\|_2 \geq \Delta \right) \leq e^{-d\Delta^2/2}, \forall v \in S_2(1).
\]

Choosing \( \Delta = u/\sqrt{d} \) gives

\[
P\left( \langle \mathbf{U}, v \rangle \geq u \right) \leq e^{-u^2/2}, \forall v \in S_2(1), \forall u > 0.
\]

By an equivalent definition of sub-Gaussian random variables (Lemma 5.5 of [Ver10a]), this inequality implies that \( \| \langle \mathbf{U}, v \rangle \|_{\psi^2} \leq C \), hence finishing the proof.

With the previous lemma in hand, we now establish the following result.

**Lemma 3.5.7.** Under the assumptions of Theorem 3.3.1, there exists a constant \( C = C(\kappa, \phi) > 0 \) such that

\[
\left| E \left( yx - q\mathbf{U}, v \right) \right| \leq C/\sqrt{N},
\]

for all \( v \in S_2(1) \).

Proof. By (3.6), we have that \( yx = q\mathbf{U} \), thus the claim is equivalent to

\[
\left| E \left( \langle \mathbf{U}, v \rangle (\mathbf{q} - q) \right) \right| \leq C/\sqrt{N}.
\]
Since $\tilde{q} = \text{sign}(q)(|q| \wedge \tau)$, we have $|\tilde{q} - q| = (|q| - \tau)1(|q| \geq \tau) \leq |q|1(|q| \geq \tau)$, and it follows that

$$\left| \mathbb{E} \left\langle \tilde{U}, \nu \right\rangle (\tilde{q} - q) \right| \leq \mathbb{E} \left| \left\langle \tilde{U}, \nu \right\rangle \tilde{q} - q \right| \leq \mathbb{E} \left( \left| \left\langle \tilde{U}, \nu \right\rangle q \cdot 1_{(|q| \geq \tau)} \right| \right) \leq \mathbb{E} \left[ \left| \left\langle \tilde{U}, \nu \right\rangle q \right|^2 \right]^{1/2} \mathbb{P} \left( |q| \geq \tau \right)^{1/2} \leq \mathbb{E} \left[ \left| \left\langle \tilde{U}, \nu \right\rangle \left( 2(1+\kappa) \right) \right|^ {-\frac{\kappa}{2(1+\kappa)}} \left\lbrace \frac{1}{2} \leq \frac{\phi}{\sqrt{N}} \right\rbrace \right] \leq \frac{2(1 + \kappa)}{\kappa} \mathbb{P} \left( |q| \geq \tau \right)^{1/2},$$

where the second to last inequality uses Cauchy-Schwarz, and the last inequality follows from Hölder’s inequality.

For the first term, by Lemma 3.5.6, $\tilde{U}$ is sub-Gaussian with $\|\tilde{U}\|_{\psi_2}$ independent of $d$. Thus, by the definition of the $\| \cdot \|_{\psi_2}$ norm and the fact that $\nu \in S_2(1)$,

$$\mathbb{E} \left[ \left| \left\langle \tilde{U}, \nu \right\rangle \right|^{2(1+\kappa)} \right] \leq \frac{2(1 + \kappa)}{\kappa} \mathbb{E} \left| \frac{\phi}{\sqrt{N}} \right| = \frac{2(1 + \kappa)}{\kappa} \|\tilde{U}\|_{\psi_2}.$$

Recall that $\phi = \mathbb{E}|q|^{2(1+\kappa)}$. Then, the second term is bounded by $\phi^{2(1+\kappa)}$. For the final term, since $\tau = m \frac{1}{\kappa}$, Markov’s inequality implies that

$$(\mathbb{P} \left( |q| > \tau \right))^{1/2} \leq \left( \frac{\mathbb{E}|q|^{2(1+\kappa)}}{\tau^{2(1+\kappa)}} \right)^{1/2} \leq \frac{\phi^{1/2}}{\sqrt{N}}.$$}

Combining these inequalities yields

$$\left| \mathbb{E} \left\langle yx - \tilde{q}\tilde{U}, \nu \right\rangle \right| \leq \sqrt{\frac{2(1+\kappa)}{\kappa} \|\tilde{U}\|_{\psi_2} \phi^{2+\kappa}} \frac{\phi^{1/2}}{\sqrt{N}} = C(\kappa, \phi) / \sqrt{N},$$

completing the proof.

### 3.5.5 Concentration via generic chaining.

In the following sections, we will use $c, C, C'$ to denote constants that are either absolute, or depend on underlying parameters $\kappa$ and $\phi$ (in the latter case, we specify such dependence). To make notation less cumbersome, constants denoted by the same letter ($c, C, C'$, etc.) might be different in various parts of the proof.
The goal of this subsection is to prove the following inequality:

**Lemma 3.5.8.** Suppose $\bar{U}_i$ and $\bar{q}_i$ are as defined according to (3.6) and (3.7) respectively. Then, for any bounded subset $T \subset \mathbb{R}^d$,

$$
\mathbb{P}\left( \sup_{t \in T} \left| \frac{1}{N} \sum_{i=1}^{N} \langle \bar{U}_i, t \rangle \bar{q}_i - \mathbb{E}\left[ \langle \bar{U}, t \rangle \bar{q} \right] \right| \geq C \frac{(\omega(T) + \Delta_d(T))\beta}{\sqrt{N}} \right) \leq ce^{-\beta/2},
$$

for any $\beta \geq 8$, a positive constant $C = C(\kappa, \phi)$ and an absolute constant $c > 0$. Here

$$
\Delta_d(T) := \sup_{t \in T} \|t\|_2.
$$

(3.28)

The main technique we apply is the generic chaining method developed by M. Talagrand [Tal14b] for bounding the supremum of stochastic processes. Recently, [MPTJ07] and [Dir13] advanced the technique to obtain a sharp bound for supremum of processes indexed by squares of functions. More recently, [Men14b] proved a concentration result for the supremum of multiplier processes under weak moment assumptions. In the current work, we show that exponential-type concentration inequalities for multiplier processes, such as the one in Lemma 3.5.8, are achievable by applying truncation under a bounded $2(1 + \kappa)$-moment assumption.

Define

$$
Z(t) = \frac{1}{N} \sum_{i=1}^{N} \langle \bar{U}_i, t \rangle \bar{q}_i - \mathbb{E}\left[ \langle \bar{U}, t \rangle \bar{q} \right],
$$

$$
Z(t) = \frac{1}{N} \sum_{i=1}^{N} \xi_i \bar{q}_i \langle \bar{U}_i, t \rangle, \ \forall t \in T,
$$

where $T$ is a bounded set in $\mathbb{R}^d$ and $\{\xi_i\}_{i=1}^{m}$ is a sequence i.i.d. Rademacher random variables taking values $\pm 1$ with probability $1/2$ each, and independent of $\{\bar{U}_i, \bar{q}_i, i = 1, \ldots, m\}$. Result of Lemma 3.5.8 easily follows from the following concentration inequality:

**Lemma 3.5.9.** For any $\beta \geq 8$,

$$
\mathbb{P}\left( \sup_{t \in T} |Z(t)| \geq C \frac{(\omega(T) + \Delta_d(T))\beta}{\sqrt{N}} \right) \leq ce^{-\beta/2},
$$

(3.29)

where $C = C(\kappa, \phi)$ is another constant possibly different from that of Lemma 3.5.8, and $c > 0$ is an absolute constant.
To deduce the inequality of Lemma 3.5.8, we first apply the symmetrization inequality (Lemma 3.5.3, followed by Lemma 3.6.1 with $\beta_0 = 8$. It implies that

$$E\left[\sup_{t \in T} |Z(t)|\right] \leq 2E\left[\sup_{t \in T} |Z(t)|\right] \leq 2C \left(8 + 2e^{-4}\right) \frac{\omega(T) + \Delta d(T)}{\sqrt{N}}.$$  

Application of the second bound of the symmetrization lemma with $u = 2C(\omega(T) + \Delta d(T))\beta/\sqrt{N}$ and (3.29) completes the proof of Lemma 3.5.8.

It remains to justify (3.29). We start by picking an arbitrary point $t_0 \in T$ such that there exists an admissible sequence $\{t_0\} = A_0 \subseteq A_1 \subseteq A_2 \subseteq \cdots$ satisfying

$$\sup_{t \in T} \|\pi_l(t) - t\|_2 \leq 2\gamma_2(T), \quad (3.30)$$

where we recall that $\pi_l$ is the closest point map from $T$ to $A_l$ and the factor 2 is introduced so as to deal with the case where the infimum in the definition (3.21) of $\gamma_2(T)$ is not achieved. Then, write $Z(t) - Z(t_0)$ as the telescoping sum:

$$Z(t) - Z(t_0) = \sum_{l=1}^{\infty} Z(\pi_l(t)) - Z(\pi_{l-1}(t)) = \sum_{l=1}^{\infty} \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i \bar{q}_i \left(\bar{U}_i, \pi_l(t) - \pi_{l-1}(t)\right).$$

We claim that the telescoping sum converges with probability 1 for any $t \in T$. Indeed, note that for each fixed set of realizations of $\{x_{i_{1}}\}_{i=1}^{N}$ and $\{\varepsilon_{i_{1}}\}_{i=1}^{N}$, each summand is bounded as

$$|\varepsilon_i \bar{q}_i(\bar{U}_i, \pi_l(t) - \pi_{l-1}(t))| \leq |\bar{q}_i||\bar{U}_i||\pi_l(t) - \pi_{l-1}(t)||_2 \leq |\bar{q}_i||\bar{U}_i||(\pi_l(t) - t)||_2 + ||\pi_{l-1}(t) - t||_2.$$  

Furthermore, since $T$ is a compact subset of $\mathbb{R}^d$, its Gaussian mean width is finite. Thus, by lemma 3.5.2, $\gamma_2(T) \leq L\omega(T) < \infty$. This inequality further implies that the sum on the left hand side of (3.30) converges with probability 1.

Next, with $\beta \geq 8$ being fixed, we split the index set $\{l \geq 1\}$ into the following three subsets:

$$I_1 = \{l \geq 1 : 2^l \beta < \log eN\};$$

$$I_2 = \{l \geq 1 : \log eN \leq 2^l \beta < N\};$$

$$I_3 = \{l \geq 1 : 2^l \beta \geq N\}.$$
By the assumptions in Theorem 3.3.1 and the bound $\beta \geq 8$, we have that $m \geq (\omega(T) + 1)^2 \beta^2 \geq 64$, implying that $\log eN = 1 + \log N < N$, and hence these three index sets are well defined. Depending on $\beta$, some of them might be empty, but this only simplifies our argument by making the partial sum over such an index set equal 0.

The following argument yields a bound for $Z(\pi_l(t)) - Z(\pi_{l-1}(t))$, assuming all three index sets are nonempty. Specifically, we show that

$$
\mathbb{P} \left( \sup_{t \in T} \left| \sum_{l \in I_j} (Z(\pi_l(t)) - Z(\pi_{l-1}(t))) \right| \geq \frac{C \gamma_2(T) \beta \sqrt{N}}{\sqrt{N}} \right) \leq e^{-\beta/2}, \quad (3.31)
$$

for $C = C(\kappa, \phi)$ and $j = 1, 2, 3$, respectively.

The case $l \in I_1$.

Proof of inequality (3.31) for the index set $I_1$. Recall that $\tau = N^{2(1+\kappa)}$.

For each $t \in T$ we apply Bernstein’s inequality (Lemma 3.5.4) to estimate each summand

$$
Z(\pi_l(t)) - Z(\pi_{l-1}(t)) = \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i \widehat{q}_i \left\langle \tilde{U}_i, \pi_l(t) - \pi_{l-1}(t) \right\rangle.
$$

For any integer $p \geq 2$, we have the following chains of inequalities:

$$
\mathbb{E} \left[ \left\langle \tilde{U}_i, \pi_l(t) - \pi_{l-1}(t) \right\rangle^p \right] \\
\leq \mathbb{E} \left[ \varepsilon_i \left\langle \tilde{U}_i, \pi_l(t) - \pi_{l-1}(t) \right\rangle^p \cdot |\tilde{q}_i|^{p-2} \right] \\
\leq \mathbb{E} \left[ \left\langle \tilde{U}_i, \pi_l(t) - \pi_{l-1}(t) \right\rangle^p \cdot \tau^{p-2} \right] \\
\leq \tau^{p-2} \mathbb{E} \left[ \left\langle \tilde{U}_i, \pi_l(t) - \pi_{l-1}(t) \right\rangle^{1 + \kappa} \right] \frac{\kappa}{1 + \kappa} \left( \mathbb{E} \left[ \tilde{q}_i^{2(1+\kappa)} \right] \right)^{\frac{1}{1 + \kappa}} \\
\leq \tau^{p-2} \left\| \tilde{U}_i \right\|_{\psi_2}^p \left( \frac{(1 + \kappa)p}{\kappa} \right)^{p/2} \phi \frac{1}{1 + \kappa} \left\| \pi_l(t) - \pi_{l-1}(t) \right\|_2,
$$

where the second inequality follows from the truncation bound, the third from Hölder’s inequality, and the last from the assumption that $\mathbb{E} \left[ \tilde{q}_i^{2(1+\kappa)} \right] \leq \phi$ and the following bound: by Lemma 3.5.6, $\tilde{U}_i$ is sub-Gaussian, hence for any $p \geq 2$

$$
\left( \mathbb{E} \left\langle \tilde{U}_i, \nu \right\rangle^{1 + \kappa} \right)^{\frac{1}{1 + \kappa}} \leq \left( \frac{(1 + \kappa)p}{\kappa} \right)^{\frac{1}{2}} \left\| \tilde{U}_i \right\|_{\psi_2} \left\| \nu \right\|_2, \quad \forall \nu \in \mathbb{R}^d.
$$
We also note that \( \|\widetilde{U}_l\|_{\psi,2} \) does not depend on \( d \) by Lemma 3.5.6. Next, by Stirling’s approximation, \( p! \geq \sqrt{2\pi} \sqrt{p} (p/e)^p \), thus there exist constants \( C' = C'(\kappa, \phi) \) and \( C'' = C''(\kappa) \) such that

\[
E \left| \varepsilon \hat{q} \langle \widetilde{U}, \pi_l(t) - \pi_{l-1}(t) \rangle \right|^p \leq \frac{p!}{2} C' \| \pi_l(t) - \pi_{l-1}(t) \|_2^2 (C'' \tau \| \pi_l(t) - \pi_{l-1}(t) \|_2)^{p-2}.
\]

Bernstein’s inequality (Lemma 3.5.4), with \( \sigma = C' \| \pi_l(t) - \pi_{l-1}(t) \|_2, D = C'' \tau \| \pi_l(t) - \pi_{l-1}(t) \|_2 \) with \( \tau = N^{1/2(1+\kappa)} \) now implies

\[
\mathbb{P} \left( \left| \sum_{i=1}^N \varepsilon_i \hat{q}_i \langle \widetilde{U}_i, \pi_l(t) - \pi_{l-1}(t) \rangle \right| \geq \left( \frac{C' \sqrt{2u}}{\sqrt{N}} + \frac{C'' u}{m^{1-\frac{1}{2(1+\kappa)}}} \right) \| \pi_l(t) - \pi_{l-1}(t) \|_2 \right) \leq 2e^{-u},
\]

for any \( u > 0 \). Taking \( u = 2^l \beta \), noting that as \( \beta \geq 8 \) by assumption, we have \( m \geq (\omega(T)+1)^2 \beta^2 \geq 64 \), and since \( l \in I_1 \), \( 2^l \leq 2^l \beta < \log em \). In turn, this implies

\[
\frac{2^l}{m^{1-\frac{1}{2(1+\kappa)}}} = \frac{2^{l/2}}{m^{1/2}} \frac{2^{l/2}}{m^{\kappa/(1+\kappa)}} \leq \frac{2^{l/2}}{m^{1/2}} \sqrt{\frac{\log em}{m^{\kappa/(1+\kappa)}}} \leq \sqrt{\frac{1+\kappa}{\kappa} \frac{2^{l/2}}{m^{1/2}}} = \sqrt{\frac{1+\kappa}{\kappa} \frac{2^{l/2}}{m^{1/2}}},
\]

where the last inequality follows from the fact that \( \log em \) is dominated by \( \frac{1+\kappa}{\kappa} m^{\kappa/(1+\kappa)} \) for all \( m \geq 1 \). This inequality implies that there exists a positive constant \( C = C(\kappa, \phi) \) such that for any \( \beta \geq 8 \)

\[
\mathbb{P} (\Omega_{l,t}) \leq 2 \exp(-2^l \beta), \tag{3.32}
\]

where for all \( l \geq 1 \) and \( t \in T \) we let

\[
\Omega_{l,t} = \left\{ \omega : \left| \frac{1}{N} \sum_{i=1}^N \varepsilon_i \hat{q}_i \langle \widetilde{U}_i, \pi_l(t) - \pi_{l-1}(t) \rangle \right| \geq C \frac{2^{l/2} \beta}{\sqrt{N}} \| \pi_l(t) - \pi_{l-1}(t) \|_2 \right\}.
\]

Notice that for each \( l \geq 1 \) the number of pairs \( (\pi_l(t), \pi_{l-1}(t)) \) appearing in the sum in (3.31) can be bounded by \( |\mathcal{A}_l| : |\mathcal{A}_{l-1}| \leq 2^{2^l+1} \). Thus, by a union bound and (3.32),

\[
\mathbb{P} \left( \bigcup_{t \in T} \Omega_{l,t} \right) \leq 2 \cdot 2^{2^l+1} \exp(-2^l \beta),
\]

100
and hence,

\[
P \left( \bigcup_{l \in I_1, t \in T} \Omega_{l,t} \right) \leq \sum_{l \in I_1} 2 \cdot 2^{2l+1} \exp(-2^l \beta) \\
\leq \sum_{l \in I_1} 2 \cdot 2^{2l+1} \exp \left(-2^{l-1} \beta - \beta/2 \right) \leq c e^{-\beta/2},
\]

for some absolute constant \( c > 0 \), where in the last inequality we use the fact \( \beta \geq 8 \) to get a geometrically decreasing sequence. Thus, on the complement of the event \( \bigcup_{l \in I_1, t \in T} \Omega_{l,t} \), we have that with probability at least \( 1 - ce^{-\beta/2} \),

\[
\sup_{t \in T} \left| \sum_{l \in I_1} (Z(\pi_l(t)) - Z(\pi_{l-1}(t))) \right| \leq \sup_{t \in T} \sum_{l \in I_1} |Z(\pi_l(t)) - Z(\pi_{l-1}(t))| \\
\leq \sup_{t \in T} C \sum_{l \in I_1} 2^{l/2} \beta \frac{\|\pi_l(t) - \pi_{l-1}(t)\|_2}{\sqrt{N}} \\
\leq \sup_{t \in T} C \sum_{l=1}^{\infty} 2^{l/2} \beta \frac{\|\pi_l(t) - \pi_{l-1}(t)\|_2}{\sqrt{N}} \\
\leq 4C \gamma_2(T) \beta \frac{\sqrt{N}}{\sqrt{N}},
\]

for \( C = C(\kappa, \phi) \), where the last inequality follows from triangle inequality \( \|\pi_l(t) - \pi_{l-1}(t)\|_2 \leq \|\pi_l(t) - t\|_2 + \|\pi_t(t) - t\|_2 \) and (3.30). This proves the inequality (3.31) for \( l \in I_1 \). \( \square \)

**The case \( l \in I_2 \).**

This is the most technically involved case of the three. For any fixed \( t \in T \) and \( l \in I_2 \), we let \( X_i = \bar{q}_i \langle \bar{U}_l, \pi_l(t) - \pi_{l-1}(t) \rangle \) and \( w_i = \langle \bar{U}_l, \pi_l(t) - \pi_{l-1}(t) \rangle \). Then \( X_i = \bar{q}_i w_i \) and

\[
Z(\pi_l(t)) - Z(\pi_{l-1}(t)) = \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i X_i = \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i w_i \bar{q}_i. \tag{3.33}
\]

For every fixed \( k \in \{1, 2, \cdots, N - 1\} \) and fixed \( u > 0 \), we bound the summation using the following inequality

\[
P \left( \left| \sum_{i=1}^{N} \varepsilon_i X_i \right| \geq \sum_{i=k+1}^{N} X_i^* + u \left( \sum_{i=k+1}^{N} (X_i^*)^2 \right)^{1/2} \right) \leq 2 \exp(-u^2/2),
\]

101
where \( \{X_i^*\}_{i=1}^N \) is the non-increasing rearrangement of \( \{|X_i|\}_{i=1}^N \) and \( \{\varepsilon_i\}_{i=1}^N \) is a sequence of i.i.d. Rademacher random variables independent of \( \{X_i\}_{i=1}^N \).

**Remark 3.5.3.** This bound was first stated and proved in [MS90b] with a sequence of fixed constants \( \{X_i\}_{i=1}^N \). The current form can be obtained using independence property and conditioning on \( \{X_i\}_{i=1}^N \). Furthermore, [MS90b] tells us that the optimal choice of \( k \) is at \( O(u^2) \). Applications of this inequality to generic chaining-type arguments were previously introduced by [Men14b].

Letting \( J \) be the set of indices of the variables corresponding to the \( k \) largest coordinates of \( \{|w_i|\}_{i=1}^m \) and of \( \{|q_i|\}_{i=1}^m \), we have \( |J| \leq 2k \) and with probability at least \( 1 - 2 \exp(-u^2/2) \)

\[
\left| \sum_{i=1}^N \varepsilon_i X_i \right| \leq \sum_{i \in J} X_i^* + u \left( \sum_{i \in J^c} (X_i^*)^2 \right)^{1/2} \\
\leq 2 \sum_{i=1}^k w_i^* q_i^* + u \left( \sum_{i \in J^c} (w_i^*)^2 \right)^{1/2} \\
\leq 2 \left( \sum_{i=1}^k (w_i^*)^2 \right)^{1/2} \left( \sum_{i=1}^k (q_i^*)^2 \right)^{1/2} + u \left( \sum_{i=k+1}^N (w_i^*)^{2(1+\kappa)} \right)^{\frac{\kappa}{2(1+\kappa)}} \left( \sum_{i=k+1}^N (q_i^*)^{2(1+\kappa)} \right)^{\frac{1}{2(1+\kappa)}} \\
\leq 2 \left( \sum_{i=1}^k (w_i^*)^2 \right)^{1/2} \left( \sum_{i=1}^N (q_i^*)^2 \right)^{1/2} + u \left( \sum_{i=k+1}^N (w_i^*)^{2(1+\kappa)} \right)^{\frac{\kappa}{2(1+\kappa)}} \left( \sum_{i=1}^N (q_i^*)^{2(1+\kappa)} \right)^{\frac{1}{2(1+\kappa)}} \\
\tag{3.34}
\]

where the second to last inequality is a consequence of Hölder’s inequality. We take \( u = 2^{l+1}/\sqrt{3} \). The key is to pick an appropriate cut point \( k \) for each \( l \in I_2 \). Here, we choose \( k = \lfloor 2^l \beta / \log(eN/2^l \beta) \rfloor \), which makes \( k = O(2^l \beta) \) and also guarantees that \( k \in \{1, 2, \cdots, N-1\} \); see Lemma 4.19. Under this choice, we have the following lemma:

**Lemma 3.5.10.** Let \( k = \lfloor 2^l \beta / \log(eN/2^l \beta) \rfloor \), \( w_i = \langle \tilde{U}_i, \pi_l(t) - \pi_{l-1}(t) \rangle \) and \( \{w_i^*\}_{i=1}^N \) be the nonincreasing rearrangement of \( \{|w_i|\}_{i=1}^N \). Then there exists an absolute constant \( C > 1 \) such that for all \( \beta \geq 8 \),

\[
P \left( \left( \sum_{i=1}^k (w_i^*)^2 \right)^{1/2} \geq C 2^{l/2} \|\pi_l(t) - \pi_{l-1}(t)\|_2 \sqrt{\beta} \right) \leq 2 \exp(-2^l \beta).
\]

**Proof.** By Lemma 3.5.6 we know that \( \{w_i\}_{i=1}^N \) are i.i.d. sub-Gaussian random variables. Thus,
by Lemma 3.6.2, $w_i^2$ is sub-exponential with norm

$$\|w_i^2\|_{\psi_1} = 2\|w_i\|^2_{\psi_2} \leq 2\|\widetilde{U}_i\|^2_{\psi_2}\|\pi_1(t) - \pi_{l-1}(t)\|_2^2. \quad (3.35)$$

It then follows from Bernstein’s inequality (Lemma 3.5.4) that for any fixed set $J \subseteq \{1, 2, \cdots, N\}$ with $|J| = k,$

$$\mathbb{P}\left(\left| \frac{1}{k} \sum_{i \in J} (w_i^2 - \mathbb{E}[w_i^2]) \right| \geq 2\|\widetilde{U}_i\|^2_{\psi_2}\|\pi_1(t) - \pi_{l-1}(t)\|_2^2 \left(\sqrt{\frac{2u}{k}} + \frac{u}{k}\right) \right) \leq 2\text{exp}(-u).$$

We choose $u = 4 \cdot 2^l \beta = 2^{l+2} \beta.$ Since $2^l \beta \geq [2^l \beta / \log(eN/2^l \beta)] = k \geq 1,$ the factor $u/k$ dominates the right hand side. Noting that $\mathbb{E}[w_i^2] = \|\pi_1(t) - \pi_{l-1}(t)\|_2^2,$ we obtain

$$\mathbb{P}\left(\left(\sum_{i \in J} w_i^2\right)^{1/2} \geq C2^{l/2}\|\pi_1(t) - \pi_{l-1}(t)\|_2 \sqrt{\beta}\right) \leq 2\text{exp}(-4 \cdot 2^l \beta),$$

where $C \leq 4\|\widetilde{U}_i\|_{\psi_2};$ note that the upper bound for $C$ is independent of $d$ by Lemma 3.5.1. Thus,

$$\mathbb{P}\left(\left(\sum_{i=1}^{N} (w_i^*)^2\right)^{1/2} \geq C2^{l/2}\|\pi_1(t) - \pi_{l-1}(t)\|_2 \sqrt{\beta}\right)$$

$$\mathbb{P}\left(\exists J \subseteq \{1, \cdots, N\}, |J| = k : \left(\sum_{i \in J} w_i^2\right)^{1/2} \geq C2^{l/2}\|\pi_1(t) - \pi_{l-1}(t)\|_2 \sqrt{\beta}\right)$$

$$\leq \binom{N}{k} \cdot \mathbb{P}\left(\left(\sum_{i \in J} w_i^2\right)^{1/2} \geq C2^{l/2}\|\pi_1(t) - \pi_{l-1}(t)\|_2 \sqrt{\beta}\right)$$

$$\leq 2\left(\frac{N}{k}\right)^k \exp(-4 \cdot 2^l \beta)$$

$$\leq 2\left(\frac{eN}{k}\right)^k \exp(-4 \cdot 2^l \beta) \leq 2\exp(-2^l \beta),$$

where the last step follows from $\left(\frac{eN}{k}\right)^k \leq \exp(3 \cdot 2^l \beta),$ an inequality proved in Appendix 3.6. \hfill \Box

**Lemma 3.5.11.** Let $k = \lfloor 2^l \beta / \log(eN/2^l \beta) \rfloor,$ $w_i = \left\langle \widetilde{U}_i, \pi_1(t) - \pi_{l-1}(t) \right\rangle$ and $\{w_i^*\}_{i=1}^N$ be the non-increasing rearrangement of $\{\|w_i\|\}_{i=1}^N.$ Then

$$\mathbb{P}\left(\left(\sum_{i=k+1}^{N} (w_i^*)^2 \right)^{1/2} \pi_{\frac{i+k}{N}} \leq C(\kappa)N \pi_{\frac{i+k}{N}} \|\pi_1(t) - \pi_{l-1}(t)\|_2 \right) \leq \exp(-2^l \beta),$$

103
for any $\beta \geq 8$ and some constant $C(\kappa) > 0$.

**Proof.** To avoid possible confusion, we use $i$ to index the nonincreasing rearrangement and $j$ for the original sequence. We start by noting that $\{w_j\}_{j=1}^m$ are i.i.d. sub-Gaussian random variables with $\|w_j\|_{\psi_2} \leq \|\tilde{U}_j\|_{\psi_2} \|\pi_l(t) - \pi_{l-1}(t)\|_2$. By an equivalent definition of sub-Gaussian random variables (Lemma 5.5. of [Ver10a]), we have for any fixed $j \in \{1, 2, \ldots, N\}$,

$$P \left( |w_j| - E[|w_j|] \geq Cu\|\tilde{U}_j\|_{\psi_2} \|\pi_l(t) - \pi_{l-1}(t)\|_2 \right) \leq e^{-u^2}, \quad (3.36)$$

for any $u > 0$ and an absolute constant $C > 0$.

To establish the claim of the lemma, we bound each $w^*_i$ separately for $i = 1, 2, \ldots, m$ and then combine individual bounds. Instead of using a fixed value of $u$ in (3.36), our choice of $u$ will depend on the index $i$. Specifically, for each $w^*_i$, we choose $u = c_\kappa (N/i)^{\kappa/4(1+\kappa)}$ with

$$c_\kappa := \max \left\{ \sqrt{5} \left( 2 + \frac{1}{N} \right)^{\frac{3(1+\kappa)}{2+\kappa}} e^{1/2(1+\kappa)} \sqrt{\frac{4(1+\kappa)}{\kappa}} \right\}. \quad (3.37)$$

The reason for this choice will be clear as we proceed.

First, for a fixed nonincreasing rearrangement index $i > k$, by (3.36) and the fact that

$$E[|w_j|] \leq E[w^2_j]^{1/2} = \|\pi_l(t) - \pi_{l-1}(t)\|_2, \forall j \in \{1, 2, \ldots, N\},$$

we have

$$P \left( |w_j| \geq \left( 1 + Cc_\kappa \|\tilde{U}_j\|_{\psi_2} \right) \left( \frac{N}{i} \right)^{\frac{\kappa + 1}{2 + \kappa}} \|\pi_l(t) - \pi_{l-1}(t)\|_2 \right) \leq \exp \left( -c^2_\kappa \left( \frac{N}{i} \right)^{\frac{2(1+\kappa)}{2+\kappa}} \right),$$

$$\forall j \in \{1, 2, \ldots, N\}.$$

To simplify notation, let $C' = 1 + Cc_\kappa \|\tilde{U}_j\|_{\psi_2}$ (note that it depends only on $\kappa$). It then follows
that

\[
\mathbb{P}\left( w_i^* \geq C' \left( \frac{N}{i} \right)^{\frac{n}{1+\kappa}} \|\pi_i(t) - \pi_{i-1}(t)\|_2 \right) \\
\leq \mathbb{P}\left( \exists J \subseteq \{1, \ldots, N\}, |J| = i : w_j \geq C' \left( \frac{N}{i} \right)^{\frac{n}{1+\kappa}} \|\pi_i(t) - \pi_{i-1}(t)\|_2, \forall j \in J \right) \\
\leq \left( \frac{N}{k} \right) \mathbb{P}\left( |w_j| \geq C' \left( \frac{N}{i} \right)^{\frac{n}{1+\kappa}} \|\pi_i(t) - \pi_{i-1}(t)\|_2 \right)^i \\
\leq \left( \frac{N}{k} \right) \exp \left( -c^2 N \frac{n}{1+\kappa} i^{\frac{n}{1+\kappa}} \right) \\
\leq \left( \frac{eN}{i} \right)^i \exp \left( -c^2 N \frac{n}{1+\kappa} i^{\frac{n}{1+\kappa}} \right).
\]

By a union bound, we have

\[
\mathbb{P}\left( \exists i > k : w_i^* \geq C' \left( \frac{N}{i} \right)^{\frac{n}{1+\kappa}} \|\pi_i(t) - \pi_{i-1}(t)\|_2 \right) \\
\leq \sum_{i=k+1}^{N} \left( \frac{eN}{i} \right)^i \exp \left( -c^2 N \frac{n}{1+\kappa} i^{\frac{n}{1+\kappa}} \right) \\
= \sum_{i=k+1}^{N} \exp \left( i \log \left( \frac{eN}{i} \right) - c^2 N \frac{n}{1+\kappa} i^{\frac{n}{1+\kappa}} \right) \\
\leq N \cdot \exp \left( k \log \left( \frac{eN}{k} \right) - c^2 N \frac{n}{1+\kappa} k^{\frac{n}{1+\kappa}} \right) \\
\leq \exp \left( 4 \cdot 2^i \beta - c^2 N \frac{n}{1+\kappa} k^{\frac{n}{1+\kappa}} \right),
\]

where the second to last inequality follows since by the definition of $c_\kappa$, $c_\kappa \geq \sqrt{4(1 + \kappa)/\kappa}$, the function $v(i) = i \log \left( \frac{eN}{i} \right) - c^2 N \frac{n}{1+\kappa} i^{\frac{n}{1+\kappa}}$ is monotonically decreasing with respect to $i$ (recall that $i \leq N$), and thus is dominated by $v(k)$. The final inequality follows from Lemma 4.18 as well as the fact that $\log N \leq \log(eN) \leq 2^i \beta$. Furthermore, by Lemma 4.19 in the Appendix 3.6 and (3.37) implying $c_\kappa \geq \sqrt{5 (2 + \frac{4}{\kappa}) \frac{2+n}{2+\kappa} / e^{1/2(1+\kappa)}}$, we have

\[
c^2_\kappa N \frac{n}{1+\kappa} k^{\frac{n}{1+\kappa}} \geq 5 \cdot 2^i \beta.
\]

Overall, we have the following bound:

\[
\mathbb{P}\left[ \exists i > k : w_i^* \geq C' \left( \frac{N}{i} \right)^{\frac{n}{1+\kappa}} \|\pi_i(t) - \pi_{i-1}(t)\|_2 \right] \leq \exp \left( 4 \cdot 2^i \beta - 5 \cdot 2^i \beta \right) \leq \exp(-2^i \beta).
\]
Thus, with probability at least $1 - \exp(-2^l \beta)$,

$$w_i^* \leq C' \left( \frac{N}{t} \right)^{\frac{n}{2(1+\kappa)}} \| \pi_l(t) - \pi_{l-1}(t) \|_2, \ \forall i > k,$$

hence with the same probability

$$\left( \sum_{j=k+1}^{N} (w_j^*)^{\frac{2(1+\kappa)}{n}} \right)^{\frac{n}{2(1+\kappa)}} \leq C'' \| \pi_l(t) - \pi_{l-1}(t) \|_2 \left( \sum_{i=k+1}^{N} \left( \frac{N}{t} \right)^{1/2} \right)^{\frac{n}{2(1+\kappa)}} \leq C'' \| \pi_l(t) - \pi_{l-1}(t) \|_2 m^{\frac{1}{2(1+\kappa)}} \left( \int_{1}^{m} \frac{dx}{x^{1/2}} \right)^{\frac{n}{2(1+\kappa)}} \leq 2^{\frac{n}{2(1+\kappa)}} C'' \| \pi_l(t) - \pi_{l-1}(t) \|_2 N^{\frac{1}{2(1+\kappa)}},$$

and the desired result follows. \qed

**Lemma 3.5.12.** The following inequalities hold for any $\beta \geq 8$:

$$P \left( \left( \sum_{i=1}^{N} q_i^2 \right)^{1/2} \geq C' \sqrt{\beta N} \right) \leq 2e^{-\beta},$$

$$P \left( \sum_{i=1}^{N} \frac{q_i^{2(1+\kappa)}}{\beta^{1+\kappa}} \geq C'' (\beta N)^{\frac{1}{2(1+\kappa)}} \right) \leq 2e^{-\beta},$$

for some positive constants $C' = C'(\phi, \kappa), \ C'' = C''(\phi, \kappa)$.

**Proof.** Recall that $\tilde{q}_i = \text{sign}(q_i)(|q_i| \wedge \tau), \tau = N^{1/2(1+\kappa)}$, and $\phi = E[q_i^{2(1+\kappa)}]$. Thus, $E[q_i^2] \leq E[q_i^{2(1+\kappa)}] \leq \phi^{1/1+\kappa}$, and for any integer $p \geq 2$, we have

$$E[q_i^{2p}] = E[q_i^{2p-2(1+\kappa)} q_i^{2(1+\kappa)}] \leq m^{\frac{p-1+\kappa}{1+\kappa}} E[q_i^{2(1+\kappa)}] \leq m^{\frac{p-1+\kappa}{1+\kappa}} \phi.$$

Thus, for any $p \geq 2$,

$$E[|q_i^2 - E[q_i^2]|^p] \leq E[q_i^{2p}] + (E[q_i^2])^p \leq m^{\frac{p-1+\kappa}{1+\kappa}} \phi + \phi^{\frac{p}{1+}\kappa} \leq (m + \phi)^{\frac{p}{1+}\kappa} \phi(m + \phi)^{\frac{p}{1+}\kappa}.$$
By Bernstein’s inequality (Lemma 3.5.4), with probability at least $1 - 2e^{-\beta}$,

$$
\left| \frac{1}{N} \sum_{i=1}^{N} q_i^2 - \mathbb{E}[q_i^2] \right| \leq \left( \frac{\sqrt{2\beta} (N + \phi) 2^{1+\kappa} \phi^{1/2}}{N^{1/2}} + \frac{\beta (N + \phi)^{1+\kappa}}{n} \right)
$$

\[
\leq \frac{\sqrt{2\beta} (1 + \phi) 2^{1+\kappa} \phi^{1/2} + \beta (1 + \phi)^{1+\kappa}}{N^{1+\kappa}},
\]

which implies the first claim. To establish the second claim, note that for any $p \geq 2$,

$$
\mathbb{E} \left| q_i^{2(1+\kappa)} - \mathbb{E}[q_i^{2(1+\kappa)}] \right|^p \leq C(p) \left( \mathbb{E} \left| q_i^{2(1+\kappa)} \right|^p + \left( \mathbb{E} \left| q_i^{2(1+\kappa)} \right| \right)^p \right)
$$

\[
\leq C(p) \left( \mathbb{E} \left| q_i^{2(1+\kappa)(p-1)} q_i^{2(1+\kappa)} \right| + \phi^p \right)
\]

\[
\leq C(p)(N^{p-1} \phi + \phi^p) \leq C(p)(N + \phi)^{p-2}(N + \phi),
\]

where we used the fact that $|q_i| \leq N^{1/2(1+\kappa)}$ to obtain the third inequality. Bernstein’s inequality implies that with probability at least $1 - 2e^{-\beta}$,

$$
\left| \frac{1}{N} \sum_{i=1}^{N} q_i^{2(1+\kappa)} - \mathbb{E}[q_i^{2(1+\kappa)}] \right| \leq \sqrt{2\beta} (1 + \phi) \phi^{1/2} + (1 + \phi),
$$

which yields the second part of the claim. \[ \square \]

**Proof of inequality (3.31) for the index set $I_2$**. Combining Lemmas 3.5.10 and 3.5.11 with the inequality (3.34), and setting $u = 2^{l/2} \sqrt{\beta}$, we get that with probability at least $1 - 4 \exp(-2^l \beta)$, for all $l \in I_2$,

$$
|Z(\pi_l(t)) - Z(\pi_{l-1}(t))| \leq
$$

\[
C \|\pi_l(t) - \pi_{l-1}(t)\|_2 \frac{2^{l/2} \sqrt{\beta}}{N} \left( \left( \sum_{i=1}^{N} q_i^2 \right)^{1/2} + \frac{N 2^{l+\kappa} \left( \sum_{i=1}^{N} q_i^{2(1+\kappa)} \right)^{1/2}}{\phi^{1/2}} \right),
\]

for some constant $C = C(\kappa, \phi) > 0$; note that the factor $1/m$ appears due to equality (3.33). Next, we apply a chaining argument similar to the one used in Section 3.5.5 we obtain that with
probability at least $1 - ce^{-\beta/2}$,

$$\sup_{t \in T} \left| \sum_{l \in I_2} (Z(\pi_l(t)) - Z(\pi_{l-1}(t))) \right| \leq C \frac{\gamma_2(T)\sqrt{\beta}}{N} \left( \left( \sum_{i=1}^{N} \bar{q}_{i}^2 \right)^{1/2} + N^{\frac{\kappa}{2(1+\kappa)}} \left( \sum_{i=1}^{N} \bar{q}_{i}^2 \right)^{\frac{1}{4(1+\kappa)}} \right),$$

(3.38)

for a positive constant $C = C(\kappa, \phi)$ and an absolute constant $c > 0$. In order to handle the remaining terms involving $\bar{q}_{i}$ in (3.38), we apply Lemma 3.5.12, which gives

$$\sup_{t \in T} \left| \sum_{l \in I_2} (Z(\pi_l(t)) - Z(\pi_{l-1}(t))) \right| \leq C \frac{\gamma_2(T)\beta}{\sqrt{N}},$$

with probability at least $1 - ce^{-\beta/2}$, where $C = C(\kappa, \phi)$ and $c > 0$ are positive constants and $\beta \geq 8$. This completes the second part of the chaining argument. \qed

The case $l \in I_3$.

Proof of inequality (3.31) for the index set $I_3$. Direct application of Cauchy-Schwartz on (3.33) yields, for all $t \in T$,

$$|Z(\pi_l(t)) - Z(\pi_{l-1}(t))| \leq \left( \frac{1}{N} \sum_{i=1}^{N} w_i^2 \right)^{1/2} \left( \frac{1}{N} \sum_{i=1}^{N} \bar{q}_{i}^2 \right)^{1/2},$$

where $w_i = \langle U_i, \pi_l(t) - \pi_{l-1}(t) \rangle$ are sub-Gaussian random variables. Thus, by Lemma 3.6.2 $\omega_i^2$ are sub-exponential with norm bounded as in (3.35). Using Bernstein’s inequality again, we deduce that

$$\mathbb{P} \left( \left| \frac{1}{N} \sum_{i=1}^{N} (w_i^2 - \mathbb{E}[w_i^2]) \right| \geq 2\|U\|_{\psi_2}^2 \|\pi_l(t) - \pi_{l-1}(t)\|_2^2 \left( \sqrt{\frac{2u}{N}} + \frac{u}{N} \right) \right) \leq 2 \exp(-u).$$

Let $u = 2^l \beta$. Using the fact that $2^l \beta/N \geq 1$ as well as $\mathbb{E}[w_i^2] = \|\pi_l(t) - \pi_{l-1}(t)\|_2^2$, we see that the term $u/m$ dominates the right hand side and

$$\mathbb{P} \left( \left( \frac{1}{N} \sum_{i=1}^{N} w_i^2 \right)^{1/2} \geq C \|\pi_l(t) - \pi_{l-1}(t)\|_2 \frac{2^{l/2} \sqrt{\beta}}{\sqrt{N}} \right) \leq 2 \exp(-2^l \beta),$$

108
for some absolute constant $C > 0$. Thus, repeating a chaining argument of section 3.5.5 (namely, the argument following (3.32)), we obtain

$$\sup_{t \in T} \left| \sum_{l \in I_3} (Z(\pi_l(t)) - Z(\pi_{l-1}(t))) \right| \leq C \gamma_2(T) \beta \sqrt{\frac{N}{T}} \sqrt{\frac{1}{N} \sum_{i=1}^{N} q_i^2}$$

with probability at least $1 - ce^{-\beta/2}$ for some absolute constants $C, c > 0$. Combining this inequality with the first claim of Lemma 3.5.12 gives

$$\sup_{t \in T} \left| \sum_{l \in I_3} (Z(\pi_l(t)) - Z(\pi_{l-1}(t))) \right| \leq C \gamma_2(T) \beta \sqrt{\frac{N}{T}}$$

with probability at least $1 - ce^{-\beta/2}$ for absolute constants $C, c > 0$ and any $\beta \geq 8$. This finishes the bound for the third (and final) segment of the “chain”.

\[\square\]

**Finishing the proof of Lemma 3.5.8**

*Proof.* So far, we have shown that

$$\sup_{t \in T} |Z(t) - Z(t_0)| = \sup_{t \in T} \left| \sum_{l \geq 1} (Z(\pi_l(t)) - Z(\pi_{l-1}(t))) \right|$$

$$\leq \sum_{j \in \{1,2,3\}} \sup_{t \in T} \left| \sum_{l \in I_j} (Z(\pi_l(t)) - Z(\pi_{l-1}(t))) \right|$$

$$\leq C \gamma_2(T) \beta \sqrt{\frac{N}{T}},$$

with probability at least $1 - ce^{-\beta/2}$ for some positive constants $C = C(\kappa, \phi)$ and $c$, and any $\beta \geq 8$. To finish the proof, it remains to bound $|Z(t_0)| = \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i \bar{q}_i \langle U_i, t_0 \rangle \right|$. With $\Delta_d(T)$ defined in (3.28), and since $t_0$ is an arbitrary point in $T$, we trivially have $\|t_0\|_2 \leq \Delta_d(T)$. Applying Bernstein’s inequality in a way similar to Section 3.5.5 yields

$$\mathbb{P}\left( \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i \bar{q}_i \langle U_i, t_0 \rangle \right| \geq \left( \frac{C' \sqrt{2u}}{\sqrt{N}} + \frac{C'' u}{N^{1-\frac{1}{\kappa+\gamma}}} \right) \Delta_d(T) \right) \leq 2e^{-u},$$

109
for some constants $C' = C'(\kappa, \phi)$, $C'' = C''(\kappa, \phi) > 0$ and any $u > 0$. Choosing $u = \beta$ gives

$$
\mathbb{P} \left( \left| \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i q_i \langle \bar{U}_i, t_0 \rangle \right| \geq \frac{C \Delta_d(T) \beta}{\sqrt{N}} \right) \leq 2e^{-\beta},
$$

for a constant $C = C(\kappa, \phi) > 0$ and any $\beta \geq 0$. Combining this bound with (3.39) shows that with probability at least $1 - ce^{-\beta/2}$,

$$
\sup_{t \in T} \left| \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i (\bar{U}_i, t) \bar{q}_i \right| \leq C \left( \gamma_2(T) + \Delta_d(T) \right) \beta \frac{\sqrt{N}}{\beta} \leq C \left( L \omega(T) + \Delta_d(T) \right) \beta \frac{\sqrt{N}}{\beta},
$$

for $C = C(\kappa, \phi)$, an absolute constant $L > 0$ and all $\beta \geq \beta_0$; note that the last inequality follows from Lemma 3.5.2. We have established (3.29), thus completing the proof. \qed

### 3.6 Technical Results.

**Lemma 3.6.1.** For any nonnegative random variable $X$, if $\mathbb{P} (X > K \beta) \leq ce^{-\beta/2}$ for some constants $K, c > 0$ and all $\beta \geq \beta_0 \geq 0$, then,

$$
\mathbb{E}[X] \leq K \left( \beta_0 + 2ce^{-\beta_0/2} \right).
$$

**Proof.** Using a well known identity for the expectation of non-negative random variables,

$$
\mathbb{E}[X] = \int_{0}^{\infty} \mathbb{P} (X > u) du = K \int_{0}^{\infty} \mathbb{P} (X > K \beta) d\beta \leq K \left( \beta_0 + \int_{\beta_0}^{\infty} \mathbb{P} (X > K \beta) d\beta \right) \leq K \left( \beta_0 + \int_{\beta_0}^{\infty} ce^{-\beta/2} d\beta \right) = K \left( \beta_0 + 2ce^{-\beta_0/2} \right).
$$

\qed

**Lemma 3.6.2.** If $X$ and $Y$ are sub-Gaussian random variables, then the product $XY$ is a subexponential random variable, and

$$
\|XY\|_{\psi_1} \leq \|X\|_{\psi_2} \|Y\|_{\psi_2}.
$$

**Proof.** See [vdVW96b].
Lemma 3.6.3. Let \( k = \lfloor 2^l \beta / \log (eN / 2^l \beta) \rfloor \) and \( l \in I_2 \), then, \( (\frac{eN}{k})^k \leq \exp(3 \cdot 2^l \beta) \).

Proof. If \( k \geq 2 \), then, \( 2^l \beta / \log (eN / 2^l \beta) \geq 2 \), which implies \( 2^l \beta \geq 2 \log (eN / 2^l \beta) \). Thus,

\[
\left( \frac{eN}{k} \right)^k \leq 2 \exp \left( \frac{2^l \beta}{\log \frac{eN}{2^l \beta}} \log \left( \frac{eN}{\frac{2^l \beta}{\log \frac{eN}{2^l \beta} - 1}} \right) \right)
\]

\[
\leq 2 \exp \left( \frac{2^l \beta}{\log \frac{eN}{2^l \beta}} \log \left( \frac{eN}{2^l \beta - \log \frac{eN}{2^l \beta}} \right) \right)
\]

\[
\leq 2 \exp \left( \frac{2^l \beta}{\log \frac{eN}{2^l \beta}} \log \left( \frac{2eN}{2^l \beta} \log \frac{eN}{2^l \beta} \right) \right) \leq \exp(3 \cdot 2^l \beta),
\]

where the second from last inequality follows from \( (\frac{eN}{k})^k \leq \exp(3 \cdot 2^l \beta) \), and the last inequality follows from \( N \geq 2^l \beta \), thus, \( \log (eN / 2^l \beta) \leq 2 \).

On the other hand, if \( k = 1 \), then, since \( \log eN \leq 2^l \beta \), \( (\frac{eN}{k})^k = eN = \exp(\log eN) \leq \exp(2^l \beta) \), finishing the proof. \( \square \)

Lemma 3.6.4. With \( N \geq 1, \beta \geq 1, \kappa \in (1, 0) \) and \( l \in I_2 = \{ l \geq 1 : \log eN \leq 2^l \beta < N \} \), the integer \( k = \lfloor 2^l \beta / \log (eN / 2^l \beta) \rfloor \) satisfies \( k \geq 1 \), and

\[
\left( \frac{2 + \frac{4}{\kappa}}{e^{1/(1+\kappa)}} \right)^{\frac{2+\kappa}{2+\kappa}} N^{\frac{\kappa}{2(1+\kappa)}} k^{\frac{\kappa}{2(1+\kappa)}} \geq 2^l \beta.
\]

Proof. Since \( 2^l \beta \geq \log (eN) \geq 1 \), it follows that \( k \geq 1 \), and thus \( k \geq 2^l \beta / 2 \log (eN / 2^l \beta) \). It is then enough to show that

\[
\left( \frac{1 + \frac{2}{\kappa}}{e^{1/(1+\kappa)}} \right)^{\frac{2+\kappa}{2+\kappa}} \left( \frac{N}{2^l \beta} \right)^{\frac{\kappa}{2(1+\kappa)}} \geq \left( \log \frac{eN}{2^l \beta} \right)^{\frac{2+\kappa}{2(1+\kappa)}}.
\]

Raising both sides to the power of \( 2(1+\kappa)/\kappa \), equivalently

\[
\left( 1 + \frac{2}{\kappa} \right)^{\frac{2+\kappa}{\kappa}} / e^{\frac{2}{\kappa}} \geq \left( \log \frac{eN}{2^l \beta} \right)^{\frac{2+\kappa}{\kappa}} / \left( \frac{N}{2^l \beta} \right).
\]

Consider the function \( g(x) = (\log ex)^{\frac{2+\kappa}{\kappa}} / x \). Note that as \( m > 2^l \beta \), to prove the inequality above it suffices to show that the \( \sup_{x \geq 1} g(x) \) is upper bounded by the left hand side. Taking
the derivative of $g(x)$ yields

$$g'(x) = \frac{2+\kappa}{\kappa}(1 + \log x)^{2/\kappa} - (1 + \log x)^{(2+\kappa)/\kappa}.$$ 

Since $x \geq 1$, the only critical point at which the global maximum occurs is given by $x = e^{2/\kappa}$. As $g(e^{2/\kappa})$ is exactly equal to the left hand side the proof is complete.

### 3.7 Decomposable Norms and Restricted Compatibility.

In this section, we recall some facts about decomposable norms that have been introduced in [NRWY12].

**Definition 3.7.1.** Suppose that $\mathcal{L} \subseteq \mathcal{L}_1$ are two subspace of $\mathbb{R}^d$, and let $\mathcal{L}_1^\perp$ be the orthogonal complement of $\mathcal{L}_1$. Norm $\| \cdot \|_\mathcal{K}$ is said to be decomposable with respect to $(\mathcal{L}, \mathcal{L}_1^\perp)$ if for any $\theta \in \mathbb{R}^d$,

$$\|\theta_1 + \theta_2\|_\mathcal{K} = \|\Pi_{\mathcal{L}}\theta_1\|_\mathcal{K} + \|\Pi_{\mathcal{L}_1^\perp}\theta\|_\mathcal{K},$$

where $\Pi_\mathcal{L}$ and $\Pi_{\mathcal{L}_1^\perp}$ stand for the orthogonal projectors onto $\mathcal{L}$ and $\mathcal{L}_1^\perp$ respectively.

It is well known that many frequently used norms, including the $\ell_1$ norm of a vector and the nuclear norm of a matrix, are decomposable with respect to the appropriately chosen pair of subspaces. For instance, the $\ell_1$ norm is decomposable with respect to the pair of subspaces $(\mathcal{L}(J), \mathcal{L}(J)^\perp)$, where

$$\mathcal{L}(J) := \{ v \in \mathbb{R}^d : v_j = 0 \text{ for all } j \notin J \} \quad (3.40)$$

consists of sparse vectors with non-zero coordinates indexed by a set $J \subseteq \{1, \ldots, d\}$.

Let $W_1 \subseteq \mathbb{R}^{d_1}$, $W_2 \subseteq \mathbb{R}^{d_2}$ be two linear subspaces. Then we define the subspace $\mathcal{L}(W_1, W_2) \subseteq \mathbb{R}^{d_1 \times d_2}$ via

$$\mathcal{L}(W_1, W_2) := \{ M \in \mathbb{R}^{d_1 \times d_2} : \text{row}(M) \subseteq W_1, \text{col}(M) \subseteq W_2 \},$$

where row$(M)$ and col$(M)$ are the linear subspaces spanned by the rows and columns of $M$. 

112
respectively, and

\[
\mathcal{L}_1^+ (W_1, W_2) := \{ M \in \mathbb{R}^{d_1 \times d_2} : \text{row}(M) \subseteq W_1^\perp, \text{col}(M) \subseteq W_2^\perp \}. \tag{3.41}
\]

Then the nuclear norm \( \| \cdot \|_* \) is decomposable with respect to \((\mathcal{L}(W_1, W_2), \mathcal{L}_1^+ (W_1, W_2))\) (see \cite{NRWY12} for details).

Assume that the norm \( \| \cdot \|_\mathcal{K} \) is decomposable with respect to \((\mathcal{L}, \mathcal{L}_1^+)\), and let \( \theta \in \mathcal{L} \). It is clear that for any \( v \in S_{c_0}(\theta) \)

\[
\| \theta + v \|_\mathcal{K} = \| \Pi_\mathcal{L} \theta + \Pi_\mathcal{L}_1 v + \Pi_\mathcal{L}_1^+ v \|_\mathcal{K} \leq \| \Pi_\mathcal{L} \theta \|_\mathcal{K} + \frac{1}{c_0} \| \Pi_\mathcal{L}_1 v \|_\mathcal{K} + \| \Pi_\mathcal{L}_1^+ v \|_\mathcal{K}. \tag{3.42}
\]

Since \( \theta \in \mathcal{L} \), decomposability and the triangle inequality imply that

\[
\| \Pi_\mathcal{L} \theta + \Pi_\mathcal{L}_1 v + \Pi_\mathcal{L}_1^+ v \|_\mathcal{K} = \| \Pi_\mathcal{L} \theta + \Pi_\mathcal{L}_1 v \|_\mathcal{K} + \| \Pi_\mathcal{L}_1^+ v \|_\mathcal{K}
\geq \| \Pi_\mathcal{L} \theta \|_\mathcal{K} - \| \Pi_\mathcal{L}_1 v \|_\mathcal{K} + \| \Pi_\mathcal{L}_1^+ v \|_\mathcal{K}.
\]

Substituting this bound into (3.42) gives

\[
-\| \Pi_\mathcal{L} v \|_\mathcal{K} + \| \Pi_\mathcal{L}_1^+ v \|_\mathcal{K} \leq \frac{1}{c_0} \| \Pi_\mathcal{L}_1 v \|_\mathcal{K} + \frac{1}{c_0} \| \Pi_\mathcal{L}_1^+ v \|_\mathcal{K},
\]

which implies that for any \( v \in S_{c_0}(\theta) \)

\[
\| \Pi_\mathcal{L}_1^+ v \|_\mathcal{K} \leq \frac{c_0 + 1}{c_0 - 1} \| \Pi_\mathcal{L}_1 v \|_\mathcal{K}.
\]

It is easy to see that the set of all \( v \) satisfying the inequality above is a convex cone, which we will denote by \( C_{c_0} = C_{c_0}(\mathcal{K}) \). Since \( S_{c_0}(\theta) \subseteq C_{c_0} \),

\[
\Psi (S_{c_0}(\theta)) \leq \Psi (C_{c_0})
\]

by definition of the restricted compatibility constant. This inequality is useful due to the fact that it is often easier to estimate \( \Psi (C_{c_0}) \).

Finally, we make a remark that is useful when dealing with non-isotropic measurements. Let \( \Sigma \succ 0 \) be a \( d \times d \) matrix, and consider the norm corresponding to the convex set \( \Sigma^{1/2} \mathcal{K} \), so that
\[ \|v\|_{\Sigma^{1/2}} = \|\Sigma^{-1/2}v\|_{\mathcal{K}}. \] It is easy to see that \( C_{c_0}(\Sigma^{1/2}) = \Sigma^{1/2}C_{c_0}(\mathcal{K}) \), hence
\[
\Psi \left( C_{c_0}(\Sigma^{1/2}); \Sigma^{1/2} \right) = \sup_{v \in \Sigma^{1/2}\mathcal{K}\{0\}} \frac{\|v\|_{\Sigma^{1/2}}}{\|v\|_2} = \sup_{u \in \mathcal{K}\{0\}} \frac{\|u\|_{\mathcal{K}}}{\|\Sigma^{1/2}u\|_2} \leq \|\Sigma^{-1/2}\| \Psi \left( C_{c_0}(\mathcal{K}); \mathcal{K} \right).
\]

**Example 1:** \( \ell_1 \) norm. Let \( \mathcal{L}(J) \) be as in (3.40) with \( |J| = s \leq d \). If \( v \in \mathbb{R}^d \) belongs to the corresponding cone \( C(c_0) \), then clearly \( \|v\|_1 \leq \frac{2c_0}{c_0-1} \|v_J\|_1 \), where \( v_J := \Pi_{\mathcal{L}(J)}v \). Hence
\[
\|v\|_1 \leq \frac{2c_0}{c_0-1} \|v_J\|_1 \leq \frac{2c_0}{c_0-1} \sqrt{|J|} \|v\|_2,
\]
and \( \Psi(C_{c_0}) \leq \frac{2c_0}{c_0-1} \sqrt{s} \).

**Example 2:** nuclear norm. Let \( \mathcal{L}_1^+(W_1, W_2) \) be as in (3.41). Note that for any \( v \in \mathbb{R}^{d_1 \times d_2} \),\( \Pi_{\mathcal{L}_1^+(W_1, W_2)}v = \Pi_{W_2^\perp}v\Pi_{W_1^\perp} \), where \( \Pi_{W_1^\perp} \) and \( \Pi_{W_2^\perp} \) are the orthogonal projectors onto subspaces \( W_1 \subseteq \mathbb{R}^{d_1} \) and \( W_2 \subseteq \mathbb{R}^{d_2} \) respectively. Then for any \( v \in C_{c_0} \), we have that
\[
\|v\|_* \leq \|\Pi_{\mathcal{L}_1^+(W_1, W_2)}v\|_* + \|\Pi_{\mathcal{L}_1(W_1, W_2)}v\|_* \leq \frac{2c_0}{c_0-1} \|\Pi_{\mathcal{L}_1(W_1, W_2)}v\|_*. \tag{3.43}
\]
Note that
\[
\Pi_{\mathcal{L}_1(W_1, W_2)}v = v - \Pi_{W_2^\perp}v\Pi_{W_1^\perp} = \Pi_{W_2^\perp}v\Pi_{W_1} + \Pi_{W_2}v,
\]
hence \( \text{rank} \left( \Pi_{\mathcal{L}_1(W_1, W_2)}v \right) \leq 2 \max(\dim(W_1), \dim(W_2)) \), which yields together with (3.43) that
\[
\|v\|_* \leq \frac{2c_0}{c_0-1} \|\Pi_{\mathcal{L}_1(W_1, W_2)}v\|_* \leq \frac{2c_0}{c_0-1} \sqrt{2\max(\dim(W_1), \dim(W_2))} \|v\|_2,
\]
and \( \Psi(C_{c_0}) \leq \frac{2\sqrt{2c_0}}{c_0-1} \sqrt{\max(\dim(W_1), \dim(W_2))} \).
Chapter 4

Estimation of the Covariance Structure of Heavy-tailed Distributions

In this chapter, we propose and analyze a new estimator of the covariance matrix that admits strong theoretical guarantees under weak assumptions on the underlying distribution, such as existence of moments of only low order. While estimation of covariance matrices corresponding to sub-Gaussian distributions is well-understood, much less is known in the case of heavy-tailed data. As K. Balasubramanian and M. Yuan write [BY16], “data from real-world experiments oftentimes tend to be corrupted with outliers and/or exhibit heavy tails. In such cases, it is not clear that those covariance matrix estimators .. remain optimal” and “..what are the other possible strategies to deal with heavy tailed distributions warrant further studies.” We make a step towards answering this question and prove tight deviation inequalities for the proposed estimator that depend only on the parameters controlling the “intrinsic dimension” associated to the covariance matrix (as opposed to the dimension of the ambient space); in particular, our results are applicable in the case of high-dimensional observations.

4.1 Introduction

Estimation of the covariance matrix is one of the fundamental problems in data analysis: many important statistical tools, such as Principal Component Analysis(PCA) [Hot33] and regression analysis, involve covariance estimation as a crucial step. For instance, PCA has immediate applications to nonlinear dimension reduction and manifold learning techniques [ACMT12], genetics [NJB+08], computational biology [ABB00], among many others.
However, assumptions underlying the theoretical analysis of most existing estimators, such as various modifications of the sample covariance matrix, are often restrictive and do not hold for real-world scenarios. Usually, such estimators rely on heuristic (and often bias-producing) data preprocessing, such as outlier removal. To eliminate such preprocessing step from the equation, one has to develop a class of new statistical estimators that admit strong performance guarantees, such as exponentially tight concentration around the unknown parameter of interest, under weak assumptions on the underlying distribution, such as existence of moments of only low order. In particular, such heavy-tailed distributions serve as a viable model for data corrupted with outliers – an almost inevitable scenario for applications.

We make a step towards solving this problem: using tools from the random matrix theory, we will develop a class of robust estimators that are numerically tractable and are supported by strong theoretical evidence under much weaker conditions than currently available analogues. The term “robustness” refers to the fact that our estimators admit provably good performance even when the underlying distribution is heavy-tailed.

4.1.1 Notation

Given $A \in \mathbb{R}^{d_1 \times d_2}$, let $A^T \in \mathbb{R}^{d_2 \times d_1}$ be transpose of $A$. If $A$ is symmetric, we will write $\lambda_{\text{max}}(A)$ and $\lambda_{\text{min}}(A)$ for the largest and smallest eigenvalues of $A$. Next, we will introduce the matrix norms used in the chapter. Everywhere below, $\| \cdot \|$ stands for the operator norm $\|A\| := \sqrt{\lambda_{\text{max}}(A^T A)}$. If $d_1 = d_2 = d$, we denote by $\text{tr} A$ the trace of $A$. For $A \in \mathbb{R}^{d_1 \times d_2}$, the nuclear norm $\| \cdot \|_1$ is defined as $\|A\|_1 = \text{tr}(\sqrt{A^T A})$, where $\sqrt{A^T A}$ is a nonnegative definite matrix such that $(\sqrt{A^T A})^2 = A^T A$. The Frobenius (or Hilbert-Schmidt) norm is $\|A\|_F = \sqrt{\text{tr}(A^T A)}$, and the associated inner product is $\langle A_1, A_2 \rangle = \text{tr}(A_1^* A_2)$. For $z \in \mathbb{R}^d$, $\|z\|_2$ stands for the usual Euclidean norm of $z$. Let $A, B$ be two self-adjoint matrices. We will write $A \succeq B$ (or $A \succ B$) iff $A - B$ is nonnegative (or positive) definite. For $a, b \in \mathbb{R}$, we set $a \vee b := \max(a, b)$ and $a \wedge b := \min(a, b)$. We will also use the standard Big-O and little-o notation when necessary.

Finally, we give a definition of a matrix function. Let $f$ be a real-valued function defined on an interval $T \subseteq \mathbb{R}$, and let $A \in \mathbb{R}^{d \times d}$ be a symmetric matrix with the eigenvalue decomposition
A = UΛU* such that λ_j(A) ∈ T, j = 1, ..., d. We define f(A) as f(A) = Uf(Λ)U*, where

\[
f(Λ) = \begin{pmatrix} f(λ_1) \\ \vdots \\ f(λ_d) \end{pmatrix}.
\]

Few comments about organization of the material in the rest of the chapter: section 4.1.2 provides an overview of the related work. Section 4.2 contains the main results of the chapter. The proofs are outlined in section 4.4; longer technical arguments can be found in the supplementary material.

### 4.1.2 Problem formulation and overview of the existing work

Let \( X ∈ \mathbb{R}^d \) be a random vector with mean \( E X = μ_0 \), covariance matrix \( Σ_0 = E [(X − μ_0)(X − μ_0)^T] \), and assume \( E \|X − μ_0\|^4_2 < ∞ \). Let \( X_1, \ldots, X_m \) be i.i.d. copies of \( X \). Our goal is to estimate the covariance matrix \( Σ \) from \( X_j, j ≤ m \). This problem and its variations have previously received significant attention by the research community: excellent expository chapters by [CRZ16] and [FLL16] discuss the topic in detail. However, strong guarantees for the best known estimators hold (with few exceptions mentioned below) under the restrictive assumption that \( X \) is either bounded with probability 1 or has sub-Gaussian distribution, meaning that there exists \( σ > 0 \) such that for any \( v ∈ \mathbb{R}^d \) of unit Euclidean norm,

\[
Pr (|⟨v, X − μ_0⟩| ≥ t) ≤ 2e^{−\frac{t^2}{2σ^2}}.
\]

In the discussion accompanying the chapter by [CRZ16], [BY16] write that “data from real-world experiments oftentimes tend to be corrupted with outliers and/or exhibit heavy tails. In such cases, it is not clear that those covariance matrix estimators described in this article remain optimal” and “what are the other possible strategies to deal with heavy-tailed distributions warrant further studies.” This motivates our main goal: develop new estimators of the covariance matrix that (i) are computationally tractable and perform well when applied to heavy-tailed data and (ii) admit strong theoretical guarantees (such as exponentially tight concentration around the unknown covariance matrix) under weak assumptions on the underlying distribution. Note that, unlike the majority of existing literature, we do not impose any further conditions on the
moments of $X$, or on the “shape” of its distribution, such as elliptical symmetry.

Robust estimators of covariance and scatter have been studied extensively during the past few decades. However, majority of rigorous theoretical results were obtained for the class of elliptically symmetric distributions which is a natural generalization of the Gaussian distribution; we mention just a small subsample among the thousands of published works. Notable examples include the Minimum Covariance Determinant estimator and the Minimum Volume Ellipsoid estimator which are discussed in [HRVA08], as well Tyler’s [Tyl87] M-estimator of scatter. Works by [PLL16, WZ+16, HL16] exploit the connection between Kendall’s tau and Pearson’s correlation coefficient [FKN90] in the context of elliptical distributions to obtain robust estimators of correlation matrices. Interesting results for shrinkage-type estimators have been obtained by [LW04, LW+12]. In a recent work, [CGR15] study Huber’s $\varepsilon$-contamination model which assumes that the data is generated from the distribution of the form $(1 - \varepsilon)F + \varepsilon Q$, where $Q$ is an arbitrary distribution of “outliers” and $F$ is an elliptical distribution of “inliers”, and propose novel estimator based on the notion of “matrix depth” which is related to Tukey’s depth function [Tuk75]; a related class of problems has been studies by [DKK+16]. The main difference of the approach investigated in this chapter is the ability to handle a much wider class of distributions that are not elliptically symmetric and only satisfy weak moment assumptions. Recent papers by [Cat16, Gin15, FWZ16a, FLW17, FK17] and [Min16] are closest in spirit to this direction. For instance, [Cat16] constructs a robust estimator of the Gram matrix of a random vector $Z \in \mathbb{R}^d$ (as well as its covariance matrix) via estimating the quadratic form $\mathbb{E} \langle Z, u \rangle^2$ uniformly over all $\|u\|_2 = 1$. However, the bounds are obtained under conditions more stringent than those required by our framework, and resulting estimators are difficult to evaluate in applications even for data of moderate dimension. [FWZ16a] obtain bounds in norms other than the operator norm which the focus of the present chapter. [Min16] and [FWZ16c] use adaptive truncation arguments to construct robust estimators of the covariance matrix. However, their results are only applicable to the situation when the data is centered (that is, $\mu_0 = 0$). In the robust estimation framework, rigorous extension of the arguments to the case of non-centered high-dimensional observations is non-trivial and requires new tools, especially if one wants to avoid statistically inefficient procedures such as sample splitting. We formulate and prove such extensions in this chapter.
4.2 Main Results

Definition of our estimator has its roots in the technique proposed by [Cat12]. Let

\[ \psi(x) = (|x| \wedge 1) \text{sign}(x) \]  

(4.1)

be the usual truncation function. As before, let \( X_1, \ldots, X_m \) be i.i.d. copies of \( X \), and assume that \( \hat{\mu} \) is a suitable estimator of the mean \( \mu_0 \) from these samples, to be specified later. We define \( \hat{\Sigma} \) as

\[ \hat{\Sigma} := \frac{1}{m \theta} \sum_{i=1}^{m} \psi(\theta(X_i - \hat{\mu})(X_i - \hat{\mu})^T), \]  

(4.2)

where \( \theta \simeq m^{-1/2} \) is small (the exact value will be given later). It easily follows from the definition of the matrix function that

\[ \hat{\Sigma} = \frac{1}{m \theta} \sum_{i=1}^{m} \frac{(X_i - \hat{\mu})(X_i - \hat{\mu})^T}{\|X_i - \hat{\mu}\|_2^2} \psi(\theta \|X_i - \hat{\mu}\|_2^2), \]

hence it is easily computable. Note that \( \psi(x) = x \) in the neighborhood of 0; it implies that whenever all random variables \( \theta \|X_i - \hat{\mu}\|_2^2, \ 1 \leq i \leq m \) are “small” (say, bounded above by 1) and \( \hat{\mu} \) is the sample mean, \( \hat{\Sigma} \) is close to the usual sample covariance estimator. On the other hand, \( \psi \) “truncates” \( \|X_i - \hat{\mu}\|_2^2 \) on level \( \simeq \sqrt{m} \), thus limiting the effect of outliers. Our results (formally stated below, see Theorem 4.2.1) imply that for an appropriate choice of \( \theta = \theta(t, m, \sigma) \),

\[ \|\hat{\Sigma} - \Sigma_0\| \leq C_0 \sigma_0 \beta \frac{\sqrt{\beta}}{m} \]

with probability \( \geq 1 - de^{-\beta} \) for some positive constant \( C_0 \), where

\[ \sigma_0^2 := \left\| \mathbb{E} \|X - \mu_0\|_2^2 (X - \mu_0)(X - \mu_0)^T \right\| \]

is the "matrix variance".
4.2.1 Robust mean estimation

There are several ways to construct a suitable estimator of the mean $\mu_0$. We present the one obtained via the “median-of-means” approach. Let $x_1, \ldots, x_k \in \mathbb{R}^d$. Recall that the geometric median of $x_1, \ldots, x_k$ is defined as

$$\text{Med}x_1, \ldots, x_k := \arg\min_{z \in \mathbb{R}^d} \sum_{j=1}^{k} \| z - x_j \|_2.$$  

Let $1 < \beta < \infty$ be the confidence parameter, and set $k = \left\lceil 3.5\beta \right\rceil + 1$; we will assume that $k \leq \frac{m}{2}$. Divide the sample $X_1, \ldots, X_m$ into $k$ disjoint groups $G_1, \ldots, G_k$ of size $\left\lfloor \frac{m}{k} \right\rfloor$ each, and define

$$\hat{\mu}_j := \frac{1}{|G_j|} \sum_{i \in G_j} X_i, \ j = 1 \ldots k,$$

$$\hat{\mu} := \text{Med}\hat{\mu}_1, \ldots, \hat{\mu}_k. \quad (4.3)$$

It then follows from Corollary 4.1 in [Min15] that

$$\Pr\left( \| \hat{\mu} - \mu \|_2 \geq 11 \sqrt{\frac{\text{tr}(\Sigma_0)(\beta + 1)}{m}} \right) \leq e^{-\beta}. \quad (4.4)$$

4.2.2 Robust covariance estimation

Let $\hat{\Sigma}$ be the estimator defined in (4.2) with $\hat{\mu}$ being the “median-of-means” estimator (4.3). Then $\hat{\Sigma}$ admits the following performance guarantees:

**Lemma 4.2.1.** Assume that $\sigma \geq \sigma_0$, and set $\theta = \frac{1}{2} \sqrt{\frac{\sigma}{m}}$. Moreover, let $\tilde{d} := \sigma_0^2 / \| \Sigma_0 \|^2$, and suppose that $m \geq C\tilde{d}\beta$, where $C > 0$ is an absolute constant. Then

$$\| \hat{\Sigma} - \Sigma_0 \| \leq 3\sigma \sqrt{\frac{\beta}{m}} \quad (4.5)$$

with probability at least $1 - 5de^{-\beta}$.

**Remark 4.2.1.** The quantity $\tilde{d}$ is a measure of “intrinsic dimension” akin to the “effective rank” $r = \frac{\text{tr}(\Sigma_0)}{\| \Sigma_0 \|}$; see Lemma 4.2.3 below for more details. Moreover, note that the claim of Lemma 4.2.1 holds for any $\sigma \geq \sigma_0$, rather than just for $\sigma = \sigma_0$; this “degree of freedom” allows construction of adaptive estimators, as it is shown below.
The statement above suggests that one has to know the value of (or a tight upper bound on) the “matrix variance” $\sigma^2_0$ in order to obtain a good estimator $\widehat{\Sigma}$. More often than not, such information is unavailable. To make the estimator completely data-dependent, we will use Lepski’s method \cite{Lep92}. To this end, assume that $\sigma_{\min}$, $\sigma_{\max}$ are “crude” preliminary bounds such that

$$\sigma_{\min} \leq \sigma_0 \leq \sigma_{\max}. \tag{4.6}$$

Usually, $\sigma_{\min}$ and $\sigma_{\max}$ do not need to be precise, and can potentially differ from $\sigma_0$ by several orders of magnitude. Set

$$\sigma_j := \sigma_{\min} 2^j \text{ and } J = \{ j \in \mathbb{Z} : \sigma_{\min} \leq \sigma_j < 2\sigma_{\max} \}.$$ 

Note that the cardinality of $J$ satisfies $\text{Card}(J) \leq 1 + \log_2(\sigma_{\max}/\sigma_{\min})$. For each $j \in J$, define $\theta_j := \theta(j, \beta) = \frac{1}{\sigma_j} \sqrt{\beta/m}$. Define

$$\widehat{\Sigma}_{m,j} = \frac{1}{m\theta_j} \sum_{i=1}^{m} \psi(\theta_j(X_i - \widehat{\mu})(X_i - \widehat{\mu})^T). \tag{4.7}$$

Finally, set

$$j_* := \min \left\{ j \in J : \forall k > j \text{ s.t. } k \in J, \|\widehat{\Sigma}_{m,k} - \widehat{\Sigma}_{m,j}\| \leq 6\sigma_k \sqrt{\beta/m} \right\} \tag{4.6}$$

and $\widehat{\Sigma}_* := \widehat{\Sigma}_{m,j_*}$. Note that the estimator $\widehat{\Sigma}_*$ depends only on $X_1, \ldots, X_m$, as well as $\sigma_{\min}$, $\sigma_{\max}$.

Our main result is the following statement regarding the performance of the data-dependent estimator $\widehat{\Sigma}_*$:

**Theorem 4.2.1.** Suppose $m \geq Cd\beta$, then, the following inequality holds with probability at least $1 - 5d\log_2 \left( \frac{2\sigma_{\max}}{\sigma_{\min}} \right) e^{-\beta}$:

$$\|\widehat{\Sigma}_* - \Sigma_0\| \leq 18\sigma_0 \sqrt{\beta/m}. \tag{4.7}$$

An immediate corollary of Theorem 4.2.1 is the quantitative result for the performance of PCA based on the estimator $\widehat{\Sigma}_*$. Let $\text{Proj}_k$ be the orthogonal projector on a subspace corresponding to the $k$ largest positive eigenvalues $\lambda_1, \ldots, \lambda_k$ of $\Sigma_0$ (here, we assume for simplicity that all the eigenvalues are distinct), and $\text{Proj}_k$ – the orthogonal projector of the same rank as
Proj$_k$ corresponding to the $k$ largest eigenvalues of $\hat{\Sigma}_*$.

The following bound follows from the Davis-Kahan perturbation theorem [DK70], more specifically, its version due to [Zwa06].

**Corollary 4.2.1.** Let $\Delta_k = \lambda_k - \lambda_{k+1}$, and assume that $\Delta_k \geq 72\sigma_0\sqrt{\frac{2}{m}}$. Then

$$\|\hat{\text{Proj}}_k - \text{Proj}_k\| \leq \frac{36}{\Delta_k} \sigma_0 \sqrt{\frac{\beta}{m}}$$

with probability $\geq 1 - 5d\log_2\left(\frac{2\sigma_{\max}}{\sigma_{\min}}\right) e^{-\beta}$.

It is worth comparing the bound of Lemma 4.2.1 and Theorem 4.2.1 above to results of the paper by [FWZ16c], which constructs a covariance estimator $\hat{\Sigma}_m$ under the assumption that the random vector $X$ is centered, and $\sup_{v \in \mathbb{R}^d : \|v\|_2 \leq 1} \mathbb{E}[|\langle v, X \rangle|^4] = B < \infty$. More specifically, $\hat{\Sigma}_m$ satisfies the inequality

$$\mathbb{P}\left(\|\hat{\Sigma}_m - \Sigma_0\| \geq \sqrt{\frac{C_1Bd}{m}}\right) \leq de^{-\beta}, \quad (4.7)$$

where $C_1 > 0$ is an absolute constant. The main difference between (4.7) and the bounds of Lemma 4.2.1 and Theorem 4.2.1 is that the latter are expressed in terms of $\sigma_0^2$, while the former is in terms of $B$. The following lemma demonstrates that our bounds are at least as good:

**Lemma 4.2.2.** Suppose that $\mathbb{E}X = 0$ and $\sup_{v \in \mathbb{R}^d : \|v\|_2 \leq 1} \mathbb{E}[|\langle v, X \rangle|^4] = B < \infty$. Then $Bd \geq \sigma_0^2$.

It follows from the above lemma that $\bar{d} = \sigma_0^2/\|\Sigma_0\|^2 \lesssim d$. Hence, By Theorem 4.2.1, the error rate of estimator $\hat{\Sigma}_*$ is bounded above by $O(\sqrt{d/m})$ if $m \gtrsim d$. It has been shown (for example, see [Lou14]) that the minimax lower bound of covariance estimation is of order $\Omega(\sqrt{d/m})$. Hence, the bounds of [FWZ16c] as well as our results imply correct order of the error. That being said, the “intrinsic dimension” $\bar{d}$ reflects the structure of the covariance matrix and can potentially be much smaller than $d$, as it is shown in the next section.

### 4.2.3 Bounds in terms of intrinsic dimension

In this section, we show that under a slightly stronger assumption on the fourth moment of the random vector $X$, the bound $O(\sqrt{d/m})$ is suboptimal, while our estimator can achieve a much better rate in terms of the “intrinsic dimension” associated to the covariance matrix. This
makes our estimator useful in applications involving high-dimensional covariance estimation, such as PCA. Assume the following uniform bound on the kurtosis of linear forms $\langle Z, v \rangle$:

$$\sup_{\|v\|_2 \leq 1} \sqrt{\frac{\mathbb{E}(Z, v)^4}{\mathbb{E}(Z, v)^2}} = R < \infty. \quad (4.8)$$

The intrinsic dimension of the covariance matrix $\Sigma_0$ can be measured by the effective rank defined as

$$r(\Sigma_0) = \frac{\text{tr}(\Sigma_0)}{\|\Sigma_0\|}.$$ 

Note that we always have $r(\Sigma_0) \leq \text{rank}(\Sigma_0) \leq d$, and in some situations $r(\Sigma_0) \ll \text{rank}(\Sigma_0)$, for instance if the covariance matrix is “approximately low-rank”, meaning that it has many small eigenvalues. The constant $\sigma_0^2$ is closely related to the effective rank as is shown in the following lemma (the proof of which is included in the supplementary material):

**Lemma 4.2.3.** Suppose that (4.8) holds. Then,

$$r(\Sigma_0)\|\Sigma_0\|^2 \leq \sigma_0^2 \leq R^2 r(\Sigma_0)\|\Sigma_0\|^2.$$ 

As a result, we have $r(\Sigma_0) \leq \overline{d} \leq R^2 r(\Sigma_0)$. The following corollary immediately follows from Theorem 4.2.1 and Lemma 4.2.3.

**Corollary 4.2.2.** Suppose that $m \geq C\beta r(\Sigma_0)$ for an absolute constant $C > 0$ and that (4.8) holds. Then,

$$\|\hat{\Sigma}_n - \Sigma_0\| \leq 18R\|\Sigma_0\| \sqrt{\frac{r(\Sigma_0)\beta}{m}}$$

with probability at least $1 - 5d \log_2 \left( \frac{2\sigma_{\text{max}}}{\sigma_{\text{min}}} \right) e^{-\beta}$.

### 4.3 Applications: Low-rank Covariance Estimation

In many data sets encountered in modern applications (for instance, gene expression profiles [SJH⁺07]), dimension of the observations, hence the corresponding covariance matrix, is larger than the available sample size. However, it is often possible, and natural, to assume that the unknown matrix possesses special structure, such as low rank, thus reducing the “effective dimension” of the problem. The goal of this section is to present an estimator of the covariance
matrix that is “adaptive” to the possible low-rank structure; such estimators are well-known and have been previously studied for the bounded and sub-Gaussian observations [Lou14]. We extend these results to the case of heavy-tailed observations; in particular, we show that the estimator obtained via soft-thresholding applied to the eigenvalues of \( \hat{\Sigma}_* \) admits optimal guarantees in the Frobenius (as well as operator) norm.

Let \( \hat{\Sigma}_* \) be the estimator defined in the previous section, see equation (4.6), and set

\[
\hat{\Sigma}_*^\tau = \arg\min_{A \in \mathbb{R}^{d \times d}} \left[ \|A - \hat{\Sigma}_*\|_F^2 + \tau \|A\|_1 \right],
\]

where \( \tau > 0 \) controls the amount of penalty. It is well-known (e.g., see the proof of Theorem 1 in [Lou14]) that \( \hat{\Sigma}_*^\tau \) can be written explicitly as

\[
\hat{\Sigma}_*^\tau = \sum_{i=1}^{d} \max \left( \lambda_i(\hat{\Sigma}_*) - \tau/2, 0 \right) v_i(\hat{\Sigma}_*) v_i(\hat{\Sigma}_*)^T,
\]

where \( \lambda_i(\hat{\Sigma}_*) \) and \( v_i(\hat{\Sigma}_*) \) are the eigenvalues and corresponding eigenvectors of \( \hat{\Sigma}_* \). We are ready to state the main result of this section.

**Theorem 4.3.1.** For any \( \tau \geq 36\sigma_0 \sqrt{\frac{3}{m}} \),

\[
\|\hat{\Sigma}_*^\tau - \Sigma_0\|_F^2 \leq \inf_{A \in \mathbb{R}^{d \times d}} \left[ \|A - \Sigma_0\|_F^2 + \frac{(1 + \sqrt{2})^2}{8} \tau^2 \text{rank}(A) \right].
\]

with probability \( \geq 1 - 5d \log_2 \left( \frac{2\sigma_{\text{max}}}{\sigma_{\text{min}}} \right) e^{-\beta} \).

In particular, if \( \text{rank}(\Sigma_0) = r \) and \( \tau = 36\sigma_0 \sqrt{\frac{3}{m}} \), we obtain that

\[
\|\hat{\Sigma}_*^\tau - \Sigma_0\|_F^2 \leq 162 \sigma_0^2 \left( 1 + \sqrt{2} \right)^2 \frac{\beta r}{m}
\]

with probability \( \geq 1 - 5d \log_2 \left( \frac{2\sigma_{\text{max}}}{\sigma_{\text{min}}} \right) e^{-\beta} \).

### 4.4 Proofs

#### 4.4.1 Proof of Lemma 4.2.1

The result is a simple corollary of the following statement.
Lemma 4.4.1. Set $\theta = \frac{1}{\sigma} \sqrt{\frac{2}{m}}$, where $\sigma \geq \sigma_0$ and $m \geq \beta$. Let $d := \sigma_0^2 / \|\Sigma_0\|^2$. Then, with probability at least $1 - 5de^{-\beta}$,

$$
\left\| \hat{\Sigma} - \Sigma_0 \right\| \leq 2\sigma \sqrt{\frac{\beta}{m}} + C\|\Sigma_0\| \left( \sqrt{\frac{d\sigma}{\|\Sigma_0\|}} \left( \frac{\beta}{m} \right)^{\frac{3}{2}} + \sqrt{\frac{d\sigma}{\|\Sigma_0\|}} \left( \frac{\beta}{m} \right)^{\frac{1}{2}} + d \left( \frac{\beta}{m} \right)^{\frac{1}{2}} + \frac{d\beta^2}{m^2} + \frac{d^2}{m^3} \right),
$$

where $C' > 1$ is an absolute constant.

Now, by Corollary 4.5.1 in the supplement, it follows that $d = \sigma_0^2 / \|\Sigma_0\|^2 \geq \text{tr}(\Sigma_0) / \|\Sigma_0\| \geq 1$. Thus, assuming that the sample size satisfies $m \geq (6C')^4d\beta / \|\Sigma_0\|$, then, $d\beta/m \leq 1/(6C')^4 < 1$, and by some algebraic manipulations we have that

$$\left\| \hat{\Sigma} - \Sigma_0 \right\| \leq 2\sigma \sqrt{\frac{\beta}{m}} + \sigma \sqrt{\frac{\beta}{m}} = 3\sigma \sqrt{\frac{\beta}{m}}.$$  \hspace{1cm} (4.11)

For completeness, a detailed computation is given in the supplement. This finishes the proof.

4.4.2 Proof of Lemma 4.4.1

Let $B_\beta = 11\sqrt{2\text{tr}(\Sigma_0)\beta/m}$ be the error bound of the robust mean estimator $\hat{\mu}$ defined in (4.3). Let $Z_i = X_i - \mu_0$, $\Sigma_\mu = \mathbb{E}[(Z_i - \mu)(Z_i - \mu)^T]$, $\forall i = 1, 2, \ldots, d$, and

$$\hat{\Sigma}_\mu = \frac{1}{m\theta} \sum_{i=1}^m \frac{(X_i - \mu)(X_i - \mu)^T}{\|X_i - \mu\|_2^2} \psi \left( \theta \|X_i - \mu\|_2^2 \right),$$

for any $\|\mu\|_2 \leq B_\beta$. We begin by noting that the error can be bounded by the supremum of an empirical process indexed by $\mu$, i.e.

$$\left\| \hat{\Sigma} - \Sigma_0 \right\| \leq \sup_{\|\mu\|_2 \leq B_\beta} \left\| \hat{\Sigma}_\mu - \Sigma_0 \right\| \leq \sup_{\|\mu\|_2 \leq B_\beta} \left\| \hat{\Sigma}_\mu - \Sigma_\mu \right\| + \left\| \Sigma_\mu - \Sigma_0 \right\|$$  \hspace{1cm} (4.12)
with probability at least $1 - e^{-\beta}$. We first estimate the second term $\|\Sigma_\mu - \Sigma_0\|$. For any $\|\mu\|_2 \leq B_\beta$,

$$
\|\Sigma_\mu - \Sigma_0\| = \left\| \mathbb{E}[(Z_i - \mu)(Z_i - \mu)^T - Z_iZ_i^T] \right\| = \sup_{v \in \mathbb{R}^d : \|v\|_2 \leq 1} \left| \mathbb{E} \left[ (Z_i - \mu, v)^2 - (Z_i, v)^2 \right] \right| = (\mu^T v)^2 \leq \|\mu\|_2^2 \leq B_\beta^2 = \frac{242 \text{tr}(\Sigma_0) \beta}{m},
$$

with probability at least $1 - e^{-\beta}$. It follows from Corollary 4.5.1 in the supplement that with the same probability

$$
\|\Sigma_\mu - \Sigma_0\| \leq 242 \frac{\sigma_0^2 \beta}{\|\Sigma_0\| m} \leq 242 \frac{\sigma^2 \beta}{\|\Sigma_0\| m} = 242 \|\Sigma_0\| \frac{d \beta}{m}.
$$

(4.13)

Our main task is then to bound the first term in (4.12). To this end, we rewrite it as a double supremum of an empirical process:

$$
\sup_{\|\mu\|_2 \leq B_\beta} \left\| \hat{\Sigma}_\mu - \Sigma_\mu \right\| = \sup_{\|\mu\|_2 \leq B_\beta, \|v\|_2 \leq 1} \left| v^T \left( \hat{\Sigma}_\mu - \Sigma_\mu \right) v \right|
$$

It remains to estimate the supremum above.

**Lemma 4.4.2.** Set $\theta = \frac{1}{\sigma} \sqrt{\frac{\beta}{m}}$, where $\sigma \geq \sigma_0$ and $m \geq \beta$. Let $\overline{d} := \sigma_0^2 / \|\Sigma_0\|^2$. Then, with probability at least $1 - 4de^{-\beta}$,

$$
\sup_{\|\mu\|_2 \leq B_\beta, \|v\|_2 \leq 1} \left| v^T \left( \hat{\Sigma}_\mu - \Sigma_\mu \right) v \right| \leq 2\sigma \sqrt{\frac{\beta}{m}} + C'' \|\Sigma_0\| \left( \sqrt{\overline{d} \sigma_0} \left( \frac{\beta}{m} \right)^{\frac{3}{2}} + \sqrt{\overline{d} \sigma_0} \frac{\beta}{m} + \sqrt{\overline{d} \sigma_0} \left( \frac{\beta}{m} \right)^{\frac{3}{2}} + \overline{d} \left( \frac{\beta}{m} \right)^{\frac{3}{2}} + \overline{d} \sigma_0 \left( \frac{\beta}{m} \right)^{\frac{3}{2}} \right),
$$

where $C'' > 1$ is an absolute constant.

Note that $\sigma \geq \sigma_0$ by definition, thus, $\overline{d} \leq \sigma^2 / \|\Sigma_0\|^2$. Combining the above lemma with (4.12) and (4.13) finishes the proof.
4.4.3 Proof of Theorem 4.2.1

Define \( \overline{j} := \min \{ j \in J : \sigma_j \geq \sigma_0 \} \), and note that \( \sigma_{\overline{j}} \leq 2\sigma_0 \). We will demonstrate that

\[
\overline{j} \leq j^* \quad \text{with high probability.}
\]

Observe that

\[
\Pr (j^* > \overline{j}) \leq \Pr \left( \bigcup_{k \in J : k > \overline{j}} \{ \| \hat{\Sigma}_{m,k} - \hat{\Sigma}_{m,\overline{j}} \| > 6\sigma_k \sqrt{\frac{\beta}{m}} \} \right)
\]

\[
\leq \Pr \left( \| \hat{\Sigma}_{m,\overline{j}} - \hat{\Sigma}_0 \| > 3\sigma_{\overline{j}} \sqrt{\frac{\beta}{m}} \right) + \sum_{k \in J : k > \overline{j}} \Pr \left( \| \hat{\Sigma}_{m,k} - \hat{\Sigma}_0 \| > 3\sigma_k \sqrt{\frac{\beta}{m}} \right)
\]

\[
\leq 5de^{-\beta} + 5d \log_2 \left( \frac{\sigma_{\max}}{\sigma_{\min}} \right) e^{-\beta},
\]

where we applied (4.5) to estimate each of the probabilities in the sum under the assumption that the number of samples \( m \geq C \overline{j} \beta \) and \( \sigma_k \geq \sigma_{\overline{j}} \geq \sigma_0 \). It is now easy to see that the event

\[
\mathcal{B} = \bigcap_{k \in J : k > \overline{j}} \left\{ \| \hat{\Sigma}_{m,k} - \hat{\Sigma}_0 \| \leq 3\sigma_k \sqrt{\frac{\beta}{m}} \right\}
\]

of probability \( 1 - 5d \log_2 \left( \frac{\sigma_{\max}}{\sigma_{\min}} \right) e^{-\beta} \) is contained in \( \mathcal{E} = \{ j^* \leq \overline{j} \} \). Hence, on \( \mathcal{B} \)

\[
\| \hat{\Sigma}_* - \hat{\Sigma}_0 \| \leq \| \hat{\Sigma}_* - \hat{\Sigma}_{m,\overline{j}} \| + \| \hat{\Sigma}_{m,\overline{j}} - \hat{\Sigma}_0 \| \leq 6\sigma_{\overline{j}} \sqrt{\frac{\beta}{m}} + 3\sigma_{\overline{j}} \sqrt{\frac{\beta}{m}}
\]

\[
\leq 12\sigma_0 \sqrt{\frac{\beta}{m}} + 6\sigma_0 \sqrt{\frac{\beta}{m}} = 18\sigma_0 \sqrt{\frac{\beta}{m}},
\]

and the claim follows.

4.4.4 Proof of Theorem 4.3.1

The proof is based on the following lemma:

**Lemma 4.4.3.** Inequality (4.10) holds on the event \( \mathcal{E} = \{ \tau \geq 2 \| \hat{\Sigma}_* - \hat{\Sigma}_0 \| \} \).

To verify this statement, it is enough to repeat the steps of the proof of Theorem 1 in [Lou14], replacing each occurrence of the sample covariance matrix by its “robust analogue” \( \hat{\Sigma}_* \).

It then follows from Theorem 4.2.1 that \( \Pr(\mathcal{E}) \geq 1 - 5d \log_2 \left( \frac{2\sigma_{\max}}{\sigma_{\min}} \right) e^{-\beta} \) whenever \( \tau \geq 36\sigma_0 \sqrt{\frac{\beta}{m}} \).
4.5  Proof of Additional Technical Lemmas

4.5.1 Preliminaries

Lemma 4.5.1. Consider any function \( \phi : \mathbb{R} \to \mathbb{R} \) and \( \theta > 0 \). Suppose the following holds

\[
- \frac{1}{\theta} \log (1 - \theta x + \theta^2 x^2) \leq \phi(x) \leq \frac{1}{\theta} \log (1 + \theta x + \theta^2 x^2), \ \forall x \in \mathbb{R} \quad (4.14)
\]

then, we have for any matrix \( A \in \mathbb{R}^{d \times d} \),

\[
- \frac{1}{\theta} \log (1 - \theta A + \theta^2 A^2) \leq \phi(A) \leq \frac{1}{\theta} \log (I + \theta A + \theta^2 A^2).
\]

Proof. Note that for any \( x \in \mathbb{R} \), \( - \frac{1}{\theta} \log (1 - x \theta + x^2 \theta^2) \leq \frac{1}{\theta} \log (1 + x \theta + x^2 \theta^2) \), then, the claim follows immediately from the definition of the matrix function. \( \square \)

The above lemma is useful in our context mainly due to the following lemma,

Lemma 4.5.2. The truncation function \( \frac{1}{\theta} \psi(\theta x) = \text{sign}(x) \cdot (|x| \wedge \frac{1}{\theta}) \) satisfies the assumption \((4.14)\) in Lemma 4.5.1.

Proof. Denote \( f_1(x) = - \frac{1}{\theta} \log (1 - \theta x + \theta^2 x^2), f_2(x) = \frac{1}{\theta} \log (1 + \theta x + \theta^2 x^2) \) and \( g(x) = \text{sign}(x) \cdot (|x| \wedge \frac{1}{\theta}) \). Note first that

\[
\begin{align*}
 f_1(0) &= g(0) = f_2(0) = 0, \\
 f_1(1/\theta) &\leq g(1/\theta) \leq f_2(1/\theta), \\
 f_1(-1/\theta) &\leq g(-1/\theta) \leq f_2(-1/\theta), \\
\end{align*}
\]

and the subgradient

\[
\partial g(x) = \begin{cases} 
1, & x \in (-1/\theta, 1/\theta), \\
0, & x \in (-\infty, -1/\theta) \cup (1/\theta, +\infty), \\
[0, 1], & x = -1/\theta, 1/\theta. 
\end{cases}
\]

128
Next, we take the derivative of \( f_2(x) \) and compare it to the derivative of \( g(x) \).

\[
f'_2(x) = \frac{1}{\theta} \cdot \frac{\theta + 2x\theta^2}{1 + x\theta + x^2\theta^2} = \frac{1 + 2x\theta}{1 + x\theta + x^2\theta^2}.
\]

Note that \( f'_2(x) \geq 1, x \in (0,1/\theta), f'_2(x) \geq 0, x \geq 1/\theta, f'_2(x) \leq 1, x \in (-1/\theta,0] \) and \( f'_2(x) \leq 0, x \leq -1/\theta \). Thus, we have \( g(x) \leq f'_2(x), \forall x \in \mathbb{R} \). Similarly, we can take the derivative of \( f_1(x) \) and compare it to \( g(x) \), which results in \( f'_1(x) \leq 1, x \in (0,1/\theta), f'_1(x) \leq 0, x \geq 1/\theta, f'_1(x) \geq 1, x \in (-1/\theta,0] \) and \( f'_2(x) \geq 0, x \leq -1/\theta \). This implies \( f_1(x) \leq g(x) \) and the Lemma is proved.

The following lemma demonstrates the importance of matrix logarithm function in matrix analysis, whose proof can be found in [Bha13] and [Tro15b].

**Lemma 4.5.3.** (a) The matrix logarithm is operator monotone, that is, if \( A \succ B \succ 0 \) are two matrices in \( \mathbb{H}^{d \times d} \), then, \( \log(A) \succ \log(B) \).

(b) Given a fixed matrix \( H \in \mathbb{H}^{d \times d} \), the function

\[
A \to \text{tr} \exp(H + \log(A))
\]

is concave on the cone of positive semi-definite matrices.

The following lemma is a generalization of Chebyshev’s association inequality. See Theorem 2.15 of [BLM13] for proof.

**Lemma 4.5.4 (FKG inequality).** Suppose \( f, g : \mathbb{R}^d \to \mathbb{R} \) are two functions non-decreasing on each coordinate. Let \( Y = [Y_1, Y_2, \cdots, Y_d] \) be a random vector taking values in \( \mathbb{R}^d \), then,

\[
\mathbb{E}[f(Y)g(X)] \geq \mathbb{E}[f(Y)]\mathbb{E}[g(Y)].
\]

The following corollary follows immediately from the FKG inequality.

**Corollary 4.5.1.** Let \( Z = X - \mu_0 \), then, we have \( \sigma_0^2 = \|Z\|_2^2 \|Z\|_2^2\| \geq \text{tr} \left( \mathbb{E}[ZZ^T] \right) \| \mathbb{E}[ZZ^T] \| = \text{tr}(\Sigma_0)\|\Sigma_0\| \).

**Proof.** Consider any unit vector \( v \in \mathbb{R}^d \). It is enough to show \( \mathbb{E}[(v^T Z)^2\|Z\|_2^2] \geq \mathbb{E}[(v^T Z)^2]\mathbb{E}[\|Z\|_2^2] \).

We change the coordinate by considering an orthonormal basis \( \{v_1, \cdots, v_d\} \) with \( v_1 = v \). Let
\( Y_i = v_i^T Z, \ i = 1, 2, \ldots, d, \) then we obtain,

\[
E[(v^T Z)^2 \| Z \|_2^2] = E[Y_1^2 \| Y \|_2^2] \geq E[Y_1^2]E[\| Y \|_2^2],
\]

where the last inequality follows from FKG inequality by taking \( f(Y_1^2, \ldots, Y_d^2) = Y_1^2 \) and \( g(Y_1^2, \ldots, Y_d^2) = \| Y \|_2^2. \)

4.5.2 Additional computation in the proof of Lemma 4.2.1

In order to show (4.11), it is enough to show that

\[
C' \| \Sigma_0 \| \left( \sqrt{\frac{d \sigma}{\| \Sigma_0 \|}} \left( \frac{\beta}{m} \right)^{\frac{1}{4}} + \sqrt{\frac{d \sigma}{\| \Sigma_0 \|}} \frac{\beta}{m} \right) + \sqrt{\frac{d \sigma}{\| \Sigma_0 \|}} \left( \frac{\beta}{m} \right)^{\frac{1}{4}} + \frac{d \beta^2}{m^2} + \frac{d^3}{2} \left( \frac{\beta}{m} \right)^{\frac{3}{4}} \right) \leq \sigma \sqrt{\frac{\beta}{m}}.
\]

Note that \( \frac{d \sigma^2}{\| \Sigma_0 \|^2} \geq \text{tr}(\Sigma_0)/\| \Sigma_0 \| \geq 1, \) and assuming that the sample size satisfies \( m \geq (6C')^4 \bar{d} \beta, \) we have \( \bar{d} \beta / m \leq 1/(6C')^4 < 1. \) We then bound each of the 6 terms on the left side.

\[
C' \| \Sigma_0 \| \sqrt{\frac{d \sigma}{\| \Sigma_0 \|}} \left( \frac{\beta}{m} \right)^{\frac{1}{4}} = C' \sqrt{\sigma} \left( \frac{\beta}{m} \right)^{\frac{1}{4}} \cdot \left( \frac{\| \Sigma_0 \| \bar{d} \beta}{m} \right)^{1/4} \cdot \left( \frac{\| \Sigma_0 \| \bar{d} \beta}{m} \right)^{1/4} \\
\leq C' \sqrt{\sigma} \left( \frac{\beta}{m} \right)^{\frac{1}{4}} \cdot \left( \frac{\| \Sigma_0 \| \bar{d} \beta}{m} \right)^{1/4} \cdot \frac{1}{6C'} \\
= \frac{1}{6} \sqrt{\sigma \sigma_0} \sqrt{\frac{\beta}{m}} \leq \frac{1}{6} \sigma \sqrt{\frac{\beta}{m}}.
\]

\[
C' \| \Sigma_0 \| \cdot \sqrt{\frac{d \sigma}{\| \Sigma_0 \|}} \left( \frac{\beta}{m} \right)^{\frac{1}{2}} = C' \sigma \sqrt{\frac{\beta}{m}} \cdot \sqrt{\frac{d \beta}{m}} \leq C' \sigma \sqrt{\frac{\beta}{m}} \frac{1}{6C'} \leq \frac{1}{6} \sigma \sqrt{\frac{\beta}{m}}.
\]

\[
C' \| \Sigma_0 \| \sqrt{\frac{d \sigma}{\| \Sigma_0 \|}} \left( \frac{\beta}{m} \right)^{\frac{1}{4}} \leq C' \| \Sigma_0 \| \sqrt{\frac{d \sigma}{\| \Sigma_0 \|}} \left( \frac{\beta}{m} \right)^{\frac{1}{4}} \leq \frac{1}{6} \sigma \sqrt{\frac{\beta}{m}}.
\]

Note that we have the following

\[
C' \| \Sigma_0 \| \frac{\bar{d} \beta}{m} = C' \| \Sigma_0 \| \left( \frac{\bar{d} \beta}{m} \right)^{\frac{1}{2}} \left( \frac{\bar{d} \beta}{m} \right)^{\frac{1}{4}} \leq C' \| \Sigma_0 \| \left( \frac{\bar{d} \beta}{m} \right)^{\frac{1}{2}} \frac{1}{6C'} \leq \frac{1}{6} \sigma_0 \sqrt{\frac{\beta}{m}} \leq \frac{1}{6} \sigma \sqrt{\frac{\beta}{m}},
\]

130
thus, the rest three terms can be bounded as follows,

\[
C'\|\Sigma_0\|\mathbb{E}\left(\frac{1}{m}\right)^{\frac{3}{2}} \leq C'\|\Sigma_0\|\mathbb{E}\left(\frac{1}{m}\right)^{\frac{3}{2}} \leq \frac{1}{6}\sqrt{\frac{\beta}{m}}
\]

\[
C'\|\Sigma_0\|\frac{\beta^2}{m^2} \leq C'\|\Sigma_0\|\frac{\beta^2}{m^2} \leq \frac{1}{6}\sqrt{\frac{\beta}{m}}
\]

\[
C'\|\Sigma_0\|\overline{d}^3 \left(\frac{\beta}{m}\right)^{\frac{3}{2}} \leq C'\|\Sigma_0\|\overline{d}^3 \left(\frac{\beta}{m}\right)^{\frac{3}{2}} \leq C'\|\Sigma_0\|\overline{d}^3 \left(\frac{\beta}{m}\right)^{\frac{3}{2}} \leq \frac{1}{6}\sigma\sqrt{\frac{\beta}{m}}.
\]

Overall, we have \((4.11)\) holds.

### 4.5.3 Proof of Lemma 4.4.2

First of all, by definition of \(\hat{\Sigma}_\mu\), we have

\[
\sup_{\|\mu\|_2 \leq B_\beta, \|v\|_2 \leq 1} \left|v^T(\hat{\Sigma}_\mu - \Sigma_\mu)v\right| = \sup_{\|\mu\|_2 \leq B_\beta, \|v\|_2 \leq 1} \left|\frac{1}{m}\sum_{i=1}^{m} (Z_i - \mu)^2 \frac{\psi(\theta\|Z_i - \mu\|_2^2)}{\|Z_i - \mu\|_2^2} - E\left[(Z_i - \mu, v)^2\right]\right|.
\]

Expanding the squares on the right hand side gives

\[
\sup_{\|\mu\|_2 \leq B_\beta} \left\|\hat{\Sigma}_\mu - \Sigma_\mu\right\| \leq \sup_{\|\mu\|_2 \leq B_\beta, \|v\|_2 \leq 1} \left|\frac{1}{m}\sum_{i=1}^{m} (Z_i, v)^2 \frac{\psi(\theta\|Z_i - \mu\|_2^2)}{\theta\|Z_i - \mu\|_2^2} - E\left((Z_i, v)^2\right)\right| (I)
\]

\[
+ 2\sup_{\|\mu\|_2 \leq B_\beta, \|v\|_2 \leq 1} \left|\frac{1}{m}\sum_{i=1}^{m} (Z_i, v) (\mu, v) \frac{\psi(\theta\|Z_i - \mu\|_2^2)}{\theta\|Z_i - \mu\|_2^2} - E\left((Z_i, v) (\mu, v)\right)\right| (II)
\]

\[
+ \sup_{\|\mu\|_2 \leq B_\beta, \|v\|_2 \leq 1} \left|\frac{1}{m}\sum_{i=1}^{m} (\mu, v)^2 \frac{\psi(\theta\|Z_i - \mu\|_2^2)}{\theta\|Z_i - \mu\|_2^2} - (\mu, v)^2\right| (III)
\]

We will then bound these three terms separately. Note that given \(\|\bar{\mu} - \mu_0\|_2 \leq B_\beta\), the term (III) can be readily bounded as follows using the fact that \(0 \leq \psi(x) \leq x, \forall x \geq 0\),

\[
(III) = \sup_{\|\mu\|_2 \leq B_\beta, \|v\|_2 \leq 1} \left|\langle \mu, v \rangle^2 \left(\frac{1}{m}\sum_{i=1}^{m} \frac{\psi(\theta\|Z_i - \mu\|_2^2)}{\theta\|Z_i - \mu\|_2^2} - 1\right)\right| \leq \sup_{\|\mu\|_2 \leq B_\beta, \|v\|_2 \leq 1} \langle \mu, v \rangle^2 \leq B_\beta^2
\]

\[
= 242\left(\frac{\theta(\Sigma_0)}{m}\right)^2 \leq 242\left(\frac{\sigma_{1/2}^2}{\|\Sigma_0\|^2}\right) \leq 242\|\Sigma_0\|\overline{d}\beta, \frac{\beta}{m}. (4.15)
\]

where the second from the last inequality follows from Corollary 4.5.1 and the last inequality follows from \(\overline{d} = \sigma_{1/2}^2/\|\Sigma_0\|^2\).

The rest two terms are bounded through the following lemma whose proof is delayed to the next section:
Lemma 4.5.5. Given $\|\hat{\mu} - \mu_0\|_2 \leq B_\beta$, with probability at least $1 - 4de^{-\beta}$, we have the following two bounds hold,

$$(I) \leq 2\sigma \sqrt{\frac{\beta}{m}} + 22\|\Sigma_0\| \left( \sqrt{2d} \frac{\beta}{m} \right)^\frac{3}{4} + 2\sqrt{2} \sqrt{\frac{\beta}{\|\Sigma_0\|}} \left( \frac{\beta}{m} \right)^\frac{3}{4} + 11d^\frac{1}{4} \left( \frac{\beta}{m} \right)^\frac{3}{4} + 22 \frac{d\beta^2}{m^2},$$

$$(II) \leq 11\|\Sigma_0\| \left( \sqrt{2} \sqrt{\frac{\beta}{\|\Sigma_0\|}} \left( \frac{\beta}{m} \right)^\frac{3}{4} + 3\sqrt{2} \sqrt{d} \frac{\beta}{\|\Sigma_0\|} \frac{\beta}{m} + 44d^\frac{3}{4} \left( \frac{\beta}{m} \right)^\frac{3}{4} \right.$$

$$+ 44\sqrt{2d} \left( \frac{\beta}{m} \right)^\frac{3}{4} + 242\sqrt{2} \frac{d\beta^2}{m^2} + 484d^\frac{5}{4} \left( \frac{\beta}{m} \right)^\frac{5}{4} \left.$$.}

Note that since $\sigma \geq \sigma_0$, we have $\sigma/\|\Sigma_0\| \geq \sigma_0/\|\Sigma_0\| = \sqrt{d}$. Combining the above lemma with (4.15) finishes the proof of Lemma 4.4.2.

4.5.4 Proof of Lemma 4.5.5

Before proving the Lemma, we introduce the following abbreviations:

$$g_v(Z_i) = \langle Z_i, v \rangle^2 \frac{\psi(\theta\|Z_i\|_2^2)}{\theta\|Z_i\|_2^2}, \quad h_\mu(Z_i) = \frac{\|Z_i\|_2^2}{\psi(\theta\|Z_i\|_2^2)} \frac{\psi(\theta\|Z_i - \mu\|_2^2)}{\theta\|Z_i - \mu\|_2^2},$$

$$\tilde{g}_v(Z_i) = \langle Z_i, v \rangle \frac{\psi(\theta\|Z_i\|_2^2)}{\theta\|Z_i\|_2^2}.$$

Our analysis relies on the following simply yet important fact which gives deterministic upper and lower bound of $h_\mu(Z_i)$ around 1. Its proof is delayed to the next section.

Lemma 4.5.6. For any $\mu$ such that $\|\mu\|_2 \leq B_\beta$, the following holds:

$$1 - 2B_\beta \sqrt{\theta} - B_\beta^2 \theta \leq h_\mu(Z_i) \leq 1 + 2B_\beta \sqrt{\theta} + B_\beta^2 \theta.$$
assumption (4.14) of Lemma 4.5.1. Then, for any \( t > 0 \),
\[
Pr \left( \sum_{i=1}^{m} (\phi(A_i) - \mathbb{E}[A_i]) \geq t\sqrt{m} \right) \leq 2d \exp \left( -t\theta \sqrt{m} + m\theta^2 \sigma_A^2 \right).
\]

Specifically, if the assumption (4.14) holds for \( \theta = \frac{t}{2\sqrt{m}\sigma_A} \), then we obtain the subgaussian tail
\[
2d \exp(-t^2/4\sigma_A^2).
\]

The intuition behind this lemma is that the \( \log(1 + x) \) tends to “robustify” a random variable by implicitly trading the bias for a tight concentration. A scalar version of such lemma with a similar idea is first introduced in the seminal work [Cat12]. The proof of the current matrix version is similar to Lemma 3.1 and Theorem 3.1 of [Min16] by modifying only the constants. We omitted the details here for brevity. Note that this lemma is useful in our context by choosing \( \phi(x) = \frac{1}{\theta} \psi(\theta x) \). Next, we prove two parts of Lemma 4.5.5 separately.

**Proof of (I) in Lemma 4.5.5.** Using the abbreviation introduced at the beginning of this section, we have
\[
(I) = \sup_{\|\mu\|_2 \leq B_\beta, \|v\|_2 \leq 1} \left| \frac{1}{m} \sum_{i=1}^{m} g_v(Z_i) h_\mu(Z_i) - \mathbb{E} \left[ \langle Z_i, v \rangle^2 \right] \right|
\]

We further split it into two terms as follows:
\[
(I) \leq \sup_{\|\mu\|_2 \leq B_\beta, \|v\|_2 \leq 1} \left| \frac{1}{m} \sum_{i=1}^{m} g_v(Z_i) (h_\mu(Z_i) - 1) \right| + \sup_{\|v\| \leq 1} \left| \frac{1}{m} \sum_{i=1}^{m} g_v(Z_i) - \mathbb{E} \left[ \langle Z_i, v \rangle^2 \right] \right| \tag{4.16}
\]

The two terms in (4.16) are bounded as follows:

1. For the second term in (4.16), note that we can write it back into the matrix form as
\[
\left\| \frac{1}{m\theta} \sum_{i=1}^{m} Z_i Z_i^T \psi(\theta \|Z_i\|_2^2) - \mathbb{E} \left[ Z_i Z_i^T \right] \right\|.
\]

Note that the matrix \( Z_i Z_i^T \) is a rank one matrix with the eigenvalue equal to \( \|Z_i\|_2^2 \), so it follows from the definition of matrix function,
\[
Z_i Z_i^T \psi(\theta \|Z_i\|_2^2) = \frac{1}{\theta} \psi(\theta Z_i Z_i^T).
\]
Now, applying Lemma 4.5.2 setting \( \theta = \frac{t}{2\sigma^2 \sqrt{m}} \) together with Lemma 4.5.7 gives
\[
Pr \left( \left\| \frac{1}{m \theta} \sum_{i=1}^{m} Z_i Z_i^T \psi \left( \frac{\theta \|Z_i\|_2^2}{\|Z_i\|_2^2} \right) - \mathbb{E} [Z_i Z_i^T] \right\| \geq \frac{t}{\sqrt{m}} \right) \leq 2d \exp \left( -\frac{t^2}{4\sigma^2} \right).
\]

Setting \( t = 2\sigma \sqrt{\beta} \) (which results in \( \theta = \frac{1}{\sigma^2 \sqrt{\frac{\beta}{m}}} \)) gives
\[
\left\| \frac{1}{m \theta} \sum_{i=1}^{m} Z_i Z_i^T \psi \left( \frac{\theta \|Z_i\|_2^2}{\|Z_i\|_2^2} \right) - \mathbb{E} [Z_i Z_i^T] \right\| \leq 2\sigma \sqrt{\frac{\beta}{m}} \tag{4.17}
\]
with probability at least \( 1 - 2d e^{-\beta} \).

2. For the first term in (4.16), by the fact that \( g_v(Z_i) \geq 0 \) and Lemma 4.5.6
\[
\sup_{\|\mu\|_2 \leq B_\beta, \|v\|_2 \leq 1} \left| \frac{1}{m} \sum_{i=1}^{m} g_v(Z_i) (h_{\mu}(Z_i) - 1) \right|
\leq \sup_{\|\mu\|_2 \leq B_\beta, \|v\|_2 \leq 1} \left\| \frac{1}{m} \sum_{i=1}^{m} g_v(Z_i) \|h_{\mu}(Z_i) - 1\| \right\|
\leq \sup_{\|v\|_2 \leq 1} \frac{1}{m} \sum_{i=1}^{m} g_v(Z_i) \left( 2B_\beta \sqrt{\theta} + B_\beta^2 \theta \right)
\leq \left( \left\| \mathbb{E} [Z_i Z_i^T] \right\| + 2\sigma \sqrt{\frac{\beta}{m}} \right) \left( 2B_\beta \sqrt{\theta} + B_\beta^2 \theta \right),
\]
with probability at least \( 1 - 2d e^{-\beta} \), where the last inequality follows from the same argument leading to (4.17). Note that \( \mathbb{E} [Z_i Z_i^T] = \Sigma_0 \).

Overall, we get
\[
(I) \leq 2\sigma \sqrt{\frac{\beta}{m}} + \left( \|\Sigma_0\| + 2\sigma \sqrt{\frac{\beta}{m}} \right) \left( 2B_\beta \sqrt{\theta} + B_\beta^2 \theta \right),
\]
with probability at least \( 1 - 2d \). Now we substitute \( B_\beta = 11 \sqrt{2 \text{tr}(\Sigma_0) \beta / m} \) and \( \theta = \frac{1}{\sigma^2 \sqrt{\frac{\beta}{m}}} \) into the above bound gives
\[
(I) \leq 2\sigma \sqrt{\frac{\beta}{m}} + 22 \sqrt{2\|\Sigma_0\| \sqrt{\frac{\text{tr}(\Sigma_0)}{\sigma}}} \left( \frac{\beta}{m} \right)^{\frac{1}{4}} + 242 \|\Sigma_0\| \frac{\text{tr}\Sigma_0}{\sigma} \left( \frac{\beta}{m} \right)^{\frac{1}{2}}
+ 44 \sqrt{2 \sigma \text{tr}(\Sigma_0)} \left( \frac{\beta}{m} \right)^{\frac{1}{2}} + 484 \text{tr}(\Sigma_0) \left( \frac{\beta}{m} \right)^2
\]

134
Using Corollary 4.5.1, we have

\[
\frac{\text{tr}(\Sigma_0)}{\sigma} \leq \frac{\text{tr}(\Sigma_0)}{\sigma_0} \leq \frac{\text{tr}(\Sigma_0)}{\sqrt{\text{tr}(\Sigma_0)\|\Sigma_0\|}} \leq \frac{\|\Sigma_0\|}{\sigma_0} \leq \mathcal{A},
\]

and also,

\[
\text{tr}(\Sigma_0) \leq \|\Sigma_0\| \sigma_0^2 \leq \|\Sigma_0\| \mathcal{A}.
\]

Substitute these two bounds into the bound of (I) gives the final bound for (I) stated in Lemma 4.5.5 with probability at least $1 - 2de^{-\beta}$.

**Proof of (II) in Lemma 4.5.5.** First of all, using the definition of $\tilde{g}_\nu(Z_i)$ and $h_\mu(Z_i)$, we can rewrite (II) as follows:

\[
(\text{II}) = \sup_{\|\mu\|_2 \leq B_\beta, \|v\|_2 \leq 1} \left| \frac{1}{m} \sum_{i=1}^m \tilde{g}_\nu(Z_i)h_\mu(Z_i) \langle \mu, v \rangle - E[\langle Z_i, v \rangle] \langle \mu, v \rangle \right|
\]

\[
\leq B_\beta \cdot \sup_{\|\mu\|_2 \leq B_\beta, \|v\|_2 \leq 1} \left| \frac{1}{m} \sum_{i=1}^m \tilde{g}_\nu(Z_i)h_\mu(Z_i) - E[\langle Z_i, v \rangle] \right|
\]

Similar to the analysis of (I), we further split the above term into two terms and get

\[
(\text{II}) \leq B_\beta \sup_{\|\mu\|_2 \leq B_\beta, \|v\|_2 \leq 1} \left| \frac{1}{m} \sum_{i=1}^m \tilde{g}_\nu(Z_i)(h_\mu(Z_i) - 1) \right| + B_\beta \sup_{\|v\|_2 \leq 1} \left| \frac{1}{m} \sum_{i=1}^m \tilde{g}_\nu(Z_i) - E[\langle Z_i, v \rangle] \right|
\]

(4.20)

For the first term, by Cauchy-Schwarz inequality and then Lemma 4.5.6, we get

\[
(\text{IV}) \leq B_\beta \sup_{\|\mu\|_2 \leq B_\beta, \|v\|_2 \leq 1} \left| \frac{1}{m} \sum_{i=1}^m \tilde{g}_\nu(Z_i)(h_\mu(Z_i) - 1) \right|
\]

\[
\leq B_\beta \sup_{\|\mu\|_2 \leq B_\beta, \|v\|_2 \leq 1} \left( \frac{1}{m} \sum_{i=1}^m \tilde{g}_\nu(Z_i)^2 \right)^{1/2} \left( \frac{1}{m} \sum_{i=1}^m |h_\mu(Z_i) - 1|^2 \right)^{1/2}
\]

\[
\leq B_\beta \sup_{\|v\|_2 \leq 1} \left( \frac{1}{m} \sum_{i=1}^m \tilde{g}_\nu(Z_i)^2 \right)^{1/2} \left( 2B_\beta \sqrt{\theta} + B_\beta^2 \theta \right)
\]

Note that $\frac{1}{2} \psi \left( \theta \|Z_i\|_2^2 \right) / \|Z_i\|_2^2 \leq 1$, then, it follows,

\[
\tilde{g}_\nu(Z_i)^2 = \langle Z_i, v \rangle^2 \left( \frac{1}{2} \psi \left( \theta \|Z_i\|_2^2 \right) \right)^2 \leq \|Z_i, v \|^2 \frac{1}{2} \psi \left( \theta \|Z_i\|_2^2 \right) \left( \frac{|Z_i, v|^2}{\|Z_i\|_2^2} \right)
\]

(4.21)
Thus, by the same analysis leading to (4.17), we get

$$ (IV) \leq B_\beta \left( \|E[Z_i Z_i^T]\| + 2\sigma \sqrt{\frac{\beta}{m}} \right)^{1/2} \left( 2B_\beta \sqrt{\theta} + B_\beta^2 \theta \right), \quad (4.21) $$

with probability at least $1 - 2de^{-\beta}$. For the second term (V), notice that $E[Z_i] = 0$, thus we have

$$ (V) \leq B_\beta \sup_{\|v\|_2 \leq 1} \left\| \frac{1}{m} \sum_{i=1}^m \frac{Z_i}{\|Z_i\|_2} \psi(\theta \|Z_i\|_2^2, v) \right\| \leq B_\beta \left\| \frac{1}{m} \sum_{i=1}^m \frac{Z_i}{\|Z_i\|_2} \|Z_i\|_2^2 \wedge \frac{1}{\theta} \right\|_2 $$

$$ \leq B_\beta \left\| \frac{1}{m} \sum_{i=1}^m \frac{Z_i}{\|Z_i\|_2} \|Z_i\|_2^2 \wedge \frac{1}{\theta} - E \left[ \frac{Z_i}{\|Z_i\|_2} \|Z_i\|_2^2 \wedge \frac{1}{\theta} \right] \right\|_2 + B_\beta \left\| E \left[ \frac{Z_i}{\|Z_i\|_2} \|Z_i\|_2^2 \wedge \frac{1}{\theta} \right] \right\|_2. \quad (4.22) $$

For the second term, which measures the bias, we have by the fact $E[Z_i] = 0$,

$$ \left\| E \left[ \frac{Z_i}{\|Z_i\|_2} \|Z_i\|_2^2 \wedge \frac{1}{\theta} - 1 \right] \right\|_2 = \sup_{\|v\|_2 \leq 1} E \left[ \langle Z_i, v \rangle \left( \|Z_i\|_2^2 \wedge \frac{1}{\theta} - 1 \right) \right] \leq \sup_{\|v\|_2 \leq 1} E \left[ \langle Z_i, v \rangle \mathbf{1}_{\|Z_i\|_2 \geq 1/\sqrt{\theta}} \right]. $$

Now by Cauchy-Schwarz inequality and then Markov inequality, we obtain,

$$ \sup_{\|v\|_2 \leq 1} E \left[ \langle Z_i, v \rangle \mathbf{1}_{\|Z_i\|_2 \geq 1/\sqrt{\theta}} \right] \leq \sqrt{\sup_{\|v\|_2 \leq 1} E \left[ \langle Z_i, v \rangle^2 \right]} Pr(\|Z_i\|_2 \geq 1/\sqrt{\theta})^{1/2} \leq \sqrt{\|\Sigma_0\| E \left[ \|Z_i\|_2^2 \right]^{1/2} \sqrt{\theta}} $$

$$ = \sqrt{\|\Sigma_0\| \text{tr}(\Sigma_0)^{1/2} \beta^{1/4}} = \sqrt{\frac{\|\Sigma_0\| \text{tr}(\Sigma_0)^{1/4} \beta^{1/4}}{m^{1/4}} \leq \left( \frac{\sigma^2}{m^2} \right)^{1/4}, $$

where the last two inequalities both follow from Lemma 4.5. This gives the second term in (4.22) is given by $B_\beta \left( \frac{\sigma^2}{m^2} \right)^{1/4}$.

For the first term in (4.22), note that for any vector $x \in \mathbb{R}^d$,

$$ \|x\|_2 = \left\| \begin{bmatrix} 0 & x^T \\ x & 0 \end{bmatrix} \right\|, $$

and furthermore, the matrix $\begin{bmatrix} 0 & x^T \\ x & 0 \end{bmatrix}$ has two same eigenvalues equal to $\|x\|_2$, which follows
\[
\begin{bmatrix}
0 & x^T \\
0 & 0
\end{bmatrix}^2 = \begin{bmatrix}
\|x\|^2 & 0 \\
0 & xx^T
\end{bmatrix}.
\]

Thus, if we take
\[
A_i = \begin{bmatrix}
0 & Z_i^T \\
Z_i & 0
\end{bmatrix} \frac{\|Z_i\|_2^2 \wedge \frac{1}{\theta}}{\|Z_i\|_2^2},
\]

Then, the first term of (4.22) is equal to \(\frac{1}{m} \sum_{i=1}^{m} A_i - \mathbb{E}[A_i]\). For this \(A_i\), we have
\[
\|\mathbb{E}[A_i^2]\| \leq \mathbb{E}[\|Z_i\|_2^2] = \text{tr}(\Sigma_0), \quad \|A_i\| \leq \frac{1}{\sqrt{\theta}} = \frac{m^{1/4} \sigma^{1/2}}{\beta^{1/4}}.
\]

By matrix Bernstein’s inequality ([Tro12]), we obtain the bound
\[
\text{Pr}\left(\left\|\frac{1}{m} \sum_{i=1}^{m} A_i - \mathbb{E}[A_i]\right\| \geq t\right) \leq d \exp\left(- \frac{3}{8} \left(\frac{mt^2}{\sigma^2} \wedge m\sqrt{\theta} t\right)\right) = d \exp\left(- \frac{3}{8} \left(\frac{mt^2}{\sigma^2} \wedge m^{3/4} \beta^{1/4} t\right)\right),
\]

where \(c\) is a fixed positive constant. Taking \(t = 3\sqrt{\frac{\sigma^2 \beta}{\|\Sigma_0\| m}}\) gives
\[
\text{Pr}\left(\left\|\frac{1}{m} \sum_{i=1}^{m} A_i - \mathbb{E}[A_i]\right\| \geq 3\sqrt{\frac{\sigma^2 \beta}{m}}\right) \leq d \exp\left(-3\beta \wedge \left(\frac{m^{1/4} \beta^{3/4} \sigma^2}{\sigma^{1/2}}\right)^{1/4}\right) \leq d \exp(-\beta),
\]

where \(\beta = \sigma^2 / \|\Sigma_0\|^2 \geq \sigma_0^2 / \|\Sigma_0\|^2 \geq \text{tr}(\Sigma_0) / \|\Sigma_0\| \geq 1\) and the last inequality follows from the assumption that \(m \geq \beta\). Overall, term (V) is bounded as follows
\[
(V) \leq B_{\beta} \left(\frac{\sigma^2 \beta}{m}\right)^{1/4} + 3B_{\beta} \sqrt{\frac{\sigma^2 \beta}{\|\Sigma_0\| m}},
\]

with probability at least \(1 - de^{-\beta}\). Note that \(\mathbb{E}[Z_iZ_i^T] = \Sigma_0\), then, combining with (4.21), the term (II) is bounded as
\[
(\text{II}) \leq B_{\beta} \left(\|\Sigma_0\|^\frac{3}{4} + \sqrt{2} \sigma^\frac{1}{2} \left(\frac{\beta}{m}\right)^\frac{1}{4}\right) \left(2B_{\beta} \sqrt{\theta} + B_{\beta}^2 \theta\right) + B_{\beta} \left(\frac{\sigma^2}{m}\beta\right)^{1/4} + 3B_{\beta} \sqrt{\frac{\sigma^2 \beta}{\|\Sigma_0\| m}},
\]

137
with probability at least $1 - 2de^{-\beta}$. Substituting $B_\beta = 11\sqrt{\frac{2tr(\Sigma_0)\beta}{m}}$ and $\theta = \frac{1}{2}\sqrt{\frac{\beta}{m}}$ gives

$$
(\text{II}) \leq 11\sqrt{2} \sqrt{\text{tr}(\Sigma_0)\sigma} \left(\frac{\beta}{m}\right)^{\frac{3}{2}} + 33\sqrt{2} \frac{\sqrt{\text{tr}(\Sigma_0)\sigma}}{\|\Sigma_0\|^{1/2}} \frac{\beta}{m} + 484\|\Sigma_0\|^{1/2} \frac{\text{tr}(\Sigma_0)}{\sigma}\left(\frac{\beta}{m}\right)^{\frac{3}{2}}
$$

$$
+ 484\sqrt{2}\text{tr}(\Sigma_0) \left(\frac{\beta}{m}\right)^{\frac{3}{2}} + 2\sqrt{2 \cdot 11^3} \|\Sigma_0\|^{\frac{3}{2}} \frac{\text{tr}(\Sigma_0)}{\sigma}\left(\frac{\beta}{m}\right)^{\frac{3}{2}} + 4 \cdot 11^3 \frac{\text{tr}(\Sigma_0)^{3/2}}{\sigma^{1/2}} \left(\frac{\beta}{m}\right)^{9/4}.
$$

Using the bounds (4.18) and (4.19) with some algebraic manipulations, we have the second bound in Lemma 4.5.5 holds with probability at least $1 - 2de^{-\beta}$.

4.5.5 Proof of Lemma 4.5.6

We divide our analysis into the following four cases:

1. If $\|Z_i\|^2 \leq 1/\theta$ and $\|Z_i - \mu\|^2 \leq 1/\theta$, then, we have $h_\mu(Z_i) = 1$.

2. If $\|Z_i\|^2 \leq 1/\theta$ and $\|Z_i - \mu\|^2 > 1/\theta$. Since $\|\mu\| \leq B_\beta$, it follows $\|Z_i - \mu\|^2 \leq \sqrt{1/\theta} + B_\beta$, and we have

$$
h_\mu(Z_i) = \frac{1/\theta}{\|Z_i - \mu\|^2} \leq 1,
$$

$$
h_\mu(Z_i) \geq \left(\sqrt{1/\theta} + B_\beta\right)^2 = \frac{1}{1 + 2B_\beta \sqrt{\theta} + B_\beta^2 \theta}
$$

$$
\geq 1 - 2B_\beta \sqrt{\theta} - B_\beta^2 \theta,
$$

where the last inequality follows from the fact $\frac{1}{1 + x} \geq 1 - x$, $\forall x \geq 0$.

3. If $\|Z_i\|^2 > 1/\theta$ and $\|Z_i - \mu\|^2 \leq 1/\theta$. Since $\|\mu\| \leq B_\beta$, it follows $\|Z_i\|^2 \leq \sqrt{1/\theta} + B_\beta$, and we have

$$
h_\mu(Z_i) = \frac{\|Z_i\|^2}{1/\theta} \geq 1,
$$

$$
h_\mu(Z_i) \leq \left(\sqrt{1/\theta} + B_\beta\right)^2 \frac{1/\theta}{1/\theta} = 1 + 2B_\beta \sqrt{\theta} + B_\beta^2 \theta.
$$

138
4. If $\|Z_i\|_2^2 > 1/\theta$ and $\|Z_i - \mu\|_2^2 > 1/\theta$. Then, we have

\[
h_{\mu}(Z_i) = \frac{\|Z_i\|_2^2}{\|Z_i - \mu\|_2^2} \leq \frac{(\|Z_i - \mu\|_2 + B_\beta)^2}{\|Z_i - \mu\|_2^2}
\]

\[
\leq \left(\frac{1/\sqrt{\theta} + B_\beta}{1/\sqrt{\theta}}\right)^2 \leq 1 + 2B_\beta \sqrt{\theta} + B_\beta^2 \theta,
\]

\[
h_{\mu}(Z_i) \geq \frac{\|Z_i\|_2^2}{(\|Z_i\|_2 + B_\beta)^2} \geq \left(\frac{1/\sqrt{\theta}}{1/\sqrt{\theta} + B_\beta}\right)^2
\]

\[
= \frac{1}{1 + 2B_\beta \sqrt{\theta} + B_\beta^2 \theta} \geq 1 - 2B_\beta \sqrt{\theta} - B_\beta^2 \theta,
\]

Overall, we proved the lemma.

4.5.6 Proof of Lemma 4.2.2

By definition,

\[
B = \sup_{\|v\|_2 \leq 1} \mathbb{E}[|\langle v, X \rangle|^4] \geq \mathbb{E}[|X_j|^4], \quad \forall j = 1, 2, \cdots, d,
\]

where $X_j$ denotes the $j$-th entry of the random vector $X$. Also, for any fixed vector $v \in \mathbb{R}^d$, we have

\[
0 \leq \mathbb{E}\left[\left(|\langle v, X \rangle|^2 - |X_j|^2\right)^2\right] = \mathbb{E}[|\langle v, X \rangle|^4] + \mathbb{E}[|X_j|^4] - 2\mathbb{E}[|\langle v, X \rangle|^2 |X_j|^2]
\]

\[
\Rightarrow \mathbb{E}[|\langle v, X \rangle|^4] + \mathbb{E}[|X_j|^4] \geq 2\mathbb{E}[|\langle v, X \rangle|^2 |X_j|^2], \quad \forall j = 1, 2, \cdots, d.
\]

Taking the supremum from both sides of the above inequality and use the previous bound on $B$, we get

\[
\sup_{\|v\|_2 \leq 1} \mathbb{E}[|\langle v, X \rangle|^4] \geq \sup_{\|v\|_2 \leq 1} \mathbb{E}[|\langle v, X \rangle|^2 |X_j|^2], \quad \forall j = 1, 2, \cdots, d.
\]

Summing over $i = 1, 2, \cdots, d$ gives

\[
Bd = \sup_{\|v\|_2 \leq 1} \mathbb{E}[|\langle v, X \rangle|^4] d \geq \sup_{\|v\|_2 \leq 1} \mathbb{E}[|\langle v, X \rangle|^2 |X_j|^2] \geq \sup_{\|v\|_2 \leq 1} \mathbb{E}[|\langle v, X \rangle|^2 |X|^2]
\]

\[
= \|XX^T\|_2 \geq \sigma_0^2.
\]
4.5.7 Proof of Lemma 4.2.3

First of all, let \( Z = X - \mu_0 \), then, we have \( \mathbb{E}[Z] = 0 \). The lower bound of \( \sigma_0^2 \) follows directly from Corollary 4.5.1. It remains to show the upper bound. Note that by Cauchy-Schwarz inequality,

\[
\sigma_0^2 = \|ZZ^T\|_2^2 = \sup_{\|v\|_2 \leq 1} \mathbb{E}[\langle Z, v \rangle^2 \|Z\|_2^2] \\
\leq \sup_{\|v\|_2 \leq 1} \mathbb{E}[(Z, v)^4]^{1/2} \mathbb{E}[\|Z\|_2^4]^{1/2}.
\]

We then bound the two terms separately. For any vector \( x \in \mathbb{R}^d \), let \( x^j \) be the \( j \)-th entry. Note that for any \( v \in \mathbb{R}^d \) such that \( \|v\|_2 \leq 1 \), we have

\[
\mathbb{E}[(Z, v)^4]^{1/2} \leq R \cdot \mathbb{E}[(Z, v)^2] \leq R \sup_{\|v\|_2 \leq 1} \mathbb{E}[(Z, v)^2] \leq R \|\Sigma_0\|,
\]

where the first inequality uses the fact that the kurtosis is bounded.

Also, we have

\[
\mathbb{E}[\|Z\|_2^4]^{1/2} = \left( \sum_{j=1}^d \mathbb{E}[(Z_j)^4] + \sum_{j,k=1, j \neq k}^d \mathbb{E}[(Z_j)^2(Z_k)^2] \right)^{1/2} \\
\leq \left( \sum_{j=1}^d \mathbb{E}[(Z_j)^4] + \sum_{j,k=1, j \neq k}^d \mathbb{E}[(Z_j)^4]^{1/2} \mathbb{E}[(Z_k)^4]^{1/2} \right)^{1/2} \\
\leq \sum_{j=1}^d \sqrt{\mathbb{E}[(Z_j)^4]} \leq R \cdot \sum_{j=1}^d \mathbb{E}[(Z_j)^2] = R \cdot \text{tr}(\Sigma_0)
\]

Combining the above two bounds gives

\[
\sigma_0^2 \leq R^2 \|\Sigma_0\| \text{tr}(\Sigma_0),
\]

which implies the result.
Bibliography

[ABB00] O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–10106, 2000.

[ACM12] W. K. Allard, G. Chen, and M. Maggioni. Multi-scale geometric methods for data sets ii: Geometric multi-resolution analysis. *Applied and Computational Harmonic Analysis*, 32(3):435–462, 2012.

[ALPV14] A. Ai, A. Lapanowski, Y. Plan, and R. Vershynin. One-bit compressed sensing with non-Gaussian measurements. *Linear Algebra and its Applications*, 441:222–239, 2014.

[Bal97] K. Ball. *An elementary introduction to modern convex geometry*. Cambridge University Press, New York, 1997.

[BBM+05] P. L. Bartlett, O. Bousquet, S. Mendelson, et al. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[BCFS14] A. Banerjee, S. Chen, F. Fazayeli, and V. Sivakumar. Estimation with norm regularization. *Advances Neural Information Processing Systems (NIPS)* 27, 2014.

[Bha13] R. Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[BM02] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.
D. R. Brillinger. A generalized linear model with “Gaussian” regressor variables. In *A Festschrift for Erich L. Lehmann*, Wadsworth Statist./Probab. Ser., pages 97–114. Wadsworth, Belmont, CA, 1983.

L. D. Brown. Fundamentals of statistical exponential families: with applications in statistical decision theory. Ims, 1986.

P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

K. Balasubramanian and M. Yuan. Discussion of “Estimating structured high-dimensional covariance and precision matrices: optimal rates and adaptive estimation”. *Electronic Journal of Statistics*, 10(1):71–73, 2016.

E. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589–592, 2008.

O. Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*, volume 48, pages 1148–1185, 2012.

O. Catoni. PAC-Bayesian bounds for the Gram matrix and least squares regression with a random design. *arXiv preprint arXiv:1603.05229*, 2016.

M. Chen, C. Gao, and Z. Ren. Robust covariance matrix estimation via matrix depth. *arXiv preprint arXiv:1506.00691*, 2015.

S. Cambanis, S. Huang, and G. Simons. On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 11:368–385, 1981.

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.

E. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *arXiv preprint math/0409186*, 2004.
[CRT06] E. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

[CRZ16] T. T. Cai, Z. Ren, and H. H. Zhou. Estimating structured high-dimensional covariance and precision matrices: optimal rates and adaptive estimation. *Electron. J. Statist.*, 10(1):1–59, 2016.

[DDS+09] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[Dir13] S. Dirksen. Tail bounds via generic chaining. *arXiv preprint arXiv:1309.3522*, 2013.

[DK70] C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[DKK+16] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Robust estimators in high dimensions without the computational intractability. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 655–664. IEEE, 2016.

[Dur19] R. Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

[FK17] J. Fan and D. Kim. Robust high-dimensional volatility matrix estimation for high-frequency factor model. *Journal of the American Statistical Association*, 2017.

[FKN90] K.-T. Fang, S. Kotz, and K. W. Ng. *Symmetric multivariate and related distributions*. Chapman and Hall, 1990.

[FLL16] J. Fan, Y. Liao, and H. Liu. An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, 19(1):C1–C32, 2016.

[FLW17] J. Fan, Q. Li, and Y. Wang. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):247–265, 2017.
[FWZ16a] J. Fan, W. Wang, and Y. Zhong. An $\ell_\infty$ eigenvector perturbation bound and its application to robust covariance estimation. *arXiv preprint arXiv:1603.03516*, 2016.

[FWZ16b] J. Fan, W. Wang, and Z. Zhu. Robust low-rank matrix recovery. *arXiv:1603.08315*, 2016.

[FWZ16c] J. Fan, W. Wang, and Z. Zhu. Robust low-rank matrix recovery. *arXiv preprint arXiv:1603.08315*, 2016.

[FWZ17] J. Fan, W. Wang, and Z. Zhu. A Shrinkage Principle for Heavy-Tailed Data: High-Dimensional Robust Low-Rank Matrix Recovery. *arXiv preprint arXiv:1603.08315v2*, 2017.

[Gen16] M. Genzel. High-dimensional estimation of structured signals from non-linear observations with general convex loss functions. *arXiv preprint arXiv:1602.03436*, 2016.

[Giu15] I. Giulini. PAC-Bayesian bounds for Principal Component Analysis in Hilbert spaces. *arXiv preprint arXiv:1511.06263*, 2015.

[GMW16] L. Goldstein, S. Minsker, and X. Wei. Structured signal recovery from non-linear and heavy-tailed measurements. *arXiv preprint arXiv:1609.01025*, 2016.

[Gor88] Y. Gordon. On milman’s inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. In *Geometric Aspects of Functional Analysis*, pages 84–106. Springer, 1988.

[GW18] L. Goldstein and X. Wei. Non-gaussian observations in nonlinear compressed sensing via stein discrepancies. *Information and Inference: A Journal of the IMA*, 8(1):125–159, 2018.

[GW19] L. Goldstein and X. Wei. Non-gaussian observations in nonlinear compressed sensing via stein discrepancies. *Information and Inference: A Journal of the IMA*, 8(1):125–159, 2019.

[HHI+93] W. Hardle, P. Hall, H. Ichimura, et al. Optimal smoothing in single-index models. *The annals of Statistics*, 21(1):157–178, 1993.
[HJS01] M. Hristache, A. Juditsky, and V. Spokoiny. Direct estimation of the index coefficient in a single-index model. *Annals of Statistics*, pages 595–623, 2001.

[HL16] F. Han and H. Liu. Eca: High dimensional elliptical component analysis in non-gaussian distributions. *Journal of the American Statistical Association*, (just-accepted), 2016.

[Hot33] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[HRVA08] M. Hubert, P. J. Rousseeuw, and S. Van Aelst. High-breakdown robust multivariate methods. *Statistical Science*, pages 92–119, 2008.

[HTW15] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[KLT11] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

[KM15] V. Koltchinskii and S. Mendelson. Bounding the smallest singular value of a random matrix without concentration. *International Mathematics Research Notices*, 2015(23):12991–13008, 2015.

[LD89] K.-C. Li and N. Duan. Regression analysis under link violation. *The Annals of Statistics*, pages 1009–1052, 1989.

[Lep92] O. Lepski. Asymptotically minimax adaptive estimation. I: Upper bounds. optimally adaptive estimates. *Theory of Probability & Its Applications*, 36(4):682–697, 1992.

[LL17] G. Lecué and M. Lerasle. Robust machine learning by median-of-means: theory and practice. *arXiv preprint arXiv:1711.10306*, 2017.

[LM16] G. Lugosi and S. Mendelson. Risk minimization by median-of-means tournaments. *arXiv preprint arXiv:1608.00757*, 2016.
[LM17a] G. Lecué and S. Mendelson. Regularization and the small-ball method ii: complexity dependent error rates. *The Journal of Machine Learning Research*, 18(1):5356–5403, 2017.

[LM17b] G. Lecué and S. Mendelson. Sparse recovery under weak moment assumptions. *Journal of the European Mathematical Society*, 19(3):881–904, 2017.

[LMPV17] C. Liaw, A. Mehrabian, Y. Plan, and R. Vershynin. A simple tool for bounding the deviation of random matrices on geometric sets. In *Geometric aspects of functional analysis*, pages 277–299. Springer, 2017.

[Lou14] K. Lounici. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058, 2014.

[LT91] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer-Verlag, Berlin, 1991.

[LT13] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[LW04] O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

[LW+12] O. Ledoit, M. Wolf, et al. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024–1060, 2012.

[M+00] P. Massart et al. About the constants in talagrand’s concentration inequalities for empirical processes. *The Annals of Probability*, 28(2):863–884, 2000.

[Men14a] S. Mendelson. Learning without concentration. In *Conference on Learning Theory*, pages 25–39, 2014.

[Men14b] S. Mendelson. Upper bounds on product and multiplier empirical processes. *arXiv preprint arXiv:1410.8003*, 2014.

[Men16] S. Mendelson. Upper bounds on product and multiplier empirical processes. *Stochastic Processes and their Applications*, 126(12):3652–3680, 2016.
[Min15] S. Minsker. Geometric median and robust estimation in Banach spaces. Bernoulli, 21(4):2308–2335, 2015.

[Min16] S. Minsker. Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. arXiv preprint arXiv:1605.07129, 2016.

[MP12] S. Mendelson and G. Paouris. On generic chaining and the smallest singular value of random matrices with heavy tails. Journal of Functional Analysis, 262(9):3775–3811, 2012.

[MPTJ07] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Reconstruction and sub-gaussian operators in asymptotic geometric analysis. Geometric and Functional Analysis, 17(4):1248–1282, 2007.

[MS90a] S. J. Montgomery-Smith. The distribution of rademacher sums. Proceedings of the American Mathematical Society, 109(2):517–522, 1990.

[MS90b] S. J. Montgomery-Smith. The distribution of rademacher sums. In Proceedings of the AMS, pages 517–522, 109(2), 1990.

[MW+20] S. Minsker, X. Wei, et al. Robust modifications of u-statistics and applications to covariance estimation problems. Bernoulli, 26(1):694–727, 2020.

[NJB+08] J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann, M. R. Nelson, et al. Genes mirror geography within Europe. Nature, 456(7218):98–101, 2008.

[NRW+12] S. N. Negahban, P. Ravikumar, M. J. Wainwright, B. Yu, et al. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. Statistical Science, 27(4):538–557, 2012.

[NRWY12] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. Statistical Science, 27(4):538–557, 2012.

[Oli13] R. I. Oliveira. The lower tail of random quadratic forms, with applications to ordinary least squares and restricted eigenvalue properties. arXiv preprint arXiv:1312.2903, 2013.
[PV16] Y. Plan and R. Vershynin. The generalized Lasso with non-linear observations. *IEEE Transactions on Information Theory*, 62(3):1528–1537, 2016.

[PVY14] Y. Plan, R. Vershynin, and E. Yudovina. High-dimensional estimation with geometric constraints. *arXiv preprint arXiv:1404.3749*, 2014.

[PVY16] Y. Plan, R. Vershynin, and E. Yudovina. High-dimensional estimation with geometric constraints. *Information and Inference: A Journal of the IMA*, 6(1):1–40, 2016.

[PZ30] R. Paley and A. Zygmund. On some series of functions,(1). In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 26, pages 337–357. Cambridge University Press, 1930.

[RV08] M. Rudelson and R. Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 61(8):1025–1045, 2008.

[RWY11] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over $\ell_q$-balls. *IEEE transactions on information theory*, 57(10):6976–6994, 2011.

[SJH+07] L. H. Saal, P. Johansson, K. Holm, S. K. Gruvberger-Saal, Q.-B. She, M. Maurer, S. Koujak, A. A. Ferrando, P. Malmström, L. Memeo, et al. Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity. *Proceedings of the National Academy of Sciences*, 104(18):7564–7569, 2007.

[Sto86] T. M. Stoker. Consistent estimation of scaled coefficients. *Econometrica: Journal of the Econometric Society*, pages 1461–1481, 1986.

[SZF17] Q. Sun, W. Zhou, and J. Fan. Adaptive Huber Regression: Optimality and Phase Transition. *arXiv preprint arXiv:1706.06991*, 2017.

[TAH15] C. Thrampoulidis, E. Abbasi, and B. Hassibi. Lasso with non-linear measurements is equivalent to one with linear measurements. In *Advances in Neural Information Processing Systems*, pages 3420–3428, 2015.

148
[Tal95] M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l’Institut des Hautes Etudes Scientifiques*, 81(1):73–205, 1995.

[Tal14a] M. Talagrand. *Upper and lower bounds for stochastic processes: modern methods and classical problems*, volume 60. Springer Science & Business Media, 2014.

[Tal14b] M. Talagrand. *Upper and lower bounds for stochastic processes: modern methods and classical problems*. Ergebnisse der Mathematik und ihrer Grenzgebiete, Springer, 2014.

[Tib96] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[Tro12] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, 2012.

[Tro15a] J. A. Tropp. Convex recovery of a structured signal from independent random linear measurements. In *Sampling Theory, a Renaissance*, pages 67–101. Springer, 2015.

[Tro15b] J. A. Tropp. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.

[Tuk75] J. W. Tukey. Mathematics and the picturing of data. In *Proceedings of the international congress of mathematicians*, volume 2, pages 523–531, 1975.

[Tyl87] D. E. Tyler. A distribution-free M-estimator of multivariate scatter. *The Annals of Statistics*, pages 234–251, 1987.

[VDVW96a] A. W. Van Der Vaart and J. A. Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.

[vdVW96b] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

[Ver10a] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. C. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applications*. Cambridge University Press, 2010.
[Ver10b] R. Vershynin. Lectures in functional analysis. *Department of Mathematics, University of Michigan*, 2010.

[Ver15] R. Vershynin. Estimation in high dimensions: a geometric perspective. In *Sampling Theory, a Renaissance*, pages 3–66. Springer, 2015.

[W+13] J. Wellner et al. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 2013.

[Wei18] X. Wei. Structured recovery with heavy-tailed measurements: A thresholding procedure and optimal rates. *arXiv preprint arXiv:1804.05959*, 2018.

[WM17] X. Wei and S. Minsker. Estimation of the covariance structure of heavy-tailed distributions. In *Advances in Neural Information Processing Systems*, pages 2859–2868, 2017.

[WYG+09] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. PAMI*, 31(2):210–227, 2009.

[WZ+16] M. Wegkamp, Y. Zhao, et al. Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. *Bernoulli*, 22(2):1184–1226, 2016.

[YBL17] Z. Yang, K. Balasubramanian, and H. Liu. On stein’s identity and near-optimal estimation in high-dimensional index models. *arXiv preprint arXiv:1709.08795*, 2017.

[YBWL17] Z. Yang, K. Balasubramanian, Z. Wang, and H. Liu. Learning non-gaussian multi-index model via second-order stein’s method. *Advances in Neural Information Processing Systems*, 30:6097–6106, 2017.

[YWCL15] X. Yi, Z. Wang, C. Caramanis, and H. Liu. Optimal linear estimation under unknown nonlinear transform. In *Advances in Neural Information Processing Systems*, pages 1549–1557, 2015.

150