Optimal Mean Estimators

Thanh C. Nguyen

Advisors: **Prof. Linh Tran** **Prof. Hoi H. Nguyen**

Fulbright University Vietnam

A thesis submitted to partially fulfill the requirements for the degree of

Bachelor of Science in Applied Mathematics

April 2025

Contents

Abstract

"Probability is the science of measuring uncertainty with precision."
Andrey Kolmogorov

Mean estimation often fails in real-world settings due to heavy-tailed distributions or adversarial contamination, rendering classical methods like the empirical mean unreliable. This thesis investigates robust mean estimation through the lens of two optimal estimators: the median-of-means (MoM) and the trimmed mean presented in the works of Lugosi and Mendelson [LM19a; LM19b; LM21]. We aim to elucidate their construction, theoretical guarantees, computational challenges, and practical implications in both univariate and multivariate settings. The MoM estimator achieves sub-Gaussian performance under minimal assumptions, requiring only finite variance in the univariate case and finite second moments in the multivariate case. The trimmed mean extends this robustness by effectively handling adversarial contamination, maintaining sub-Gaussian bounds even with up to a fraction of corrupted data points. In general, sub-Gaussian performance is the best one could hope for in mean estimation, which is why we refer to these estimators as optimal.

We provide detailed proofs of their performance bounds, demonstrating their optimality in terms of error rates. Additionally, we analyze their computational complexity, highlighting the MoM's relative simplicity and the trimmed mean's exponential challenges in high dimensions. Through a comparative study, we contrast their strengths and limitations, showing that the MoM excels in clean, heavy-tailed scenarios, while the trimmed

mean is superior in noisy environments. This work underscores the significance of robust mean estimation in applications like finance and machine learning, offering insights into the trade-offs between statistical performance and computational feasibility, and suggesting directions for future research in efficient robust estimation techniques.

Foreword

"It always seems impossible until it's done."
Nelson Mandela

As noted on the cover page, this thesis is a partial fulfillment of the requirements for my Bachelor of Science in Applied Mathematics at Fulbright University Vietnam. When I first envisioned this project at the end of my third year, I had a different topic in mind—something in the realm of Random Matrix Theory, to be developed as my capstone project under the supervision of my intended advisor, Prof. Linh Tran, at Fulbright. At that time, I could not have imagined the journey that would lead me to where I am today.

Everything changed when summer arrived, and I was fortunate to participate in the Application Driven Mathematics program hosted by VinBigData. There, I had the incredible opportunity to study High-Dimensional Probability with Prof. Hoi H. Nguyen from The Ohio State University. His course opened my eyes to the beauty of probability in high dimensions, and he kindly guided me on a research project to introduce me to the world of academic research. That project, over time, evolved into this capstone project at Fulbright.

Our initial goal was ambitious: we aimed to develop a sub-Gaussian estimator for the covariance matrix, inspired by existing work on mean vector estimators, specifically the Median-of-Means estimator discussed by Lugosi in [LM19a]. We were aware, however, that this would be a significant challenge for someone like me, with limited research experience.

As it turned out, the task was indeed too difficult for me to produce original results. This thesis, in its final form, is not a presentation of new findings but rather a careful summary of a portion of the literature I explored during my research journey. None of the proofs within these pages are my own. Yet, I find the topic of sub-Gaussian estimators deeply fascinating—their results are elegant, powerful, and hold significant importance in statistics. I've worked to present these ideas in a clear and accessible way, especially for undergraduate students like myself, who might be new to this field. My hope is that this thesis, alongside the foundational works like those of Lugosi, can offer something meaningful to others starting their journey in this area.

The path to completing this capstone project, and the broader research on sub-Gaussian estimators for covariance matrices, has been an unforgettable experience for me—one filled with both challenges and growth. The early stages were particularly overwhelming. I've always wanted to pursue a PhD immediately after completing my undergraduate studies, but this ambition brought immense pressure. Even though I knew my research project was mainly for learning and gaining experience, I felt compelled to produce a novel result to support my goal of advancing to graduate school. I constantly worried that my work wouldn't measure up. At the same time, I was preparing my applications during the fall semester of my fourth year, a critical period for securing a spot in a U.S. program right after graduation. Balancing these demands made the initial phase of my research incredibly stressful. Toward the end of that semester, I was advised to pause my research to focus entirely on submitting my applications.

On February 14, I received the wonderful news that I had been accepted into the top program on my list. It was a moment of immense relief and joy, lifting a heavy weight off my shoulders. With a clearer mind, I returned to my research project, hoping that this newfound peace would pave the way for progress. Unfortunately, despite my renewed efforts, the results I hoped for didn't materialize, for reasons beyond my control. As I write this, I haven't achieved the breakthroughs I initially aimed for.

I'm not entirely sure how I feel about the outcome. There's a mix of disappointment and acceptance in my heart. But when I reflect on the journey itself—the late nights, the moments of doubt, the small victories of understanding a complex concept—I feel an overwhelming sense of gratitude. This experience has taught me resilience, patience, and the value of perseverance, lessons I know will stay with me as I move forward in my academic career. To those who read this thesis, I hope you find something valuable within its pages—whether it's a clearer understanding of sub-Gaussian estimators, an appreciation for their beauty, or simply a sense of shared experience as a student navigating the ups and downs of research. This journey, though not what I expected, has been a meaningful one, and I'm honored to share it with you.

Acknowledgements

"As iron sharpens iron, so one person sharpens another."
Proverb, 27:17

I've always considered myself a lucky person—perhaps even ridiculously lucky. Throughout my journey of growing up, I've met and been supported by so many wonderful people who have shaped me into who I am today. Without them, I can't imagine where I'd be. As I close an important chapter of my life with this thesis and prepare to begin a new adventure, I want to take this opportunity to express my heartfelt gratitude to them. This section might be a bit long, because, as I said, I've been truly fortunate to have so many people to thank.

First, I want to express my deepest gratitude to Prof. Linh Tran, my advisor at Fulbright University Vietnam, who has been so much more than an academic guide. He's supported me in many parts of life, and to me, he's the perfect role model of who I hope to become—professional, dedicated, caring, and always with a great sense of humor. Over the past four years, his guidance helped me grow and tackle challenges, especially in my final year. During that tough time, as I juggled preparing graduate applications and finishing this thesis, his patience, constant encouragement, and unshakable belief in me kept me going. I'm truly grateful to have had him by my side on this journey.

I also want to offer my sincere thanks from the bottom of my heart to Prof. Hoi H. Nguyen, who introduced me to the world of professional research. His passion during every lecture at the Application Driven Mathematics

program, his patience as he listened to my weekly progress reports, and his unwavering support whenever I needed guidance are memories I'll always cherish. I'm certain that my research experience would have been far more difficult without him as my advisor. I tend to overthink and get stuck in my worries, which often makes things harder for me. But Prof. Hoi's kindness and patience gave me the strength to push through those tough moments. I feel incredibly lucky to have been his student.

I'm grateful to all the friends who have been by my side, especially Triệu Phước, Trọng Toàn, Trung Nguyên, and Quốc Hưng. In particular, I want to thank my friends at Fulbright—Hoàng Ân, Nhật Tân, Hà Huy, Lan Phương, Minh Quân, Gia Khang, Đăng Thức, Kim Ngân, Nhất Phương, Hồng Hà, and Tuấn Lĩnh—for being such amazing companions. They've not only been wonderful friends but have also helped me become a better person, which, in some ways, I think is even more valuable than academic knowledge for a student's experience.

I also want to thank all the teachers who have supported me throughout my academic journey. Meeting them and learning from their guidance has been a significant milestone in my life.

I'm thankful for Hoàng Trang, who has been my emotional rock, standing by me through the hardest, happiest, and even the most ordinary days.

Finally, I want to express my gratitude to my family, who have given me the greatest support and unconditional encouragement throughout nearly 23 years of my life.

Chapter 1

Introduction

1.1 Problem Definition

Mean estimation plays a fundamental role in statistical analysis, supporting a wide range of applications from scientific research to machine learning. The problem involves estimating the expected value $\mu = \mathbb{E}X$ of a random variable X based on n independent, identically distributed (i.i.d.) samples $X_1, X_2, \ldots, X_n$.

Let $\widehat{\mu}_n = \widehat{\mu}_n(X_1, \ldots, X_n)$ be our estimator. We prefer $\widehat{\mu}_n$ that are close to μ *with high probability.*

Formally, our aim is to understand, for any given sample size n and confidence parameter $\delta \in (0,1)$, the *smallest possible value* $\varepsilon = \varepsilon(n, \delta) > 0$ such that

$$\mathbb{P}\{|\widehat{\mu}_n - \mu| \geqslant \varepsilon\} \leqslant \delta$$

in one dimension, or

$$\mathbb{P}\{\|\widehat{\mu}_n - \mu\| \geqslant \varepsilon\} \leqslant \delta$$

in higher dimensions, where $\|\cdot\|$ denote the Euclidean norm. For the sake of convenience, we sometimes also state the above property as: with the probability of at least $1 - \delta$, we have the following inequality

$$|\widehat{\mu}_n - \mu| \leqslant \varepsilon, \quad \text{or} \quad \|\widehat{\mu}_n - \mu\| \leqslant \varepsilon.$$

1.2 Motivation and Significance

Mean estimation is a fundamental task in statistics, yet real-world applications often present significant challenges that render traditional methods ineffective. Consider estimating the average income in a population where a few individuals have extremely high incomes: these outliers can drastically skew the empirical mean, $\frac{1}{n}\sum_{i=1}^{n} X_i$, leading to an inaccurate estimate of the typical income. This issue arises because real-world data frequently follow *heavy-tailed distributions*, where extreme values occur more often than in a normal distribution, as seen in domains like finance (e.g., asset returns) and biological measurements (e.g., gene expression data). In machine learning, such outliers can distort model training, for example, when processing user-generated content with erroneous entries. The need for robust mean estimation methods that perform reliably under these conditions is thus critical, a challenge addressed by the estimators studied in this thesis [LM19b; LM19a].

The empirical mean's poor performance in such scenarios stems from its sensitivity to outliers. As noted by Lugosi and Mendelson [LM19a], for a univariate random variable with finite variance σ^2, the empirical mean's error is bounded by $\sigma\sqrt{\frac{1}{n\delta}}$ with probability at least $1-\delta$, according to Chebyshev's inequality. This error rate scales poorly with the confidence parameter δ, making the estimator impractical for heavy-tailed datasets. For instance, to achieve a confidence of $1-\delta = 0.95$ with $n = 1000$ samples, the error could be as large as 0.14σ, which is substantial for datasets with outliers.

In contrast, *sub-Gaussian estimators* achieve a much tighter error bound in the univariate case. For the same example with $n = 1000$ and $\delta = 0.05$, if we consider an error bound of the form $\varepsilon = L\sigma\sqrt{\frac{\log(2/\delta)}{n}}$ with a constant $L = 1$ for simplicity, the error is approximately 0.061σ, a significant improvement over the 0.14σ of the empirical mean. In the multivariate case, the error depends on the covariance matrix Σ, balancing the trace $\mathrm{Tr}(\Sigma)$ (total variance) and the largest eigenvalue $\lambda_{\max}$ (worst-case directional variance). These sub-Gaussian rates are optimal, as no estimator

can achieve better performance even for Gaussian data [LM19a, Theorem 1], making them a powerful solution for heavy-tailed distributions.

Beyond heavy-tailed data, real-world datasets often face an additional challenge: *adversarial contamination*, where up to a fraction η of the samples may be arbitrarily altered due to measurement errors or malicious tampering [LM21]. In engineering, for example, sensor data in autonomous driving systems may include faulty readings that corrupt the sample, while in finance, fraudulent transactions can skew financial datasets. The empirical mean is particularly vulnerable to such contamination, as even a single corrupted value can arbitrarily distort the estimate, leading to unreliable results in critical applications.

The survey by Lugosi and Mendelson [LM19a] emphasizes how recent advances, fueled by big data and statistical learning, have revitalized the field of robust statistics, building on classical foundations laid by researchers like Huber [Hub64]. By achieving sub-Gaussian error rates under minimal assumptions and addressing adversarial contamination, the methods explored in this thesis—developed in recent works by Lugosi and Mendelson [LM19b; LM19a; LM21]—offer practical solutions to these pressing challenges, making them highly relevant to contemporary applications.

1.3 Overview of the Estimators

This thesis examines two robust mean estimators—the median-of-means tournament estimator and the trimmed mean estimator—that directly address the challenges of heavy-tailed distributions and adversarial contamination, as developed in three key papers [LM19b; LM19a; LM21]. The *median-of-means tournament estimator*, introduced by Lugosi and Mendelson [LM19b], tackles the issue of heavy-tailed distributions by partitioning the sample into blocks and computing the mean of each block, effectively reducing the influence of outliers. In the univariate case, it selects the standard median of these block means, achieving the sub-Gaussian error rate of $\sigma\sqrt{\frac{\log(1/\delta)}{n}}$ with only finite variance required

[LM19a, Section 2.1]. In the multivariate case, it defines a novel median concept by choosing a point that minimizes the radius of a set defined by the block means, ensuring sub-Gaussianity in high-dimensional spaces with an error dependent on $\mathrm{Tr}(\Sigma)$ and $\lambda_{\max}$ [LM19b]. This approach resembles a tournament where block means compete to find a winner close to the true mean, mitigating the impact of extreme values in heavy-tailed data.

The *trimmed mean estimator*, detailed by Lugosi and Mendelson [LM21], addresses both heavy-tailed distributions and adversarial contamination by discarding extreme values from the sample. In the univariate case, it splits the data into two parts: one estimates truncation levels (quantiles), and the other computes the average of values within these bounds, achieving sub-Gaussian error rates even when up to a fraction η of the samples are corrupted [LM19a; LM21, Section 2.3]. In the multivariate case, it applies trimming to projections of the data onto all directions, intersecting the resulting constraints to estimate the mean, maintaining robustness to both outliers and contamination [LM21]. This method is akin to pruning a tree, cutting off extreme branches—whether due to heavy tails or corrupted samples—to reveal the core structure. While both estimators achieve sub-Gaussian accuracy under minimal assumptions, the trimmed mean's ability to handle adversarial contamination makes it particularly versatile for noisy datasets, such as those encountered in engineering and finance applications.

1.4 Thesis Objectives and Structure

The objective of this thesis is to provide a comprehensive analysis of the median-of-means (MoM) and trimmed mean estimators, summarizing their methods, mathematical foundations, performance guarantees, and computational aspects as developed by Lugosi and Mendelson [LM19b; LM19a; LM21]. We aim to elucidate their construction and proofs in both univariate and multivariate settings, offering insights into their robustness, optimality, and practical trade-offs. The thesis balances accessibility, through intuitive explanations and numerical examples, with

rigor, via detailed proofs and technical discussions, making it suitable for an undergraduate mathematics audience.

The thesis is organized as follows

- **Chapter 1: Introduction** provides the background and motivation for robust mean estimation, introduces the median-of-means and trimmed mean estimators, and outlines the thesis structure, setting the stage for the technical content that follows.
- **Chapter 2: Background and Theoretical Foundations** introduces key concepts, including estimators, sub-Gaussianity, concentration inequalities (e.g., Chebyshev's, Bernstein's, Talagrand's), and geometric tools (e.g., dual Sudakov inequality), providing the groundwork for understanding the estimators. It also demonstrates that sub-Gaussian estimators generally offer the best expected performance under minimal assumptions.
- **Chapter 3: Median-of-Means Estimator** details the MoM estimator's method and proofs in the univariate and multivariate cases, highlighting its sub-Gaussian performance under finite variance or second moments.
- **Chapter 4: Trimmed Mean Estimator** presents the trimmed mean estimator's construction, performance bounds, and proofs for both settings, emphasizing its robustness to adversarial contamination while maintaining sub-Gaussian performance.
- **Chapter 5: Computational Considerations and Comparison of Estimators** analyzes the theoretical computational complexity of both estimators, highlighting challenges in high dimensions, and compares their performance, robustness, and applicability, contrasting their strengths and limitations in various scenarios.
- **Chapter 6: Conclusion** summarizes the findings, reflects on the significance of robust mean estimation in practical applications, and suggests directions for future research, including computational improvements and extensions to other contamination models.

Chapter 2

Background and Theoretical Foundations

To understand the robust mean estimation methods presented in this thesis—the median-of-means tournament estimator and the trimmed mean estimator—it is essential to establish the mathematical foundations that underpin their development and analysis. These estimators, as explored in Lugosi and Mendelson [LM19b; LM19a; LM21], rely on a variety of statistical and probabilistic tools to achieve sub-Gaussian performance under minimal assumptions, addressing challenges such as heavy-tailed distributions and adversarial contamination. This chapter introduces the key concepts and techniques required to follow the estimators' construction, performance guarantees, and proofs, ensuring a solid groundwork for the subsequent chapters.

We begin by introducing concentration inequalities, which are fundamental tools for bounding the deviation of a random variable from its expected value, ensuring that estimators are reliable even in the presence of heavy-tailed data or contamination. Next, we cover geometric and empirical process tools, which are essential for handling high-dimensional data in multivariate settings. We then define estimators and their performance metrics, focusing on high-probability error bounds that are central to robust estimation. Finally, we explore the concept of sub-Gaussian estimators, which achieve optimal error rates even for heavy-tailed data, and

discuss their significance in the context of this thesis, setting the stage for a detailed examination of the estimators in the chapters that follow.

2.1 Concentration Inequalities

Concentration inequalities are probabilistic tools that bound the deviation of a random variable from its expected value, playing a crucial role in the proofs of the estimators' performance in Lugosi and Mendelson [LM19b; LM19a; LM21]. These inequalities allow us to ensure that estimators are close to the true mean with high probability, even in the presence of heavy-tailed data or contamination. Below, we introduce the key concentration inequalities used in the papers, providing intuitive explanations and their mathematical formulations. Note that we will not present the proofs of these inequalities here to maintain focus on their applications; detailed derivations can be found in Boucheron, Lugosi, and Massart [BLM13] or Vershynin [Ver18].

Theorem 1 (Markov's Inequality)**.** *For a non-negative random variable* $Y \geqslant 0$*, Markov's inequality states:*

$$\mathbb{P}\{Y \geqslant t\} \leqslant \frac{\mathbb{E}Y}{t},$$

for any $t > 0$ *[Che67].*

Markov's inequality is a simple yet powerful tool that provides an upper bound on the probability of a non-negative random variable exceeding a certain threshold, based solely on its expected value. It is often used as a building block for deriving other concentration inequalities, such as Chebyshev's inequality, and will be crucial in our analysis of tail bounds for sub-Gaussian distributions.

Theorem 2 (Chebyshev's Inequality)**.** *For a random variable* Y *with mean* $\mathbb{E}[Y]$ *and variance* $\mathrm{Var}(Y)$*, Chebyshev's inequality states:*

$$\mathbb{P}\{|Y - \mathbb{E}Y| \geqslant t\} \leqslant \frac{\mathrm{Var}(Y)}{t^2},$$

for any $t > 0$ *[Che67].*

In the context of mean estimation, Lugosi and Mendelson [LM19a] uses Chebyshev's inequality to bound the deviation of block means in the median-of-means estimator. For a block mean Z_j based on m samples with variance σ^2, the variance is $\operatorname{Var}(Z_j) = \frac{\sigma^2}{m}$, so $\mathbb{P}\{|Z_j - \mu| \geqslant t\} \leqslant \frac{\sigma^2}{mt^2}$. This ensures that most block means are close to the true mean, a key step in achieving sub-Gaussian rates.

Theorem 3 (Hoeffding's Inequality)**.** *For independent random variables $Y_1, \ldots, Y_n$ with $Y_i \in [a_i, b_i]$, and $S = \sum_{i=1}^n Y_i$, Hoeffding's inequality states:*

$$\mathbb{P}\{|S - \mathbb{E}[S]| \geqslant t\} \leqslant 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right),$$

for any $t > 0$ [Hoe63; BLM13].

Lugosi and Mendelson [LM19a] uses Hoeffding's inequality in the median-of-means estimator to bound the probability that many block means deviate significantly, treating the number of deviating blocks as a binomial random variable. This exponential tail bound is stronger than Chebyshev's, ensuring high-probability guarantees with fewer samples.

Theorem 4 (Bernstein's Inequality)**.** *For independent random variables $Y_1, \ldots, Y_n$ with $|Y_i - \mathbb{E}[Y_i]| \leqslant M$, variance $\operatorname{Var}(Y_i) \leqslant \sigma^2$, and $S = \sum_{i=1}^n (Y_i - \mathbb{E}[Y_i])$, Bernstein's inequality states:*

$$\mathbb{P}\{|S| \geqslant t\} \leqslant 2\exp\left(-\frac{t^2}{2(n\sigma^2 + Mt/3)}\right),$$

for any $t > 0$ [Ber24; BLM13].

Corollary 5. *Let $X_1, \ldots, X_n$ be i.i.d. random variables, and let $v_i = X_i - \mathbb{E}[X_i] \in \mathbb{R}$. Let $\sigma^2 = \operatorname{Var}(X_i)$. Then, with probability $1 - \delta$,*

$$\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| \leqslant \sqrt{\frac{2\sigma^2 \ln\frac{2}{\delta}}{n}} + \frac{4R\ln\frac{2}{\delta}}{3n},$$

where R is such that $|v_i| \leqslant R$ almost surely [Ber24].

Lugosi and Mendelson [LM21] applies Bernstein's inequality to the trimmed mean estimator, where the trimmed values are bounded (e.g., within quantiles), ensuring concentration around the mean even with heavy-tailed data. Corollary 5 provides a specific bound for the sample mean, useful in sub-Gaussian contexts.

Theorem 6 (Talagrand's Inequality)**.** *Let X_i, $i = 1, \ldots, n$, be independent $\mathcal{X}$-valued random variables. Let $\mathcal{F}$ be a countable family of measurable real-valued functions on $\mathcal{X}$ such that $\|f\|_\infty \leqslant U < \infty$ and $\mathbb{E}[f(X_1)] = \cdots = \mathbb{E}[f(X_n)] = 0$, for all $f \in \mathcal{F}$. Let*

$$Z := \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} f(X_i) \quad \text{or} \quad Z = \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} f(X_i) \right|,$$

and let the parameters σ^2 and ν_n be defined as

$$U^2 \geqslant \sigma^2 \geqslant \frac{1}{n} \sum_{i=1}^{n} \sup_{f \in \mathcal{F}} \mathbb{E}\left[f^2(X_i)\right] \quad \text{and} \quad \nu_n := 2U\mathbb{E}[Z] + n\sigma^2.$$

Then, for all $t \geqslant 0$,

$$\mathbb{P}\{Z \geqslant \mathbb{E}[Z] + t\} \leqslant \exp\left(-\frac{t^2}{2\nu_n + 2tU/3}\right),$$

and

$$\mathbb{P}\left\{Z \geqslant \mathbb{E}[Z] + \sqrt{2\nu_n x} + Ux/3\right\} \leqslant e^{-x}, \quad x \geqslant 0.$$

[Tal96; BLM13].

Lugosi and Mendelson [LM21] apply Talagrand's inequality to control uniform deviations across directions in $\mathbb{R}^d$, ensuring that the estimators' errors are bounded consistently over all projections. In the multivariate trimmed mean estimator, Talagrand's inequality helps bound the deviation of the trimmed projections over all unit vectors in the sphere S^{d-1}, a critical step in achieving sub-Gaussian performance.

Theorem 7 (Gaussian Concentration Inequality of Tsirelson, Ibragimov, and Sudakov)**.** *Let $X \in \mathbb{R}^d$ be a Gaussian random vector with mean 0 and covariance matrix Σ. If $f : \mathbb{R}^d \to \mathbb{R}$ is a Lipschitz function with*

Lipschitz constant L (i.e., $|f(x) - f(y)| \leqslant L \|x - y\|$ for all $x, y \in \mathbb{R}^d$), then for any $t \geqslant 0$,

$$\mathbb{P}\left\{|f(X) - \mathbb{E}\left[f(X)\right]| \geqslant t\right\} \leqslant 2\exp\left(-\frac{t^2}{2L^2\lambda_{\max}}\right),$$

where $\lambda_{\max} = \lambda_{\max}(\Sigma)$ is the largest eigenvalue of Σ [TIS76].

This inequality is particularly useful for functions of Gaussian random vectors, such as the Euclidean norm of a centered Gaussian vector, which is a Lipschitz function. It will be applied in the analysis of sub-Gaussian estimators in the multivariate case to control the deviation of norms.

In the context of robust mean estimation, Lugosi and Mendelson [LM21] leverages the Bounded Differences Inequality to analyze functions of independent random variables where the impact of any single variable is limited, for example, the sum of indicator functions.

Definition 8 (Bounded Differences Property)**.** *A function $f : \mathcal{X}^n \to \mathbb{R}$ defined on a product space $\mathcal{X}^n$ satisfies the* bounded differences property *with constants $c_1, \ldots, c_n \geqslant 0$ if for all $i \in \{1, \ldots, n\}$, and for all $x_1, \ldots, x_n, x_i' \in \mathcal{X}$,*

$$\sup_{x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n} |f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leqslant c_i.$$

This means that changing the i-th coordinate of the input changes the value of f by at most c_i, regardless of the values of the other coordinates. [BLM13]

Theorem 9 (Bounded Differences Inequality)**.** *Assume that the function f satisfies the bounded differences assumption with constants $c_1, \ldots, c_n$, and denote*

$$v = \frac{1}{4}\sum_{i=1}^{n} c_i^2.$$

Let $Z = f(X_1, \ldots, X_n)$, where the X_i are independent. Then

$$\mathbb{P}\{Z - \mathbb{E}Z > t\} \leqslant e^{-t^2/(2v)}.$$

[BLM13]

These inequalities collectively ensure the robust performance of the estimators by tightly controlling deviations—whether of block means, trimmed averages, or uniform bounds—under minimal assumptions. Specifically, Chebyshev's and Hoeffding's inequalities bound the deviations of sums of random variables, Bernstein's inequality (and its corollary) sharpens these bounds for bounded differences, Talagrand's inequality ensures uniformity in empirical processes, Tsirelson's inequality manages Gaussian-specific deviations, and the Bounded Differences Inequality limits the impact of individual observations, all supporting the sub-Gaussian properties outlined in the next section.

2.2 Geometric and Empirical Process Tools

To estimate the mean of a random vector in $\mathbb{R}^d$, we need to control the estimator's error uniformly across all directions, a challenge in high-dimensional spaces. The papers by Lugosi and Mendelson [LM19b] and Lugosi and Mendelson [LM21] use geometric and empirical process tools to achieve sub-Gaussian performance, even for heavy-tailed distributions. This section introduces key concepts—convex bodies, covering numbers, ε-maximal separated sets, the dual Sudakov inequality, symmetrization inequalities, and the contraction lemma—which enable uniform control over directions. We provide formal definitions, intuitive explanations for an undergraduate audience, and their application to multivariate mean estimation, drawing on [Ver20] and the appendices of Lugosi and Mendelson [LM19b].

We begin with geometric objects that shape high-dimensional analysis.

Definition 10 (Convex Body)**.** *A set $K \subset \mathbb{R}^d$ is a* convex body *if it is compact, convex, and has non-empty interior. That is, for any $x, y \in K$ and $\lambda \in [0,1]$, the point $\lambda x + (1-\lambda)y \in K$, and there exists a ball of positive radius contained in K. [Ver20]*

Think of a convex body as a solid, rounded shape like a ball or ellipsoid. In our thesis, convex bodies define norms used to measure distances in

high-dimensional spaces.

Definition 11 (Norm Induced by a Convex Body)**.** *For a convex body $K \subset \mathbb{R}^d$ that is symmetric about the origin (i.e., $K = -K$) and has non-empty interior, the norm $\|x\|_K$ is defined as $\|x\|_K = \inf\{t > 0 : x \in tK\}$ for $x \in \mathbb{R}^d$. This norm measures how much K must be scaled to include x. [Ver20]*

In this thesis, we are particularly interested in the $L_2(X)$-norm, which arises naturally when studying the geometry of a random vector $X \in \mathbb{R}^d$ with mean μ and covariance matrix $\Sigma = \mathbb{E}\left[(X-\mu)(X-\mu)^T\right]$. The $L_2(X)$-norm of a vector $v \in \mathbb{R}^d$ is defined as

$$\|v\|_{L_2(X)} = \sqrt{\mathbb{E}\left[\langle X-\mu, v\rangle^2\right]} = \sqrt{\langle v, \Sigma v\rangle},$$

where $\langle v, \Sigma v\rangle = v^T \Sigma v$. The associated convex body is the unit ball in this norm

$$K = \left\{v \in \mathbb{R}^d : \|v\|_{L_2(X)} \leqslant 1\right\} = \left\{v \in \mathbb{R}^d : \langle v, \Sigma v\rangle \leqslant 1\right\},$$

which is an ellipsoid centered at the origin, symmetric because Σ is positive semi-definite. This $L_2(X)$-norm measures the variability of the projection of $X - \mu$ onto the direction v, and it is the norm we use to define distances on the sphere rS^{d-1} in Section 3.3.

To approximate complex sets like the sphere $rS^{d-1} = \left\{v \in \mathbb{R}^d : \|v\| = r\right\}$, we use finite sets of points, quantified by covering numbers and maximal separated sets.

Definition 12 (Covering Number)**.** *For a set $A \subset \mathbb{R}^d$ and a norm $\|\cdot\|$, the* covering number *$N(A, \|\cdot\|, \varepsilon)$ is the smallest number of balls of radius ε in the norm $\|\cdot\|$ needed to cover A. Formally, it is the smallest integer N such that there exist points $a_1, \ldots, a_N \in \mathbb{R}^d$ with $A \subseteq \bigcup_{i=1}^N \{x : \|x - a_i\| \leqslant \varepsilon\}$. [Ver20]*

Imagine covering a sphere with small patches of radius ε. The covering number counts how many patches you need. In our proofs later, this helps discretize directions to bound the estimator's error.

Definition 13 (ε-Maximal Separated Set)**.** *For a set $A \subset \mathbb{R}^d$ and a norm $\|\cdot\|$, a subset $S \subset A$ is an* ε-maximal separated set *if*

- *For all distinct $x, y \in S$, $\|x - y\| \geqslant \varepsilon$.*
- *For any $z \in A \setminus S$, there exists $x \in S$ such that $\|z - x\| < \varepsilon$.*

The set S is a collection of points in A at least ε apart, and no point in A is farther than ε from some point in S. [Ver20]

Picture S as a net of points on a sphere, spread out so each pair is at least ε apart, yet close enough to cover the entire sphere. In our thesis, such sets approximate the sphere rS^{d-1} to control the supremum of $\langle X - \mu, v\rangle$ over all directions v.

The dual Sudakov inequality bounds the size of covering numbers or maximal separated sets, crucial for high-dimensional analysis.

Theorem 14 (Dual Sudakov Inequality)**.** *Let $K \subset \mathbb{R}^d$ be a convex body, and let $N(B^d, \|\cdot\|_K, \varepsilon)$ be the covering number of the Euclidean unit ball $B^d = \left\{x \in \mathbb{R}^d : \|x\| \leqslant 1\right\}$ in the norm $\|\cdot\|_K$. The dual Sudakov inequality states*

$$\sqrt{\log N(B^d, \|\cdot\|_K, \varepsilon)} \leqslant \frac{1}{\sqrt{32}} \frac{\mathbb{E}\left[\|G\|_K\right]}{\varepsilon},$$

where $G \sim \mathcal{N}(0, I_d)$ is a standard Gaussian vector. Equivalently, for an ε-maximal separated set $S \subset rS^{d-1} = \left\{v \in \mathbb{R}^d : \|v\| = r\right\}$ in a norm $\|\cdot\|$, the number of points $|S|$ satisfies

$$\log(|S|/2) \leqslant \frac{1}{32} \left(\frac{\mathbb{E}\left[\|G\|\right]}{\varepsilon/r}\right)^2.$$

In the context of the $L_2(X)$-norm, where $\|v\|_{L_2(X)} = \sqrt{\langle v, \Sigma v\rangle}$, this becomes

$$\log(|S|/2) \leqslant \frac{1}{32} \left(\frac{\mathbb{E}\left[\sqrt{\langle G, \Sigma G\rangle}\right]}{\varepsilon/r}\right)^2.$$

[Ver20]

This theorem limits how many points can fit in a maximal separated set on a sphere or how many balls are needed to cover a set. In Lugosi and Mendelson [LM19b], it bounds the number of points in a net on rS^{d-1}, enabling uniform error bounds for the multivariate median-of-means estimator (Section 3.3).

We also need to use two familiar tools in the context of bounding the empirical process, which are the symmetrization inequalities and contraction lemma. They involve *Rademacher random variables*, which simplify the analysis of empirical processes.

Definition 15 (Rademacher Random Variable)**.** *A* Rademacher random variable σ *takes values* $+1$ *or* -1 *with equal probability, i.e.,* $\mathbb{P}\{\sigma = 1\} = \mathbb{P}\{\sigma = -1\} = 1/2$.

Theorem 16 (Symmetrization Inequalities)**.** *Let* $X_1, \ldots, X_n$ *be i.i.d. random vectors in* $\mathbb{R}^d$*, and let* $\mathcal{F}$ *be a class of real-valued functions defined on* $\mathbb{R}^d$*. Let* $\sigma_1, \ldots, \sigma_n$ *be independent Rademacher random variables, independent of the* X_i*. The symmetrization inequalities state*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - \mathbb{E} f(X_i)) \leqslant 2\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i),$$

and if $\mathbb{E} f(X_i) = 0$ *for all* $f \in \mathcal{F}$*, then*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i) \leqslant 2\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

[Ver20]

Symmetrization inequalities relate the deviation of an empirical process to a simpler process involving Rademacher random variables, which are easier to analyze. Lugosi and Mendelson [LM19b] uses these in the multivariate median-of-means tournament estimator to control uniform deviations over directions, ensuring the estimator's error is sub-Gaussian across all projections.

The contraction lemma involves *Lipschitz functions*, which constrain the rate of change of a function.

Definition 17 (Lipschitz Function)**.** *A function* $\phi : \mathbb{R} \to \mathbb{R}$ *is* Lipschitz *with constant* L *if for all* $x, y \in \mathbb{R}$*, it satisfies*

$$|\phi(x) - \phi(y)| \leqslant L\,|x - y|\,.$$

The constant L *is called the* Lipschitz constant *of* ϕ*.*

Theorem 18 (Contraction Lemma)**.** *Let* $X_1, \ldots, X_n$ *be i.i.d. random vectors in* $\mathbb{R}^d$*, and let* $\mathcal{F}$ *be a class of real-valued functions defined on* $\mathbb{R}^d$*. Let* $\sigma_1, \ldots, \sigma_n$ *be independent Rademacher random variables, independent of the* X_i*. If* $\phi : \mathbb{R} \to \mathbb{R}$ *is a function with* $\phi(0) = 0$ *and Lipschitz constant* L*, then*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \phi(f(X_i)) \leqslant L \cdot \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i).$$

[Ver20]

The contraction lemma bounds the expected supremum of a transformed empirical process, which is useful for functions that modify the original data (e.g., in the multivariate trimmed mean). Lugosi and Mendelson [LM19b] applies this to control the effect of Lipschitz transformations in their proofs, ensuring manageable uniform bounds.

These tools collectively enable the analysis of high-dimensional data by providing uniform control over all directions, a necessity for the sub-Gaussian performance of multivariate estimators.

2.3 Estimators and Performance Metrics

An *estimator* is a function that uses observed data to approximate an unknown parameter of a distribution. In the context of mean estimation, we consider the following setting.

Definition 19 (Mean Estimator)**.** *Given* n *i.i.d. samples* $X_1, X_2, \ldots, X_n$ *from a random variable* X *with mean* $\mu = \mathbb{E}X$*, a* mean estimator *is a function* $\widehat{\mu}_n = \widehat{\mu}_n(X_1, X_2, \ldots, X_n)$ *that aims to approximate* μ*.*

In the univariate case ($X \in \mathbb{R}$), the error is measured by the absolute difference $|\widehat{\mu}_n - \mu|$, while in the multivariate case ($X \in \mathbb{R}^d$), it is measured by the Euclidean norm $\|\widehat{\mu}_n - \mu\|$.

The performance of an estimator is typically evaluated by measuring its error. Classical statistical literature often uses the *mean squared error* as a metric.

Definition 20 (Mean Squared Error)**.** *The* mean squared error *of an estimator $\widehat{\mu}_n$ for the mean μ is defined as $\mathbb{E}\left[(\widehat{\mu}_n - \mu)^2\right]$.*

Example 21. For the empirical mean $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ in the univariate case, the mean squared error is $\mathbb{E}\left[(\widehat{\mu}_n - \mu)^2\right] = \frac{\sigma^2}{n}$, where $\sigma^2 = \mathrm{Var}(X)$.

However, the mean squared error can be misleading for heavy-tailed distributions, where the error may not be well-concentrated around the mean, leading to unreliable estimates. As highlighted in Lugosi and Mendelson [LM19a], this metric does not reflect the *typical* behavior of the error in such cases, motivating a focus on high-probability bounds. In robust mean estimation, we prioritize estimators that are close to the true mean with high probability.

Definition 22 (High-Probability Bound)**.** *An estimator $\widehat{\mu}_n$ satisfies a* high-probability bound *if, for a given sample size n and confidence parameter $\delta \in (0,1)$, there exists a small error $\varepsilon = \varepsilon(n,\delta)$ such that*

$$\mathbb{P}\left\{|\widehat{\mu}_n - \mu| > \varepsilon\right\} \leqslant \delta$$

in the univariate case, or

$$\mathbb{P}\left\{\|\widehat{\mu}_n - \mu\| > \varepsilon\right\} \leqslant \delta$$

in the multivariate case.

The goal is to minimize ε for a given n and δ, ensuring that the estimator is close to the true mean with high probability.

It is important to stress that this high-probability bound is a *non-asymptotic criterion*, meaning it provides guarantees for finite sample sizes n, rather than relying on asymptotic behavior as $n \to \infty$. This type of estimate is reminiscent of the Probably Approximately Correct (PAC) framework, commonly adopted in statistical learning theory [Val79; VC82; Blu+89]. The PAC framework focuses on obtaining quantitative bounds on the error that hold with high probability, which is particularly useful for robust estimation in the presence of heavy-tailed distributions and contaminated samples, as we will explore in the estimators' analyses.

2.4 Sub-Gaussian Estimators

Before diving into the definition of sub-Gaussian estimators, we need to understand why they are called "sub-Gaussian" and what makes them special for robust mean estimation. This involves looking at the behavior of probability distributions, specifically their *tail behavior*, and introducing the concept of a *sub-Gaussian distribution*, which will help us appreciate the motivation behind using these estimators.

2.4.1 Tail Behavior and Sub-Gaussian Distributions

The *tail behavior* of a random variable describes how likely it is to observe extreme values, far from the mean. For a random variable X with mean $\mu = \mathbb{E}[X]$, the tail probability is the chance that X deviates significantly from μ, i.e., $\mathbb{P}\{|X - \mu| > t\}$ for large t. In mean estimation, tail behavior is crucial because extreme values (outliers) can heavily distort the estimate, especially for distributions with heavy tails.

We start with a Gaussian random variable $X \sim \mathcal{N}(\mu, \sigma^2)$. To analyze the tail behavior, we first standardize the variable. Define

$$g = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1),$$

since $X - \mu \sim \mathcal{N}(0, \sigma^2)$. The tail probability we want to bound is

$$\mathbb{P}\{|X - \mu| > t\} = \mathbb{P}\left\{|g| > \frac{t}{\sigma}\right\}.$$

Let $u = \frac{t}{\sigma} > 0$. We need to compute $\mathbb{P}\{|g| > u\}$, where $g \sim \mathcal{N}(0, 1)$.

Since g follows a standard normal distribution, it is symmetric about 0. Thus

$$\mathbb{P}\{|g| > u\} = \mathbb{P}\{g > u\} + \mathbb{P}\{g < -u\} = 2\mathbb{P}\{g > u\},$$

because $\mathbb{P}\{g < -u\} = \mathbb{P}\{g > u\}$ for $u \geqslant 0$.

Now, we bound $\mathbb{P}\{g > u\}$. For any $\lambda > 0$, apply Markov's inequality to the moment generating function (MGF) of g

$$\mathbb{P}\{g > u\} = \mathbb{P}\left(e^{\lambda g} \geqslant e^{\lambda u}\right) \leqslant \frac{\mathbb{E}\left[e^{\lambda g}\right]}{e^{\lambda u}}.$$

Since the MGF of $g \sim \mathcal{N}(0, 1)$ is $\mathbb{E}[e^{\lambda g}] = e^{\lambda^2/2}$, we have

$$\mathbb{P}\{g > u\} \leqslant \frac{e^{\lambda^2/2}}{e^{\lambda u}} = e^{\lambda^2/2 - \lambda u}.$$

To get the tightest bound, minimize the exponent $\lambda^2/2 - \lambda u$ over $\lambda > 0$, we have

$$\mathbb{P}\{g > u\} \leqslant e^{-u^2/2}.$$

In the end, the tail bound for X is

$$\mathbb{P}\{|X - \mu| > t\} \leqslant 2e^{-t^2/\left(2\sigma^2\right)}.$$

This bound captures the exponential decay of the tail, showing that the Gaussian distribution has a *light tail*. This property means extreme values are unlikely, making the Gaussian distribution useful for mean estimation, as outliers have a reduced impact.

which still captures the exponential decay of the tail. This bound shows that the probability of extreme values decreases exponentially fast, meaning the Gaussian distribution has a *light tail*. This property makes the

Gaussian distribution an ideal benchmark for mean estimation, as extreme values are very unlikely, reducing the impact of outliers on estimates.

In contrast, consider a distribution with much heavier tails, such as the Pareto distribution. A Pareto random variable $Y \sim \text{Pareto}(x_m, \alpha)$ (with minimum value $x_m > 0$ and shape parameter $\alpha > 0$) has density $f(y) = \frac{\alpha x_m^\alpha}{y^{\alpha+1}}$ for $y \geqslant x_m$, and its tail probability is

$$\mathbb{P}\{Y > t\} = \left(\frac{x_m}{t}\right)^\alpha,$$

for $t \geqslant x_m$. Choose $x_m = 1$ and $\alpha = 1.5$ (so the mean exists but the variance does not, since the mean exists for $\alpha > 1$ and the variance exists for $\alpha > 2$). The mean is

$$\mathbb{E}[Y] = \frac{\alpha x_m}{\alpha - 1} = \frac{1.5 \cdot 1}{1.5 - 1} = 3.$$

The tail decays like a power law ($t^{-\alpha}$), which is much slower than the Gaussian's e^{-t^2} decay, indicating a very heavy tail. For example, with $t = 5$, the tail probability is $\mathbb{P}\{Y > 5\} = \left(\frac{1}{5}\right)^{1.5} = 5^{-1.5} \approx 0.089$, compared to a Gaussian $X \sim \mathcal{N}(3, 1)$, where $\mathbb{P}\{X > 5\} \approx 0.0228$. This shows that extreme values are far more likely for the Pareto distribution. When estimating the mean using the sample mean $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n Y_i$, these extreme values can dominate the sum, leading to high variability in the estimate, especially since the variance of Y is infinite. This makes mean estimation for the Pareto distribution far less reliable than for the Gaussian.

The importance of tail behavior lies in its impact on estimator reliability. For light-tailed distributions like the Gaussian, the sample mean $\widehat{\mu}_n$ performs well because extreme values are rare. However, for heavy-tailed distributions, such as the Cauchy or real-world data (e.g., financial returns), the sample mean can be heavily biased by outliers. Our goal is to design estimators whose error $|\widehat{\mu}_n - \mu|$ or $\|\widehat{\mu}_n - \mu\|$ has a light tail, similar to a Gaussian distribution, *even if the underlying data has a heavy tail*. This leads us to the concept of sub-Gaussian estimators, which are designed to achieve such error bounds.

With this understanding of tail behavior, we can define a *sub-Gaussian distribution*, a class of distributions with light tails similar to a Gaussian distribution.

Definition 23 (Sub-Gaussian Distribution)**.** *A random variable X with mean $\mu = \mathbb{E}[X]$ is* sub-Gaussian *if there exists a constant $L > 0$ such that its tail probability satisfies*

$$\mathbb{P}\{|X - \mu| > t\} \leqslant 2\exp\left(-\frac{t^2}{2L^2}\right)$$

for all $t > 0$. The parameter L acts like a variance proxy, controlling the rate at which the tail decays.

Example 24. A sub-Gaussian distribution includes distributions whose tails decay at least as fast as a Gaussian distribution's tails. For example

- If $X \sim \mathcal{N}(\mu, \sigma^2)$, then X is sub-Gaussian with $L = \sigma$, as derived above.
- If X is a bounded random variable, e.g., $X \in [-a, a]$, then

  $$\mathbb{P}\{|X - \mu| > t\} = 0$$

  for $t > a$, so X is sub-Gaussian with an appropriate L depending on a.

The key property of sub-Gaussian distributions is their ability to control the probability of extreme values, even for distributions that are not Gaussian but still have a finite variance. This property is crucial for mean estimation: if the error of an estimator $|\widehat{\mu}_n - \mu|$ has a sub-Gaussian tail, the probability of large errors is very small, making the estimator reliable even for heavy-tailed or contaminated data.

2.4.2 Sub-Gaussian Estimators: Definition and Motivation

With the understanding of tail behavior and sub-Gaussian distributions, we can now define *sub-Gaussian estimators*. These are estimators whose

error distribution has a sub-Gaussian tail, ensuring that large errors are unlikely, even under challenging conditions like heavy-tailed data or adversarial contamination. We will define sub-Gaussian estimators separately for the univariate and multivariate cases, and explain the motivation behind their error bounds, as done in Lugosi and Mendelson [LM19a; LM19b].

Univariate Case. In the univariate setting, where $X \in \mathbb{R}$, we want an estimator $\widehat{\mu}_n$ to approximate the true mean $\mu = \mathbb{E}[X]$ with an error $|\widehat{\mu}_n - \mu|$ that is small with high probability. The empirical mean $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$ has an error bound given by Chebyshev's inequality: if $\mathrm{Var}(X) = \sigma^2$, then $\mathrm{Var}(\widehat{\mu}_n) = \frac{\sigma^2}{n}$, so

$$\mathbb{P}\{|\widehat{\mu}_n - \mu| > t\} \leqslant \frac{\mathrm{Var}(\widehat{\mu}_n)}{t^2} = \frac{\sigma^2}{nt^2}.$$

To achieve a confidence level of $1-\delta$, we set $\frac{\sigma^2}{nt^2} = \delta$, yielding $t = \sigma\sqrt{\frac{1}{n\delta}}$. Thus, the error of the empirical mean is $\varepsilon = \sigma\sqrt{\frac{1}{n\delta}}$. For $n = 1000$ and $\delta = 0.05$, this gives $\varepsilon \approx 0.14\sigma$, as shown in Chapter 1. However, this bound grows quickly as δ becomes small (e.g., for $\delta = 0.01$, $\varepsilon \approx 0.22\sigma$), and it is ineffective if σ^2 is large or infinite (as in heavy-tailed cases like the Cauchy distribution).

Now, consider an ideal scenario: if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\widehat{\mu}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$, and the error $\widehat{\mu}_n - \mu \sim \mathcal{N}(0, \frac{\sigma^2}{n})$. Using the Gaussian tail bound derived earlier, we have

$$\mathbb{P}\{|\widehat{\mu}_n - \mu| > t\} \leqslant 2\exp\left(-\frac{t^2}{2(\sigma^2/n)}\right) = 2\exp\left(-\frac{nt^2}{2\sigma^2}\right).$$

Setting this probability equal to δ, and solve it to find

$$t = \sigma\sqrt{\frac{2\ln(2/\delta)}{n}}.$$

For $\delta = 0.05$, $\ln(2/0.05) \approx 3.69$, so $t \approx \sigma\sqrt{\frac{2\times 3.69}{n}} \approx 2.72\frac{\sigma}{\sqrt{n}}$, or about 0.086σ when $n = 1000$, much smaller than the 0.14σ from Chebyshev's

bound. Moreover, this bound scales with $\sqrt{\log(1/\delta)}$, which grows much more slowly than $\sqrt{1/\delta}$, making it more effective for small δ.

From this, we can derive an ideal form for ε that ensures $\widehat{\mu}_n$ is a robust estimator, known as a *sub-Gaussian estimator*.

Definition 25 (Sub-Gaussian Estimator: Univariate Case)**.** *Let X be a univariate random variable with mean μ and finite variance $\sigma^2 = \mathrm{Var}(X) < \infty$. An estimator $\widehat{\mu}_n$ of μ based on n samples is called* sub-Gaussian *if there exists a constant $L > 0$ such that for all $\delta \in (0, 1)$, the following holds*

$$\mathbb{P}\left\{|\widehat{\mu}_n - \mu| > \varepsilon\right\} \leqslant \delta,$$

where $\varepsilon = L\sigma\sqrt{\frac{\log(2/\delta)}{n}}$ [LM19a].

To justify that the bound $\varepsilon = L\sigma\sqrt{\frac{\log(1/\delta)}{n}}$ for the univariate sub-Gaussian estimator is optimal, we present a minimax lower bound that shows it is the best possible error rate achievable by any estimator, even under ideal conditions. This result, established in Lugosi and Mendelson [LM19a], demonstrates that no estimator can achieve a significantly better bound under the given assumptions (finite variance and a fixed confidence level δ).

Theorem 26 (Minimax Lower Bound for Univariate Mean Estimation)**.** *Let $n > 5$ be a positive integer. Let $\mu \in \mathbb{R}$, $\sigma > 0$, and $\delta \in (2e^{-n/4}, 1/3)$. Then, for any mean estimator $\widehat{\mu}_n$, there exists a distribution with mean μ and variance σ^2 such that*

$$\mathbb{P}\left\{|\widehat{\mu}_n - \mu| > \sigma\sqrt{\frac{\log(1/\delta)}{n}}\right\} \geqslant \delta.$$

Proof. To derive this "minimax" lower bound, it suffices to consider two distributions P_+ and P_-, both concentrated on two points, defined by

$$P_+(0) = P_-(0) = 1 - p, \quad P_+(c) = P_-(-c) = p,$$

where $p \in [0,1]$ and $c > 0$. Note that the means of the two distributions are $\mu_+ = \mathbb{E}[X \sim P_+] = pc$ and $\mu_- = \mathbb{E}[X \sim P_-] = -pc$, and both have variance $\sigma^2 = c^2 p(1-p)$.

For $i = 1, \ldots, n$, let (X_i, Y_i) be independent pairs of real-valued random variables such that

$$\mathbb{P}\{X_i = Y_i = 0\} = 1 - p \quad \text{and} \quad \mathbb{P}\{X_i = c, Y_i = -c\} = p.$$

Note that X_i is distributed as P_+ and Y_i as P_-. Let $\delta \in (2e^{-n/4}, 1/3)$ and $p = \frac{1}{2n}\log(2/\delta)$, then, using $1 - p \geqslant \exp(-p/(1-p))$, we have

$$\mathbb{P}\{(X_1, \ldots, X_n) = (Y_1, \ldots, Y_n)\} = (1-p)^n \geqslant 2\delta.$$

Let $\widehat{\mu}_n$ be any estimator, possibly depending on δ. Then

$$\begin{aligned}
&\max\left(\mathbb{P}\left\{|\widehat{\mu}_n(X_1, \ldots, X_n) - \mu_+| > cp\right\}, \mathbb{P}\left\{|\widehat{\mu}_n(Y_1, \ldots, Y_n) - \mu_-| > cp\right\}\right) \\
&\geqslant \frac{1}{2}\mathbb{P}\left\{\left|\widehat{\mu}_n(X_1, \ldots, X_n) - \mu_{P_+}\right| > cp \quad \text{or} \quad \left|\widehat{\mu}_n(Y_1, \ldots, Y_n) - \mu_{P_-}\right| > cp\right\} \\
&\geqslant \frac{1}{2}\mathbb{P}\left\{\widehat{\mu}_n(X_1, \ldots, X_n) = \widehat{\mu}_n(Y_1, \ldots, Y_n)\right\} \\
&\geqslant \frac{1}{2}\mathbb{P}\left\{(X_1, \ldots, X_n) = (Y_1, \ldots, Y_n)\right\} \geqslant \delta.
\end{aligned}$$

Thus, for any estimator $\widehat{\mu}_n$, possibly depending on δ, there exists a distribution (either P_+ or P_-) with mean $\mu = \pm pc$ and variance $\sigma^2 = c^2 p(1-p)$ such that

$$\mathbb{P}\{|\widehat{\mu}_n - \mu| > cp\} \geqslant \delta.$$

Now, choose c that $cp = \sigma\sqrt{\frac{\log(1/\delta)}{n}}$, where $p = \frac{1}{2n}\log(2/\delta)$, the theorem holds as stated

$$\mathbb{P}\left\{|\widehat{\mu}_n - \mu| > \sigma\sqrt{\frac{\log(1/\delta)}{n}}\right\} \geqslant \delta.$$

This completes the proof. □

This minimax lower bound shows that no estimator can achieve an error smaller than $\sigma\sqrt{\frac{\log(1/\delta)}{n}}$ with probability greater than $1 - \delta$, under the given constraints. Since the sub-Gaussian estimator in Definition 25

achieves this bound (up to a constant L), the form $\varepsilon = L\sigma\sqrt{\frac{\log(1/\delta)}{n}}$ is optimal, and attempting to find a stronger form would be futile. This result also highlights that if the underlying distribution is Gaussian, the sample mean is optimal for all sample sizes and confidence levels [Cat12].

Multivariate Case. Now, we extend the concept of sub-Gaussian estimators to the multivariate setting. Consider a random vector $X \in \mathbb{R}^d$ with mean $\mu = \mathbb{E}[X]$ and covariance matrix Σ such that $\mathrm{Tr}(\Sigma) < \infty$. We aim to construct an estimator $\widehat{\mu}_n \in \mathbb{R}^d$ for μ based on n independent samples $X_1, \ldots, X_n$, such that the error $\|\widehat{\mu}_n - \mu\|$ (in the Euclidean norm) is small with high probability. To motivate the form of a sub-Gaussian estimator in this setting, consider the ideal case where $X \sim \mathcal{N}(\mu, \Sigma)$, a multivariate normal distribution.

If $X \sim \mathcal{N}(\mu, \Sigma)$, the empirical mean $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$ has distribution $\widehat{\mu}_n \sim \mathcal{N}\left(\mu, \frac{1}{n}\Sigma\right)$, and the error satisfies

$$\mathbb{P}\left\{\|\widehat{\mu}_n - \mu\| > t\right\} = \mathbb{P}\left\{\left\|\frac{1}{n}\sum_{i=1}^n (X_i - \mu)\right\| > t\right\} = \mathbb{P}\left\{\|\overline{X}\| > t\sqrt{n}\right\},$$

where $\overline{X} = \frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \mu) \sim \mathcal{N}(0, \Sigma)$. A key property of Gaussian vectors is that $\overline{X}$ has the same distribution as $\Sigma^{1/2}Y$, where $Y \sim \mathcal{N}(0, I)$, and $\Sigma^{1/2}$ is the positive semidefinite square root of Σ. For any $y, y' \in \mathbb{R}^d$,

$$\|\Sigma^{1/2}y\| - \|\Sigma^{1/2}y'\| \leqslant \|\Sigma^{1/2}(y - y')\| \leqslant \|\Sigma^{1/2}\|_{2\to 2}\|y - y'\|,$$

where $\|\Sigma^{1/2}\|_{2\to 2} = \sqrt{\lambda_{\max}(\Sigma)}$, and $\lambda_{\max}(\Sigma)$ is the largest eigenvalue of Σ. Thus, $\Sigma^{1/2}y$ is a Lipschitz function of y with Lipschitz constant $\sqrt{\lambda_{\max}(\Sigma)}$. Using the Gaussian concentration inequality of Tsirelson, Ibragimov, and Sudakov [TIS76; LT91; BLM13], we have

$$\mathbb{P}\left\{\|\overline{X}\| - \mathbb{E}\|\overline{X}\| \ \geqslant t\sqrt{n}\right\} \leqslant e^{-nt^2/(2\lambda_{\max})}.$$

Noting that $\mathbb{E}\|\overline{X}\| \leqslant \sqrt{\mathbb{E}\|\overline{X}\|^2} = \sqrt{\mathrm{Tr}(\Sigma)}$, the trace of the covariance matrix Σ, we obtain, for $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\|\widehat{\mu}_n - \mu\| \leqslant \sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}} + \sqrt{\frac{2\lambda_{\max}(\Sigma)\log(1/\delta)}{n}}.$$

This motivates the form of a sub-Gaussian estimator in the multivariate case. We define a multivariate sub-Gaussian estimator as follows, generalizing Definition 25

Definition 27 (Sub-Gaussian Estimator: Multivariate Case)**.** *Let* $X \in \mathbb{R}^d$ *be a random vector with mean* μ *and covariance matrix* Σ *such that* $\mathrm{Tr}(\Sigma) < \infty$*. An estimator* $\widehat{\mu}_n \in \mathbb{R}^d$ *of* μ *based on* n *samples is called* sub-Gaussian *if there exists a constant* $L > 0$ *such that for all* $\delta \in (0, 1)$*, the following holds:*

$$\mathbb{P}\left\{\|\widehat{\mu}_n - \mu\| > \varepsilon\right\} \leqslant \delta,$$

where

$$\varepsilon = L\left(\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}} + \sqrt{\frac{\lambda_{\max}(\Sigma)\log(2/\delta)}{n}}\right),$$

and $\lambda_{\max}(\Sigma)$ *is the largest eigenvalue of* Σ *[LM19a].*

This form of ε accounts for the multidimensional nature of the data: $\mathrm{Tr}(\Sigma)$ captures the total variance across all dimensions, while $\lambda_{\max}(\Sigma)$ reflects the largest directional variance, influencing the tail behavior in the worst-case direction. In the special case where $d = 1$, this reduces to the univariate form, as $\mathrm{Tr}(\Sigma) = \lambda_{\max}(\Sigma) = \sigma^2$, and the bound becomes $\varepsilon \propto \sigma\sqrt{\frac{\log(1/\delta)}{n}}$, consistent with Definition 25 (up to constant factors in the logarithm).

2.5 Notation and Assumptions

To support the analysis of sub-Gaussian estimators for the mean, as developed in Lugosi and Mendelson [LM19b; LM21], we establish the standard notation and minimal assumptions used in this thesis. These conventions facilitate univariate and multivariate mean estimation, accommodating heavy-tailed distributions and, for certain estimators, adversarial contamination, while enabling sub-Gaussian tail bounds under weak conditions.

We consider a sample of n independent and identically distributed (i.i.d.) observations $X_1, \ldots, X_n$, drawn from a random variable $X \in \mathbb{R}^d$ with

true mean $\mu = \mathbb{E}X$. The estimator for the mean is denoted by $\widehat{\mu}_n$. In the univariate case ($d = 1$), the variance of X is $\sigma^2 = \mathrm{Var}(X)$, and the estimation error is given by $|\widehat{\mu}_n - \mu|$. In the multivariate case ($d \geqslant 1$), the covariance matrix is $\Sigma = \mathbb{E}[(X - \mu)(X - \mu)^\top]$, with total variance $\mathrm{Tr}(\Sigma) = \mathbb{E}\|X - \mu\|^2$ and worst-case directional variance given by the largest eigenvalue $\lambda_{\max}(\Sigma)$. The error is quantified by the Euclidean norm $\|\widehat{\mu}_n - \mu\|$.

The confidence parameter $\delta \in (0, 1)$ specifies the probability of exceeding an error threshold, such as $\mathbb{P}\{|\widehat{\mu}_n - \mu| > \varepsilon\} \leqslant \delta$ in the univariate case or $\mathbb{P}\{\|\widehat{\mu}_n - \mu\| > \varepsilon\} \leqslant \delta$ in the multivariate case. For the trimmed mean estimator in Chapter 4, we adopt the adversarial contamination model of Lugosi and Mendelson [LM21], where up to ηn samples, with $\eta \in [0, 1)$, may be arbitrarily corrupted. This contamination model is not used in the median-of-means tournament estimator of Chapter 3, which focuses solely on heavy-tailed distributions.

The following minimal assumptions ensure the applicability of the estimators

- **Univariate Case**: The random variable $X \in \mathbb{R}$ has finite variance, i.e., $\sigma^2 < \infty$.
- **Multivariate Case**: The random variable $X \in \mathbb{R}^d$ has finite second moments, i.e., $\mathbb{E}\|X\|^2 < \infty$, ensuring a well-defined covariance matrix Σ.
- **Contamination Model (Trimmed Mean Estimator)**: For the trimmed mean estimator in Chapter 4, up to ηn samples may be corrupted, with $0 \leqslant \eta < 1$, as per Lugosi and Mendelson [LM21].

These assumptions require only finite second moments, accommodating heavy-tailed distributions without higher moment constraints. They enable estimators like the median-of-means tournament (Chapter 3) to achieve sub-Gaussian tail bounds, in high-dimensional settings, and the trimmed mean (Chapter 4) to handle both heavy-tailed distributions and adversarial contamination.

Chapter 3

Median-of-Means Estimator

This chapter explores the median-of-means (MoM) estimator for mean estimation, a robust method that achieves sub-Gaussian performance under minimal assumptions, as developed in Lugosi and Mendelson [LM19b; LM19a]. We present its construction, derive error bounds, and provide detailed proofs for both univariate and multivariate settings, building on the theoretical foundations of Chapter 2. The estimator excels for heavy-tailed distributions, requiring only finite variance or second moments, without involving adversarial contamination models, unlike the trimmed mean estimator of Chapter 4.

3.1 Introduction

The median-of-means (MoM) estimator is a cornerstone of robust statistics, designed to estimate the mean of a random variable or vector when the underlying distribution may have heavy tails, such as the Pareto distribution discussed in Chapter 2. Unlike the empirical mean, which is sensitive to outliers and requires conditions on the distribution to achieve sub-Gaussian error bounds, the MoM estimator offers optimal sub-Gaussian performance with only finite second moments, as shown in Lugosi and Mendelson [LM19b; LM19a], regardless of the distribution.

In the univariate case, the MoM estimator, described by Lugosi and Mendelson [LM19a], partitions the sample into groups, computes group means, and takes their median. This leverages the robustness of the median to outliers, achieving sub-Gaussian tails with minimal assumptions. In the multivariate case, Lugosi and Mendelson [LM19b] introduces a tournament-based MoM estimator, selecting a group mean that minimizes the radius of a set defined by pairwise distance comparisons, ensuring uniformity over high-dimensional directions.

This chapter details these approaches. Section 3.2 covers the univariate MoM estimator, presenting its algorithm, error bound, and proof. Section 3.3 extends to the multivariate tournament estimator, detailing its construction, performance, and proof. We rely on tools from Chapter 2, such as concentration inequalities and geometric methods, to establish sub-Gaussian guarantees.

3.2 Univariate Case

In the univariate case ($d = 1$), the median-of-means (MoM) estimator estimates the mean $\mu = \mathbb{E}X$ of a random variable $X \in \mathbb{R}$ with finite variance $\sigma^2 = \mathbb{E}(X - \mu)^2 < \infty$. Given i.i.d. samples $X_1, \ldots, X_n$, the algorithm, as described in Lugosi and Mendelson [LM19a], partitions the sample into groups and computes the median of their means, offering robustness against heavy-tailed outliers.

3.2.1 Construction of the Estimator

1. **Partition the Data**: Divide the indices $[n]$ disjoint blocks $B_1, \ldots, B_k$, each of size at least $\lfloor n/k \rfloor$, where $k = \lceil 8 \log(1/\delta) \rceil$ for a confidence parameter $\delta \in (0, 1)$. Assume for simplicity that $n = mk$, so each block has size $m = n/k$.

2. **Compute Block Means**: For each block $i = 1, \ldots, k$, compute the sample mean:
$$Z_i = \frac{1}{m} \sum_{j \in B_i} X_j.$$

3. **Evaluate Median**: Define the median of the block means $Z_1, \ldots, Z_k$ as
$$M(Z_1, \ldots, Z_k) := Z_c,$$
where $c \in [k]$ such that Z_c satisfies
$$|\{j \in [k] : Z_j \leqslant Z_c\}| \geqslant \frac{k}{2}, \quad |\{j \in [k] : Z_j \geqslant Z_c\}| \geqslant \frac{k}{2}.$$
If multiple indices qualify, take the smallest. Set the estimator:
$$\widehat{\mu}_n = M(Z_1, \ldots, Z_k).$$

The choice of $k \approx \log(1/\delta)$ ensures enough blocks to achieve high confidence, while $m \approx n/\log(1/\delta)$ keeps each block large for reliable means. The median step guards against outlying block means, unlike the empirical mean.

3.2.2 Performance Bound

The performance bound, presented below, quantifies the estimator's error $|\widehat{\mu}_n - \mu|$ in terms of the sample size n, variance σ^2, and confidence level δ. It assumes n i.i.d. samples $X_1, \ldots, X_n$, divided into k blocks of size m, where $n = mk$, as described in earlier. The bound is expressed in two forms: a general form depending on block parameters, and a specific form optimized for a chosen number of blocks. This result, adapted from Lugosi and Mendelson [LM19b], highlights the MoM's effectiveness under minimal assumptions.

Theorem 28. *Let X_1, X_2, …, X_n be independent, identically distributed random variables with mean μ and variance σ^2. Let m, k be positive*

integers such that $n = mk$. Then the median-of-means estimator $\widehat{\mu}_n$ with k blocks satisfies:

$$\mathbb{P}\left\{|\widehat{\mu}_n - \mu| > \sigma\sqrt{\frac{4}{m}}\right\} \leqslant e^{-k/8}.$$

In particular, for any $\delta \in (0,1)$, if $k = \lceil 8\log(1/\delta)\rceil$, then, with probability at least $1-\delta$:

$$|\widehat{\mu}_n - \mu| \leqslant \sigma\sqrt{\frac{32\log(1/\delta)}{n}}.$$

The theorem provides two error bounds, each offering insight into the MoM estimator's performance

- **General Bound**: The first bound, $\mathbb{P}\left\{|\widehat{\mu}_n - \mu| > \sigma\sqrt{\frac{4}{m}}\right\} \leqslant e^{-k/8}$, relates the error to the block size m and number of blocks k. The error term $\sigma\sqrt{\frac{4}{m}}$ depends on the variance σ^2 and the number of samples per block m, decreasing as m increases because larger blocks produce more accurate block means. The probability bound $e^{-k/8}$ decays exponentially with k, the number of blocks, reflecting the robustness of taking the median of more block means. For example, with $\sigma^2 = 1$, $m = 100$, and $k = 10$, the error is $\sqrt{\frac{4}{100}} = 0.2$, and the probability of exceeding this error is $e^{-10/8} \approx 0.287$, indicating moderate confidence.

- **Optimized Bound**: The second bound, $|\widehat{\mu}_n - \mu| \leqslant \sigma\sqrt{\frac{32\log(1/\delta)}{n}}$, is achieved by setting $k = \lceil 8\log(1/\delta)\rceil$. This expresses the error in terms of the total sample size n and confidence level δ, making it more intuitive for practical use. The term $\sigma\sqrt{\frac{32\log(1/\delta)}{n}}$ scales with the standard deviation σ, decreases with n, and increases logarithmically with $1/\delta$, reflecting higher precision for larger samples or lower confidence requirements. For instance, with $\sigma^2 = 1$, $n = 1000$, and $\delta = 0.05$, we have $\log(1/0.05) \approx 3$, $k = \lceil 8 \cdot 3\rceil = 24$, and the error bound is $\sqrt{\frac{32\cdot 3}{1000}} \approx 0.31$, ensuring the estimate is accurate with 95% probability.

3.2.3 Proof of Performance Bound

The proof is based on the proof for Theorem 2 in Lugosi and Mendelson [LM19a] (pages 7–8). We provide a detailed explanation of each step to enhance clarity, and verify correctness using tools from Chapter 2 (Section 2.1).

Step 1: Concentration of Block Means.

For each block $i = 1, \ldots, k$, the block mean is

$$Z_i = \frac{1}{m} \sum_{j \in B_i} X_j.$$

Since the X_j are i.i.d. with $\mathbb{E}X_j = \mu$ and $\mathrm{Var}(X_j) = \sigma^2$, we have

$$\mathbb{E}Z_i = \mu, \quad \mathrm{Var}(Z_i) = \frac{\sigma^2}{m}.$$

By Chebyshev's inequality (Chapter 2), for any $\theta > 0$,

$$\mathbb{P}\left\{|Z_i - \mu| > \theta\right\} \leqslant \frac{\mathrm{Var}(Z_i)}{\theta^2} = \frac{\sigma^2}{m\theta^2}.$$

Set $\theta = \sigma\sqrt{4/m}$, we have

$$\mathbb{P}\left\{|Z_i - \mu| > \sigma\sqrt{\frac{4}{m}}\right\} \leqslant \frac{\sigma^2/m}{(\sigma\sqrt{4/m})^2} = \frac{\sigma^2/m}{4\sigma^2/m} = \frac{1}{4}.$$

Thus, with probability at least $3/4$,

$$|Z_i - \mu| \leqslant \sigma\sqrt{\frac{4}{m}}.$$

This bound shows that each block mean is likely close to the true mean, with the error scaling inversely with the block size m.

Step 2: Median and Bad Blocks.

The estimator

$$\widehat{\mu}_n = M(Z_1, \ldots, Z_k) = Z_c$$

selects a block mean such that at least $k/2$ block means are at least Z_c and at least $k/2$ are at most Z_c.

If the estimator deviates significantly, i.e., $|\widehat{\mu}_n - \mu| > \sigma\sqrt{4/m}$, then $|Z_c - \mu| > \sigma\sqrt{4/m}$. This implies that at least $k/2$ block means Z_j are far from μ. For example, if $Z_i > \mu + \sigma\sqrt{4/m}$, then at least $k/2$ block means satisfy $Z_j \geqslant Z_i > \mu + \sigma\sqrt{4/m}$, all deviating by more than $\sigma\sqrt{4/m}$.

Define a block i as "bad" if

$$|Z_i - \mu| > \sigma\sqrt{\frac{4}{m}}.$$

From Step 1, each block is bad with probability at most $1/4$. Thus, the number of bad blocks, $B = \sum_{i=1}^{k} \mathbf{1}_{\{|Z_i - \mu| > \sigma\sqrt{4/m}\}}$, follows a binomial distribution, $\mathrm{Bin}(k, p)$, with $p \leqslant 1/4$. The event $|\widehat{\mu}_n - \mu| > \sigma\sqrt{4/m}$ requires at least $k/2$ bad blocks, so

$$\mathbb{P}\left\{|\widehat{\mu}_n - \mu| > \sigma\sqrt{\frac{4}{m}}\right\} \leqslant \mathbb{P}\left\{B \geqslant \frac{k}{2}\right\}.$$

This step leverages the median's property: a large deviation in $\widehat{\mu}_n$ implies many block means are far from μ, which is unlikely because μ is the true mean.

Step 3: Binomial Tail Bound.

We bound the probability of having at least $k/2$ bad blocks. Since $\mathbb{E}B \leqslant k/4$, we compute

$$\mathbb{P}\left\{B \geqslant \frac{k}{2}\right\} = \mathbb{P}\left\{B - \mathbb{E}B \geqslant \frac{k}{4}\right\}.$$

By Hoeffding's inequality for a binomial random variable (Chapter 2), for $t \geqslant 0$

$$\mathbb{P}\left\{B - \mathbb{E}B \geqslant t\right\} \leqslant \exp\left(-\frac{2t^2}{k}\right).$$

Set $t = k/4$, so

$$\mathbb{P}\left\{B \geqslant \frac{k}{2}\right\} \leqslant \exp\left(-\frac{2(k/4)^2}{k}\right) = \exp\left(-\frac{k}{8}\right).$$

Hence,

$$\mathbb{P}\left\{|\widehat{\mu}_n - \mu| > \sigma\sqrt{\frac{4}{m}}\right\} \leqslant \exp\left(-\frac{k}{8}\right).$$

This establishes the first part of the theorem, showing that the probability of a large error decays exponentially with k.

Step 4: Confidence Parameter and Sub-Gaussian Form.

To achieve the confidence level δ, note that $k = \lceil 8\log(1/\delta)\rceil$. Since $m = n/k \approx n/(8\log(1/\delta))$, the error bound becomes

$$\sigma\sqrt{\frac{4}{m}} \approx \sigma\sqrt{\frac{4k}{n}} \approx \sigma\sqrt{\frac{4\cdot 8\log(1/\delta)}{n}} = \sigma\sqrt{\frac{32\log(1/\delta)}{n}}.$$

This gives

$$\mathbb{P}\left\{|\widehat{\mu}_n - \mu| > \sigma\sqrt{\frac{32\log(1/\delta)}{n}}\right\} \leqslant \delta.$$

This completes the proof.

3.3 Multivariate Case

In this section, we explore the multivariate extension of the median-of-means (MoM) estimator, which achieves sub-Gaussian performance for estimating the mean of a random vector $X \in \mathbb{R}^d$ with mean $\mu = \mathbb{E}[X]$ and covariance matrix $\Sigma = \mathbb{E}\left[(X-\mu)(X-\mu)^T\right]$. This estimator is robust against heavy-tailed distributions, requiring only that the second moment of X exists. We present its construction, performance bound, and a detailed proof, drawing on the theoretical tools from Chapter 2 to ensure clarity and coherence with the framework established earlier.

3.3.1 Construction of the Estimator

The multivariate MoM estimator generalizes the univariate approach by introducing a novel concept of a multivariate median, interpreted as a

tournament among points in $\mathbb{R}^d$. Given an i.i.d. sample $X_1, \ldots, X_n$ of size n, the estimator is constructed as follows:

1. **Partition the Data**: Divide the index set $\{1, \ldots, n\}$ into k disjoint blocks $B_1, \ldots, B_k$, each of size $m = n/k$, where k is a parameter chosen based on the desired confidence level δ. For simplicity, we assume n is divisible by k.

2. **Compute Block Means**: For each block B_j, calculate the sample mean:
$$Z_j = \frac{1}{m} \sum_{i \in B_j} X_i.$$

3. **Define the Tournament Set**: For any point $a \in \mathbb{R}^d$, define the set $T_a \subset \mathbb{R}^d$ as:
$$T_a = \left\{ x \in \mathbb{R}^d \mid \exists J \subset [k], |J| > \frac{k}{2} : \forall j \in J, \|Z_j - x\| \leqslant \|Z_j - a\| \right\}.$$
Here, T_a contains points x that are at least as close as a to the block means Z_j for a majority of the blocks (more than $k/2$).

4. **Minimize the Radius**: Define the radius of T_a as:
$$\text{radius}(T_a) = \sup_{x \in T_a} \|x - a\|.$$
The estimator $\widehat{\mu}_n$ is chosen as:
$$\widehat{\mu}_n \in \operatorname*{argminradius}_{a \in \mathbb{R}^d}(T_a).$$
If multiple minimizers exist, any one is selected. The minimum is achieved because $\text{radius}(T_a)$ is continuous, as T_a is an intersection of a finite union of closed balls with centers and radii continuous in a.

This construction defines a multivariate median of the block means Z_1, $\ldots$, Z_k. In the univariate case ($d = 1$), it reduces to the standard MoM estimator, where the median balances the number of block means above and below it.

3.3.2 Performance Bound

The multivariate MoM estimator achieves sub-Gaussian performance, as encapsulated in the following theorem

Theorem 29. *Let $\delta \in (0,1)$, and consider the multivariate MoM estimator $\widehat{\mu}_n$ with parameter $k = \lceil 200 \log(2/\delta) \rceil$. For i.i.d. random vectors $X_1, \ldots, X_n \in \mathbb{R}^d$ with mean μ and covariance matrix Σ, with probability at least $1 - \delta$,*

$$\|\widehat{\mu}_n - \mu\| \leqslant \max\left(960\sqrt{\frac{\operatorname{tr}(\Sigma)}{n}}, 240\sqrt{\frac{\lambda_{\max} \log(2/\delta)}{n}}\right).$$

The error bound comprises two terms, each addressing a key aspect of high-dimensional mean estimation

- **Average Variance Term**: The term $\sqrt{\frac{\operatorname{tr}(\Sigma)}{n}}$ captures the average variance across all d dimensions, where $\operatorname{tr}(\Sigma) = \sum_{i=1}^{d} \mathbb{E}\left[(X_i - \mu_i)^2\right]$ sums the variances of X's coordinates. It decreases as n increases, showing that larger samples reduce estimation error. For example, in a 3D dataset ($d = 3$) with $n = 1000$ and $\operatorname{tr}(\Sigma) = 3$, this term is approximately $\sqrt{\frac{3}{1000}} \approx 0.055$, indicating a small error contribution when the sample size is sufficient.
- **Directional Variance Term**: The term $\sqrt{\frac{\lambda_{\max}(\Sigma) \log(2/\delta)}{n}}$ accounts for the worst-case variance in any direction, where $\lambda_{\max}(\Sigma)$ is the largest eigenvalue of Σ. The factor $\log(2/\delta)$ ensures uniform error control across all directions in the unit sphere $S^{d-1} = \{v \in \mathbb{R}^d : \|v\| = 1\}$, with smaller δ (higher confidence) increasing the bound logarithmically. In the example, if $\lambda_{\max}(\Sigma) = 1$ and $\delta = 0.05$, then $\log(2/0.05) \approx\approx 3.69$, so this term is approximately $\sqrt{\frac{1 \cdot 3.69}{1000}} \approx 0.061$, slightly larger than the average variance term due to the confidence requirement.

The MoM estimator constructs k block means from n samples and takes their coordinate-wise median, as described in Chapter 3. This leverages the median's robustness to mitigate extreme values in heavy-tailed

distributions, unlike the empirical mean, which can be distorted by outliers. The bound's two terms reflect the estimator's ability to manage overall variability ($\operatorname{tr}(\Sigma)$) and worst-case variability ($\lambda_{\max}(\Sigma)$) in high dimensions. For a 10D dataset with $d = 10$, $n = 1000$, $\operatorname{tr}(\Sigma) = 10$, $\lambda_{\max}(\Sigma) = 1$, and $\delta = 0.05$, the bound is approximately

$$\max\left(960 \cdot 0.1 + 240 \cdot 0.061\right) \approx 14.64,$$

demonstrating precision despite heavy-tailed data.

3.3.3 Proof of Performance Bound

The proof relies on a geometric property of the estimator, formalized in Theorem 30, which implies Theorem 29. We provide a detailed, step-by-step proof, using tools from Chapter 2, including Chebyshev's inequality, binomial tail estimates, the dual Sudakov inequality, symmetrization inequalities, and the contraction lemma.

Theorem 30. *With the same notation as Theorem 29, set*

$$r = \max\left(960\sqrt{\frac{\operatorname{tr}(\Sigma)}{n}}, 240\sqrt{\frac{\lambda_{\max}\log(2/\delta)}{n}}\right).$$

With probability at least $1-\delta$, for any $a \in \mathbb{R}^d$ such that $\|a-\mu\| \geqslant r$, one has $\|Z_j - a\| > \|Z_j - \mu\|$ for more than $k/2$ indices j.

Proof of Theorem 30. The proof employs a "median-of-means tournament" concept, where points $a, b \in \mathbb{R}^d$ are compared based on their empirical performance across blocks. We say that a *defeats* b if

$$\frac{1}{m}\sum_{i \in B_j}\left(\|X_i - b\|^2 - \|X_i - a\|^2\right) > 0$$

for more than $k/2$ blocks B_j. Our goal is to show that μ defeats any b with $\|b-\mu\| \geqslant r$, with high probability, and relate this to the condition $\|Z_j - a\| > \|Z_j - \mu\|$.

Step 1: Express the Defeat Condition.

For any $a, b \in \mathbb{R}^d$, we have

$$\begin{aligned}\|X_i - b\|^2 - \|X_i - a\|^2 &= \|X_i - b - (a-b)\|^2 - \|X_i - b\|^2 \\ &= -2\langle X_i - b, a - b\rangle + \|a-b\|^2.\end{aligned}$$

Thus, for block B_j, we have

$$\begin{aligned}&\frac{1}{m}\sum_{i\in B_j}\left(\|X_i - b\|^2 - \|X_i - a\|^2\right) \\ =& -2\left\langle \frac{1}{m}\sum_{i\in B_j} X_i - b, a - b\right\rangle + \|a-b\|^2 \\ =& -2\langle Z_j - b, a - b\rangle + \|a-b\|^2 \\ =& \|Z_j - a\|^2 - \|Z_j - b\|^2.\end{aligned}$$

Hence, a defeats b on block B_j if and only if

$$\|Z_j - a\|^2 - \|Z_j - b\|^2 < 0 \quad \text{or equivalently,} \quad \|Z_j - a\| < \|Z_j - b\|.$$

In the end, if μ defeats b then $\|Z_j - \mu\| < \|Z_j - b\|$ for more than $k/2$ blocks.

Step 2: Analyze the Defeat of μ over b.

Set $v = b - \mu$, so $b = \mu + v$, and assume $\|v\| \geqslant r$. We need

$$\frac{1}{m}\sum_{i\in B_j}\left(\|X_i - (\mu+v)\|^2 - \|X_i - \mu\|^2\right) > 0$$

for most blocks. Operating as in step 1, we have

$$\|X_i - (\mu+v)\|^2 - \|X_i - \mu\|^2 = -2\langle X_i - \mu, v\rangle + \|v\|^2.$$

Let $\overline{X}_i = X_i - \mu$, then

$$\frac{1}{m}\sum_{i\in B_j}\left(\|X_i - b\|^2 - \|X_i - \mu\|^2\right) = -\frac{2}{m}\sum_{i\in B_j}\langle \overline{X}_i, v\rangle + \|v\|^2.$$

Thus, μ defeats b on block B_j if

$$-\frac{2}{m}\sum_{i\in B_j}\langle \overline{X}_i, v\rangle + \|v\|^2 > 0. \tag{3.1}$$

And, Theorem 30 is true if this condition holds for more than $k/2$ blocks B_j, when $\|v\| \geqslant r$. It suffices to prove the result for $\|v\| = r$.

Step 3: Control the Inner Product Term for a Fixed v.

Fix $v \in \mathbb{R}^d$ with $\|v\| = r$. The term $\frac{1}{m}\sum_{i\in B_j}\langle \overline{X}_i, v\rangle$ is the average of centered random variables. By Chebyshev's inequality (Theorem 2 in Chapter 2),

$$\mathbb{P}\left\{\left|\frac{1}{m}\sum_{i\in B_j}\langle \overline{X}_i, v\rangle\right| \geqslant t\right\} \leqslant \frac{\mathbb{E}\left[\langle \overline{X}_1, v\rangle^2\right]}{mt^2}.$$

Since $\mathbb{E}\left[\langle \overline{X}_1, v\rangle^2\right] = \langle v, \Sigma v\rangle \leqslant \lambda_{\max}\|v\|^2$, set $t = \sqrt{10}\sqrt{\frac{\lambda_{\max}\|v\|^2}{m}}$, then

$$\mathbb{P}\left\{\left|\frac{1}{m}\sum_{i\in B_j}\langle \overline{X}_i, v\rangle\right| \leqslant \sqrt{10}\|v\|\sqrt{\frac{\lambda_{\max}}{m}}\right\} \geqslant \frac{9}{10}.$$

Thus,

$$-\frac{2}{m}\sum_{i\in B_j}\langle \overline{X}_i, v\rangle \geqslant -2\sqrt{10}\|v\|\sqrt{\frac{\lambda_{\max}}{m}}$$

with probability at least 9/10. Since $\|v\| = r \geqslant 960\sqrt{\frac{\lambda_{\max}\log(2/\delta)}{n}}$ and $m = n/k = n/\lceil 200\log(2/\delta)\rceil$, we have

$$-2\sqrt{10}\|v\|\sqrt{\frac{\lambda_{\max}}{m}} \geqslant -\frac{r^2}{2}.$$

Thus,

$$-\frac{2}{m}\sum_{i\in B_j}\langle \overline{X}_i, v\rangle \geqslant -\frac{r^2}{2} \tag{3.2}$$

with probability at least $9/10$.

Step 4: Ensure Majority of Blocks for a Fixed v.

For a fixed v, the number of blocks satisfying the condition (3.2) follows a binomial distribution. With probability at least $9/10$ per block, a binomial tail estimate (related to Hoeffding's inequality, Theorem 3 in Chapter 2) gives

$$\mathbb{P}\left\{\text{at least } \frac{8k}{10} \text{ blocks satisfy (3.2)}\right\} \geqslant 1 - \exp\left(-\frac{k}{50}\right).$$

Since $k = \lceil 200 \log(2/\delta) \rceil$, we have $\exp(-k/50) \leqslant \exp(-4\log(2/\delta)) = (2/\delta)^4 \ll \delta$.

Step 5: Uniform Control Over an ε-Maximal Separated Set.

To extend the analysis to all vectors $v \in \mathbb{R}^d$ with $\|v\| = r$, we consider the sphere $rS^{d-1} = \{v \in \mathbb{R}^d : \|v\| = r\}$. Since rS^{d-1} contains infinitely many points, we approximate it with a finite set $V_1 \subset rS^{d-1}$, which we construct as an ε-maximal separated set (Definition 13) in the $L_2(X)$-norm.

Let X be a random vector in $\mathbb{R}^d$ such that X_i be the copy of X. As defined in Section 2.2, the $L_2(X)$-norm of a vector $v \in \mathbb{R}^d$ is

$$\|v\|_{L_2(X)} = \sqrt{\mathbb{E}\left[\langle X - \mu, v\rangle^2\right]} = \sqrt{\langle v, \Sigma v\rangle},$$

where $\langle v, \Sigma v\rangle = v^T \Sigma v$. For two vectors $v_1, v_2 \in rS^{d-1}$, their distance is:

$$\|v_1 - v_2\|_{L_2(X)} = \sqrt{\langle v_1 - v_2, \Sigma(v_1 - v_2)\rangle}.$$

We construct $V_1 \subset rS^{d-1}$ such that any two distinct points in V_1 are at least ε apart in the $L_2(X)$-norm, and every point on rS^{d-1} is within ε of some point in V_1.

To bound the size of V_1, we apply the dual Sudakov inequality (Theorem 14). For the $L_2(X)$-norm, the inequality gives

$$\log(|V_1|/2) \leqslant \frac{1}{32}\left(\frac{\mathbb{E}\left[\sqrt{\langle G, \Sigma G\rangle}\right]}{\varepsilon/r}\right)^2,$$

where $G \sim N(0, I_d)$ is a standard Gaussian vector.

Since

$$\begin{aligned}\mathbb{E}\left[\sqrt{\langle G, \Sigma G\rangle}\right] &= \mathbb{E}_G\left[\sqrt{\mathbb{E}_X\left[\langle G, \overline{X}\rangle^2\right]}\right] \\ &\leqslant \sqrt{\mathbb{E}_X\mathbb{E}_G\left[\langle G, \overline{X}\rangle^2\right]} \\ &= \sqrt{\mathbb{E}[\|\overline{X}\|^2]} \\ &= \sqrt{\mathrm{Tr}(\Sigma)}.\end{aligned}$$

Set $\varepsilon = 2r\sqrt{\frac{\mathrm{tr}(\Sigma)}{k}}$, then

$$\log\left(|V_1|/2\right) \leqslant \frac{1}{32} \cdot \frac{k\,\mathrm{tr}\left(\Sigma\right)}{4\,\mathrm{tr}\left(\Sigma\right)} = \frac{k}{128}.$$

Thus, $|V_1| \leqslant 2e^{k/100}$. Now, by the union bound, we have

$$\begin{aligned}&\mathbb{P}\left\{\exists v \in V_1 : \text{fewer than } \frac{8k}{10} \text{ blocks satisfy (3.2)}\right\} \\ &\leqslant 2e^{k/100} \cdot e^{-k/50} \leqslant 2e^{-k/100} \leqslant \frac{\delta}{2}.\end{aligned}$$

Step 6: Extend to All $v \in rS^{d-1}$.

In the previous step, we showed that for any fixed $v \in V_1$, a maximal ε-separated subset of the sphere $rS^{d-1} = \left\{v \in \mathbb{R}^d : \|v\| = r\right\}$, the condition (3.2)

$$-\frac{2}{m}\sum_{i \in B_j}\langle \overline{X}_i, v\rangle > -\frac{r^2}{2}$$

holds for at least $8k/10$ blocks with high probability (at least $1 - \delta/2$).

However, in the end, our goal is to prove (3.1) for *all* $v \in rS^{d-1}$

$$-\frac{2}{m}\sum_{i \in B_j}\langle \overline{X}_i, v\rangle + \|v\|^2 > 0$$

(note that this is different from the form of (3.2)).

For any $x \in rS^{d-1}$, let $v_x \in V_1$ be the point in V_1 closest to x in the $L_2(X)$-norm, defined as

$$\|x - v_x\|_{L_2(X)} = \sqrt{\mathbb{E}\left[\langle \overline{X}_1, x - v_x\rangle^2\right]} = \sqrt{\langle x - v_x, \Sigma(x - v_x)\rangle} \leqslant \varepsilon,$$

where $\varepsilon = 2r\sqrt{\frac{\operatorname{tr}(\Sigma)}{k}}$, as chosen in Step 5. We will prove that for *every* $x \in rS^{d-1}$, the original condition (3.1)

$$-\frac{2}{m}\sum_{i \in B_j} \langle \overline{X}_i, x\rangle + r^2 > 0$$

holds for at least $7k/10$ blocks, with high probability. Rewrite the expression by decomposing $x = v_x + (x - v_x)$ as

$$-\frac{2}{m}\sum_{i \in B_j} \langle \overline{X}_i, x\rangle = -\frac{2}{m}\sum_{i \in B_j} \langle \overline{X}_i, v_x\rangle - \frac{2}{m}\sum_{i \in B_j} \langle \overline{X}_i, x - v_x\rangle.$$

Thus, the condition (3.1) becomes

$$-\frac{2}{m}\sum_{i \in B_j} \langle \overline{X}_i, v_x\rangle - \frac{2}{m}\sum_{i \in B_j} \langle \overline{X}_i, x - v_x\rangle + r^2 > 0. \tag{3.3}$$

From Step 5, we know that for each $v_x \in V_1$, there are at least $8k/10$ blocks where the condition (3.2) holds, or

$$-\frac{2}{m}\sum_{i \in B_j} \langle \overline{X}_i, v_x\rangle \geqslant -\frac{r^2}{2},$$

with probability at least $1 - \delta/2$. For these blocks, the condition (3.3) reduces to

$$\frac{r^2}{2} - \frac{2}{m}\sum_{i \in B_j} \langle \overline{X}_i, x - v_x\rangle > 0,$$

which holds if

$$\left|\frac{1}{m}\sum_{i \in B_j} \langle \overline{X}_i, x - v_x\rangle\right| < \frac{r^2}{4}. \tag{3.4}$$

Since there are at least $8k/10$ blocks B_j where (3.2) is controlled, we need to ensure that (3.4) also holds for most of the blocks B_j. Specifically, we aim for this to hold for at least $9k/10$ blocks across all $x \in rS^{d-1}$, so that the intersection gives at least $8k/10 - k/10 = 7k/10$ blocks where both conditions hold, ensuring the condition (3.1) for all x.

Define the quantity to control uniformly

$$\frac{1}{k}\sum_{j=1}^{k} \mathbf{1}_{\left\{\frac{1}{m}\sum_{i\in B_j} \left|\langle \overline{X}_i, x - v_x\rangle\right| \geqslant \frac{r^2}{4}\right\}},$$

which represents the fraction of blocks where the deviation exceeds $\frac{r^2}{4}$ for a specific x. We need to bound this uniformly over all $x \in rS^{d-1}$, so we consider

$$Y = \sup_{x\in rS^{d-1}} \frac{1}{k}\sum_{j=1}^{k} \mathbf{1}_{\left\{\frac{1}{m}\sum_{i\in B_j} \left|\langle \overline{X}_i, x - v_x\rangle\right| \geqslant \frac{r^2}{4}\right\}}.$$

Our objective is to show

$$\mathbb{P}\left\{Y \leqslant 0.1\right\} \geqslant 1 - \frac{\delta}{2},$$

meaning that for every $x \in rS^{d-1}$, there are at most $0.1k$ blocks where the deviation is large, or equivalently, at least $0.9k$ blocks where

$$\left|\frac{1}{m}\sum_{i\in B_j} \langle \overline{X}_i, x - v_x\rangle\right| < \frac{r^2}{4}.$$

Combining with the $8k/10$ blocks from Step 5, this ensures at least $7k/10$ blocks satisfy the full condition.

Consider the expectation of Y

$$\mathbb{E}\left[Y\right] = \mathbb{E}\sup_{x\in rS^{d-1}} \frac{1}{k}\sum_{j=1}^{k} \mathbf{1}_{\left\{\frac{1}{m}\sum_{i\in B_j} \left|\langle \overline{X}_i, x - v_x\rangle\right| \geqslant \frac{r^2}{4}\right\}}.$$

We have

$$\sum_{j=1}^{k} \mathbf{1}\left\{ \tfrac{1}{m} \sum_{i \in B_j} \left| \langle \overline{X}_i, x - v_x \rangle \right| \geqslant \frac{r^2}{4} \right\} \leqslant \left(\frac{1}{m} \sum_{i \in B_j} \left| \langle \overline{X}_i, x - v_x \rangle \right| \right) / \frac{r^2}{4}.$$

Thus,

$$\begin{aligned} Y &\leqslant \sup_{x \in rS^{d-1}} \frac{1}{k} \sum_{j=1}^{k} \frac{\frac{1}{m} \sum_{i \in B_j} \left| \langle \overline{X}_i, x - v_x \rangle \right|}{\frac{r^2}{4}} \\ &= \frac{4}{r^2} \sup_{x \in rS^{d-1}} \frac{1}{k} \sum_{j=1}^{k} \left| \frac{1}{m} \sum_{i \in B_j} \langle \overline{X}_i, x - v_x \rangle \right|, \end{aligned}$$

and

$$\mathbb{E}[Y] \leqslant \frac{4}{r^2} \mathbb{E} \sup_{x \in rS^{d-1}} \frac{1}{k} \sum_{j=1}^{k} \left| \frac{1}{m} \sum_{i \in B_j} \langle \overline{X}_i, x - v_x \rangle \right|.$$

Define the empirical process

$$Z = \sup_{x \in rS^{d-1}} \frac{1}{k} \sum_{j=1}^{k} \left| \frac{1}{m} \sum_{i \in B_j} \langle \overline{X}_i, x - v_x \rangle \right| = \sup_{x \in rS^{d-1}} \frac{1}{k} \sum_{j=1}^{k} \left| \frac{1}{m} \sum_{i \in B_j} f_{x,v_x}(\overline{X}_i) \right|,$$

where $f_{x,v_x}(\overline{X}_i) = \langle \overline{X}_i, x - v_x \rangle$. We need to bound $\mathbb{E}[Z]$, which will be done with the symmetrization inequality.

To have the approviate form to do that, let's rewrite

$$\begin{aligned} \mathbb{E}[Z] &= \mathbb{E}\left[\sup_{x \in rS^{d-1}} \frac{1}{k} \sum_{j=1}^{k} \left| \frac{1}{m} \sum_{i \in B_j} f_{x,v_x}(\overline{X}_i) \right| \right] \\ &= \mathbb{E} \sup_{x \in \mathcal{S}^{d-1}} \frac{1}{k} \sum_{j=1}^{k} \left(\left| \frac{1}{m} \sum_{i \in B_j} f_{x,v_x}(\overline{X}_i) \right| - \mathbb{E} \left| \frac{1}{m} \sum_{i \in B_j} f_{x,v_x}(\overline{X}_i) \right| \right) \\ &\quad + \mathbb{E} \left| \frac{1}{m} \sum_{i \in B_j} f_{x,v_x}(\overline{X}_i) \right| \end{aligned}$$

$$\stackrel{\text{def}}{=} (A) + (B).$$

Since $\|x - v_x\|_{L_2(X)} = \left(\mathbb{E}\langle X, x - v_x\rangle^2\right)^{1/2} \leqslant \varepsilon$, it follows that for every j

$$(B) \leqslant \sqrt{\mathbb{E}\left[\frac{1}{m}\sum_{i\in B_j}\langle X_i, x - v_x\rangle\right]^2} \leqslant \frac{\varepsilon}{\sqrt{m}}.$$

So, we only need to bound (A).

First, apply the symmetrization inequality (Theorem 16 in Chapter 2). Let $\sigma_1, \ldots, \sigma_n$ be independent Rademacher random variables (i.e., $\mathbb{P}\{\sigma_i = 1\} = \mathbb{P}\{\sigma_i = -1\} = 1/2$), independent of the $\overline{X}_i$, we have

$$(A) \leqslant 2\mathbb{E}\sup_{x\in rS^{d-1}}\frac{1}{k}\sum_{j=1}^{k}\sigma_j\left|\frac{1}{m}\sum_{i\in B_j} f_{x,v_x}(\overline{X}_i)\right|.$$

Next, apply the contraction lemma (Theorem 18 in Chapter 2). The function $\phi(u) = |u|$ is Lipschitz with constant $L = 1$ (since $||u| - |v|| \leqslant |u - v|$) and satisfies $\phi(0) = 0$. Thus,

$$\mathbb{E}\sup_{x\in rS^{d-1}}\frac{1}{k}\sum_{j=1}^{k}\sigma_j\left|\frac{1}{m}\sum_{i\in B_j} f_{x,v_x}(\overline{X}_i)\right| \leqslant \mathbb{E}\sup_{x\in rS^{d-1}}\frac{1}{k}\sum_{j=1}^{k}\sigma_j\frac{1}{m}\sum_{i\in B_j} f_{x,v_x}(\overline{X}_i).$$

Aplpy the second part of Theorem 16, we get

$$\begin{aligned}\mathbb{E}\sup_{x\in rS^{d-1}}\frac{1}{k}\sum_{j=1}^{k}\sigma_j\frac{1}{m}\sum_{i\in B_j} f_{x,v_x}(\overline{X}_i) &\leqslant 2\mathbb{E}\sup_{x\in rS^{d-1}}\frac{1}{k}\sum_{j=1}^{k}\frac{1}{m}\sum_{i\in B_j} f_{x,v_x}(\overline{X}_i)\\ &\leqslant 2\mathbb{E}\sup_{x\in rS^{d-1}}\sum_{i=1}^{n}\left|\frac{1}{n}\sum_{i\in B_j} f_{x,v_x}(\overline{X}_i)\right|\end{aligned}$$

Now, we will bound the expectation on the right. Consider the function class $\mathcal{F} = \left\{f_{x,v_x} : x \in rS^{d-1}, v_x \in V_1\right\}$. For any x, since v_x is the closest point in V_1, we have $\|x - v_x\|_{L_2(X)} \leqslant \varepsilon$, but in the Euclidean norm, we

have $\|x - v_x\| \leqslant 2r$ (since $x, v_x \in rS^{d-1}$, the maximum distance between two points on the sphere is the diameter $2r$). Thus,

$$\sup_{x \in rS^{d-1}} \left| \frac{1}{n} \sum_{i=1}^{n} \langle \overline{X}_i, x - v_x \rangle \right| = \sup_{t : \|t\| \leqslant 2r} \left| \frac{1}{n} \sum_{i=1}^{n} \langle \overline{X}_i, t \rangle \right| = 2r \sup_{t : \|t\| \leqslant 1} \left| \frac{1}{n} \sum_{i=1}^{n} \langle \overline{X}_i, t \rangle \right|.$$

The expectation of the supremum over the unit ball is

$$\mathbb{E} \sup_{t : \|t\| \leqslant 1} \left| \frac{1}{n} \sum_{i=1}^{n} \langle \overline{X}_i, t \rangle \right| \leqslant \sqrt{\mathbb{E} \sup_{t : \|t\| \leqslant 1} \left| \frac{1}{n} \sum_{i=1}^{n} \langle \overline{X}_i, t \rangle \right|^2} = \sqrt{\mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \overline{X}_i \right\|^2} = \frac{1}{n} \sqrt{\mathbb{E} \left\| \sum_{i=1}^{n} \overline{X}_i \right\|^2}.$$

Since $\mathbb{E} \left\| \sum_{i=1}^{n} \overline{X}_i \right\|^2 = \sum_{i=1}^{n} \mathbb{E} \left[\|\overline{X}_i\|^2 \right] = n \operatorname{tr}(\Sigma)$, we have

$$\mathbb{E} \sup_{t : \|t\| \leqslant 1} \left| \frac{1}{n} \sum_{i=1}^{n} \langle \overline{X}_i, t \rangle \right| \leqslant \frac{1}{n} \sqrt{n \operatorname{tr}(\Sigma)} = \sqrt{\frac{\operatorname{tr}(\Sigma)}{n}}.$$

Therefore, in the end,

$$(A) \leqslant 8r \sqrt{\frac{\operatorname{tr}(\Sigma)}{n}}.$$

Thus, putting the bound of (A) and (B) together, we have

$$\mathbb{E}[Y] \leqslant \frac{4}{r^2} \left(8r \sqrt{\frac{\operatorname{tr}(\Sigma)}{n}} + \frac{\varepsilon}{\sqrt{m}} \right) \leqslant \frac{1}{30} + \frac{1}{60} = \frac{1}{20},$$

provided that $r \geqslant 960 \sqrt{\frac{\operatorname{tr}(\Sigma)}{n}}$ and $\varepsilon = 2r \sqrt{\frac{\operatorname{tr}(\Sigma)}{k}}$.

To achieve $\mathbb{P}\{Y \leqslant 0.1\} \geqslant 1 - \delta/2$, we will use the bounded differences inequality in Theorem 9. The random variable Y depends on $\overline{X}_1, \ldots, \overline{X}_n$, and changing one $\overline{X}_i$ affects at most one block B_j. The indicator changes

by at most 1, so Y changes by at most $\frac{1}{k}$. Thus, by the bounded differences inequality, with differences bounded by $\frac{1}{k}$, the variance of Y is effectively controlled, and

$$\mathbb{P}\left\{Y > \mathbb{E}\left[Y\right] + t\right\} \leqslant \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}\left(\frac{1}{k}\right)^2}\right) \leqslant \exp\left(-\frac{2t^2 k}{n}\right).$$

Set $t = \frac{1}{20}$, so we need

$$\exp\left(-2\left(\frac{1}{20}\right)^2 \frac{k}{n}\right) \leqslant \frac{\delta}{2}.$$

This simplifies to

$$\exp\left(-\frac{k}{200n}\right) \leqslant \frac{\delta}{2}.$$

Since $k = \lceil 200\log(2/\delta)\rceil \leqslant 200\log(2/\delta)$, we have

$$\exp\left(-\frac{k}{200n}\right) \leqslant \exp\left(-\frac{\log(2/\delta)}{n}\right).$$

Therefore, we need

$$\left(\frac{\delta}{2}\right)^{\frac{1}{n}} \leqslant \frac{\delta}{2},$$

which is true.

This ensures that for every $x \in rS^{d-1}$, the condition

$$\left|\frac{1}{m}\sum_{i\in B_j}\langle \overline{X}_i, x - v_x\rangle\right| < \frac{r^2}{4}$$

holds for at least $9k/10$ blocks. Intersecting with the $8k/10$ blocks from Step 5, we get at least $7k/10$ blocks where both hold, completing the proof of Theorem 30. □

Proof of Theorem 29. Theorem 30 implies that, with probability at least $1-\delta$, for any a with $\|a-\mu\| \geqslant r$, $\|Z_j - a\| > \|Z_j - \mu\|$ for more than $k/2$ blocks. By definition of $\widehat{\mu}_n$,

$$\text{radius}(T_{\widehat{\mu}_n}) \leqslant \text{radius}(T_\mu).$$

Since $T_\mu = \{x : \|Z_j - x\| \leqslant \|Z_j - \mu\| \text{ for } |J| > k/2\}$, and $\mu \in T_\mu$, we have $\text{radius}(T_\mu) \leqslant r$. Thus, either $\mu \in T_{\widehat{\mu}_n}$, implying $\|\widehat{\mu}_n - \mu\| \leqslant r$, or $\widehat{\mu}_n \in T_\mu$, also implying $\|\widehat{\mu}_n - \mu\| \leqslant r$. Hence,

$$\|\widehat{\mu}_n - \mu\| \leqslant r,$$

as required. □

The multivariate MoM estimator achieves a sub-Gaussian bound under minimal assumptions, making it robust to heavy-tailed distributions. The constants (960 and 240) are chosen for proof clarity and are not necessarily optimal. The dependence on δ through k is necessary, as discussed in Chapter 2, unless stronger distributional assumptions are made.

Chapter 4

Trimmed Mean Estimator

This chapter investigates the trimmed mean estimator for mean estimation, a robust method that achieves optimal sub-Gaussian performance under adversarial contamination, as developed in Lugosi and Mendelson [LM21]. We present its construction, derive error bounds, and provide detailed proofs for both univariate and multivariate settings, building on the theoretical foundations of Chapter 2. Unlike the median-of-means estimator in Chapter 3, which assumes no contamination, the trimmed mean excels in handling heavy-tailed distributions and malicious noise, requiring only finite variance or second moments.

4.1 Introduction

The trimmed mean estimator is a key method in robust statistics, designed to estimate the mean of a random variable or vector when data may be affected by heavy-tailed distributions or adversarial contamination. As discussed in Chapter 2, real-world data often deviates from idealized assumptions, such as normality, due to outliers, measurement errors, or malicious tampering. The empirical mean, which averages all observations, is highly sensitive to such deviations, leading to poor performance in the presence of extreme values, as seen in heavy-tailed distributions like

the Pareto. In contrast, the trimmed mean estimator, developed by Lugosi and Mendelson [LM21], achieves optimal sub-Gaussian performance with minimal assumptions, ensuring robustness to both heavy-tailed distributions and adversarial noise, where a fraction of the data is arbitrarily altered.

Statistical methods rely on assumptions about the underlying distribution, such as independence, randomness, or a specific shape (e.g., Gaussian). However, these assumptions are often simplified models of complex realities, and small deviations can significantly impact classical estimators [Hub81]. For example, a dataset may include outliers—observations far from the true mean—due to natural variability, errors, or contamination. Such outliers can inflate the empirical mean's error, especially in heavy-tailed distributions where extreme values are more likely. To illustrate this sensitivity, consider an example from Tukey [Tuk60], which shows how even a tiny fraction of contaminated data can disrupt classical methods, highlighting the need for *distributional robustness*, where estimators remain stable despite small deviations from the assumed model.

Example 31 (Tukey's Contamination Model)**.** Suppose we have n observations $X_1, X_2, \ldots, X_n$, each measuring a quantity μ. With probability $1-\varepsilon$, an observation is "good" and follows a normal distribution $\mathcal{N}(\mu, \sigma^2)$, but with probability ε (e.g., $\varepsilon = 0.002$, or 2 bad observations per 1000), it is "bad" and follows $\mathcal{N}(\mu, 9\sigma^2)$, with a variance nine times larger. The overall distribution is a mixture

$$F(x) = (1-\varepsilon)\Phi\left(\frac{x-\mu}{\sigma}\right) + \varepsilon\Phi\left(\frac{x-\mu}{3\sigma}\right),$$

where $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-y^2/2}\,dy$ is the standard normal cumulative distribution function. This mixture models a scenario where most data points are accurate, but a few are outliers with larger errors, mimicking a slightly longer-tailed distribution.

Consider two classical measures of variability: the mean absolute deviation,

$$d_n = \frac{1}{n}\sum_{i=1}^{n} |X_i - \overline{X}|,$$

and the root mean square deviation,

$$s_n = \left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right]^{1/2},$$

where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the empirical mean. For a purely normal distribution ($\varepsilon = 0$), s_n converges to σ, the standard deviation, and is about 12% more efficient than d_n, which converges to $\sqrt{2/\pi}\sigma \approx 0.80\sigma$ [Hub81]. Efficiency here refers to how precisely a statistic estimates the true variability, with lower variance indicating higher efficiency.

To compare d_n and s_n, we use the *Asymptotic Relative Efficiency* (ARE), which measures the relative precision of two estimators as the sample size n grows large. For estimators of scale like d_n and s_n, the ARE of d_n relative to s_n is defined as the ratio of their normalized variances

$$\mathrm{ARE}(\varepsilon) = \lim_{n\to\infty} \frac{\mathrm{var}(s_n)/(\mathbb{E}\,[s_n])^2}{\mathrm{var}(d_n)/(\mathbb{E}\,[d_n])^2}.$$

A value of $\mathrm{ARE}(\varepsilon) > 1$ means d_n has lower normalized variance and is more efficient than s_n. For the mixture distribution, the ARE is

$$\mathrm{ARE}(\varepsilon) = \frac{1}{4}\left[\frac{3(1+8\varepsilon)}{(1+8\varepsilon)^2} - \frac{1}{\pi}\cdot\frac{(1+8\varepsilon)}{2(1+2\varepsilon)^2}\right]. \tag{4.1}$$

ε	$\mathrm{ARE}(\varepsilon)$	ε	$\mathrm{ARE}(\varepsilon)$
0.000	0.876	0.050	2.035
0.001	0.948	0.100	1.903
0.002	1.016	0.150	1.689
0.005	1.198	0.250	1.371
0.010	1.439	0.500	1.017
0.020	1.752	1.000	0.876

Table 4.1: Values of ε and corresponding $\mathrm{ARE}(\varepsilon)$

For $\varepsilon = 0.002$, $\mathrm{ARE}(\varepsilon) \approx 1.016$, meaning d_n becomes more efficient than s_n. For $\varepsilon = 0.05$, $\mathrm{ARE}(\varepsilon) \approx 2.035$, indicating d_n is twice as efficient.

This shows that even a small fraction of outliers can render s_n suboptimal, as bad observations inflate its variance due to the squared terms, while d_n is less affected by extreme values. In practice, datasets in fields like physics often have ε between 0.01 and 0.1, making robust methods like d_n preferable [Hub81].

Tukey's example highlights why contamination matters: even a tiny fraction of outliers can disrupt classical estimators, necessitating distributional robustness. The trimmed mean estimator addresses this by removing extreme values, ensuring stability when a fraction η of the data is adversarially corrupted [LM21]. Adversarial contamination models a worst-case scenario where an adversary can arbitrarily alter up to ηN observations, simulating errors, tampering, or outliers. Unlike the median-of-means (MoM) estimator in Chapter 3, which assumes no contamination and relies on partitioning data into blocks, the trimmed mean is designed for such adversarial settings, making it preferred when data reliability is uncertain or distributions deviate slightly from assumptions.

In the univariate case, the trimmed mean splits the sample into two parts: one estimates truncation levels to eliminate outliers, and the other computes a truncated average, requiring only finite variance [LM21]. This achieves sub-Gaussian performance, with an error bound that accounts for contamination, such as

$$|\widehat{\mu} - \mu| \leqslant 3\mathcal{E}(4\varepsilon, X) + 2\sigma\sqrt{\frac{\log(4/\delta)}{n}},$$

where $\mathcal{E}(\varepsilon, X)$ captures the tail behavior, and will be discussed later in the next section, and ε depends on η and the confidence parameter δ.

In the multivariate case, the estimator applies this approach to projections across all directions on the unit sphere S^{d-1}, intersecting slabs to form a robust estimate, achieving

$$\|\widehat{\mu} - \mu\| \leqslant c\left(\sqrt{\frac{\operatorname{tr}(\Sigma)}{n}} + \sqrt{\frac{\lambda_1 \log(1/\delta)}{n}} + \sqrt{\lambda_1 \eta}\right), \tag{4.2}$$

where Σ is the covariance matrix, λ_1 is its largest eigenvalue, and c is a constant [LM21]. This bound extends the sub-Gaussian performance of the MoM estimator, adding a $\sqrt{\lambda_1 \eta}$ term for contamination.

Why not clean the data by rejecting outliers before using classical methods? As noted in Huber [Hub81], such two-step approaches are problematic: identifying outliers in high-dimensional data is challenging without robust estimates, cleaned data may not follow the assumed distribution, and classical rejection rules often fail with multiple outliers, where one outlier may mask another. The trimmed mean avoids these issues by smoothly truncating extreme values, offering superior performance [LM21].

This chapter explores the trimmed mean estimator's robustness to contamination, contrasting with the contamination-free MoM estimator. Section 4.2 details the univariate estimator, its algorithm, error bound, and proof, while Section 4.3 covers the multivariate extension, including its construction, performance, and proof. We employ tools from Chapter 2, such as Bernstein's inequality and empirical process techniques, to derive sub-Gaussian guarantees under adversarial conditions, ensuring clarity for practical applications in statistics and machine learning. Notably, all proof techniques utilized in this chapter have been meticulously developed and thoroughly examined in the proofs presented in Chapter 3. Leveraging this established familiarity, we present the proofs here more concisely compared to the preceding chapter, focusing on key insights while maintaining rigor.

4.2 Univariate Case

In the univariate case ($d = 1$), the trimmed mean estimator estimates the mean $\mu = \mathbb{E}[X]$ of a random variable $X \in \mathbb{R}$ with finite variance $\sigma^2 = \mathbb{E}\left[(X - \mu)^2\right] < \infty$. Given a sample of $2n$ i.i.d. copies $X_1, \ldots, X_n, Y_1, \ldots, Y_n$, where up to $2\eta n$ points may be adversarially corrupted to produce $\widetilde{X}_1, \ldots, \widetilde{X}_n, \widetilde{Y}_1, \ldots, \widetilde{Y}_n$, the estimator uses one half of the sample to determine truncation levels and the other to compute a

robust average [LM21]. Unlike the median-of-means (MoM) estimator in Chapter 3, which assumes no contamination, the trimmed mean is designed to handle adversarial noise, making it robust to outliers and malicious tampering, requiring only finite variance as per Section 2.5.

We first establish a lower bound on the performance of any mean estimator under adversarial contamination, which sets the benchmark for optimality and highlights the challenge of robust estimation [LM21]. Define $\overline{X} = X - \mu$, and for $0 < p < 1$, the quantile

$$Q_p\left(\overline{X}\right) = \sup\left\{M \in \mathbb{R} : \mathbb{P}\left\{\overline{X} \geqslant M\right\} \geqslant 1 - p\right\}.$$

For simplicity of presentation, we assume throughout the article that X has an absolutely continuous distribution. That means,

$$\mathbb{P}\left\{\overline{X} \geqslant Q_p(\overline{X})\right\} \geqslant 1 - p.$$

Consider an adversary who can corrupt up to ηn points in each half of the sample. One strategy is to truncate the upper tail: replace X_i with $\widetilde{X}_i = \min\{X_i, \mu + Q_{1-\eta/2}(\overline{X})\}$. Since $X_i - \mu = \overline{X}$, we have:

$$\mathbb{P}\left\{\overline{X} \geqslant Q_{1-\eta/2}(\overline{X})\right\} = \frac{\eta}{2},$$

meaning $\frac{\eta}{2}$ fraction of the points are above $\mu + Q_{1-\eta/2}(\overline{X})$.

For n points, the expected number of points exceeding this threshold is $\frac{\eta}{2}n$. By a binomial tail bound as we did in the proof of univariate case in section 3.2.3, with probability at least $1 - 2\exp(-c\eta n)$, at most $\frac{3}{4}\eta n$ points X_i satisfy

$$X_i \geqslant \mu + Q_{1-\eta/2}(\overline{X}),$$

where $c > 0$ is a constant. Since $\frac{3}{4}\eta N \leqslant \eta N$, the adversary can corrupt all such points, producing a sample $\widetilde{X}_1, \ldots, \widetilde{X}_N$ indistinguishable from an uncorrupted sample drawn from

$$Z = \min\{X, \mu + Q_{1-\eta/2}(\overline{X})\}.$$

In other words, on this event, no procedure can distinguish between $\mu = \mathbb{E}[X]$ and $\mathbb{E}[Z]$. Therefore, the error caused by this action is at least $|\mathbb{E}[Z] - \mu|$, which is

$$\mathbb{E}[Z - \mu] = \mathbb{E}\left[\left(\overline{X} - Q_{1-\eta/2}(\overline{X})\right)\mathbf{1}_{\overline{X} \geqslant M}\right].$$

Since we can do the similar process for the lower tail of X, it follows that, with probability at least $1 - 2\exp(-c\eta n)$ no estimator can perform with accuracy better than

$$\overline{\mathcal{E}}(n, X) \stackrel{\text{def.}}{=} \max\left\{\mathbb{E}\Big[\big|\overline{X} - Q_{n/2}(X)\big|\mathbf{1}_{\overline{X}\leqslant Q_{n/2}(X)}\Big], \right.$$
$$\left.\mathbb{E}\Big[\big|\overline{X} - Q_{1-n/2}(X)\big|\mathbf{1}_{\overline{X}\geqslant Q_{1-n/2}(X)}\Big]\right\}.$$

Another trivial action for the adversary is to do nothing. In that case, we already know that the best estimator is the sub-Gaussian estimator. Therefore, if one wishes to find a procedure that performs with probability at least $1 - \delta - 2\exp(-c\eta N)$, the best error one can hope for is

$$\overline{\mathcal{E}}(n, X) + C\sigma\sqrt{\frac{\log(2/\delta)}{n}},$$

where $C > 0$ is a constant.

Our trimmed mean estimator is, in fact, almost the optimal error, with $\overline{\mathcal{E}}(n, X)$ replaced by

$$\mathcal{E}(n, X) \stackrel{\text{def.}}{=} \max\left\{\mathbb{E}\left[|X|\,\mathbf{1}_{X\leqslant Q_{n/2}(X)}\right], \mathbb{E}\left[|X|\,\mathbf{1}_{X\geqslant Q_{1-n/2}(X)}\right]\right\}.$$

4.2.1 Construction of the Estimator

Recall our setting: The trimmed mean estimator estimates the mean $\mu = \mathbb{E}[X]$ of a real-valued random variable X with finite variance $\sigma_X^2 < \infty$. Given $2n$ independent copies $X_1, \ldots, X_n, Y_1, \ldots, Y_n$, an adversary may corrupt up to $2\eta n$ points, producing $\widetilde{X}_1, \ldots, \widetilde{X}_n, \widetilde{Y}_1, \ldots, \widetilde{Y}_n$.

The estimator is constructed as follows

1. **Split the Sample**: Divide the $2n$ observations into two equal parts of size n: $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_n\}$. The adversary may corrupt up to ηn points in each part, yielding $\left\{\widetilde{X}_1, \ldots, \widetilde{X}_n\right\}$ and $\left\{\widetilde{Y}_1, \ldots, \widetilde{Y}_n\right\}$.

2. **Set Truncation Levels**: Using the second half $\left\{\widetilde{Y}_1, \ldots, \widetilde{Y}_n\right\}$, compute the nondecreasing rearrangement $\widetilde{Y}_1^* \leqslant \cdots \leqslant \widetilde{Y}_n^*$. Given the contamination parameter η and confidence level δ, define

$$\varepsilon = 8\eta + 12 \cdot \frac{\log(4/\delta)}{n}.$$

Set the truncation levels

$$\alpha = \widetilde{Y}_{\varepsilon n}^*, \quad \beta = \widetilde{Y}_{(1-\varepsilon)n}^*,$$

where $\widetilde{Y}_{\varepsilon n}^*$ is the εn-th smallest value and $\widetilde{Y}_{(1-\varepsilon)n}^*$ is the $(1-\varepsilon)n$-th largest value in the sorted sequence.

3. **Compute the Trimmed Mean**: Define the truncation function

$$\phi_{\alpha,\beta}(x) = \begin{cases} \beta & \text{if } x > \beta, \\ x & \text{if } x \in [\alpha, \beta], \\ \alpha & \text{if } x < \alpha. \end{cases}$$

Using the first half $\{\widetilde{X}_1, \ldots, \widetilde{X}_n\}$, compute the estimator

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \phi_{\alpha,\beta}(\widetilde{X}_i).$$

This construction uses the second half to identify truncation levels α and β, which exclude extreme values, and applies the truncation function to the first half to compute a robust mean, mitigating the adversary's impact [LM21].

4.2.2 Performance Bound

The univariate trimmed mean estimator achieves robust performance under adversarial contamination, offering sub-Gaussian error bounds with minimal assumptions, as established in Lugosi and Mendelson [LM21]. The following theorem quantifies its error, ensuring reliability for heavy-tailed distributions and up to ηn corrupted samples per half of the dataset, requiring only finite variance, as outlined in Section 2.5.

Theorem 32. *Let $\delta \in (0,1)$ such that $\delta \geqslant e^{-n}/4$. Given $2n$ i.i.d. samples $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ from a random variable $X \in \mathbb{R}$ with mean $\mu = \mathbb{E}[X]$ and variance $\sigma_X^2 < \infty$, where an adversary may corrupt up to $2\eta n$ points to produce $\widetilde{X}_1, \ldots, \widetilde{X}_n, \widetilde{Y}_1, \ldots, \widetilde{Y}_n$, the trimmed mean estimator $\widehat{\mu}$ satisfies, with probability at least $1 - \delta$*

$$|\widehat{\mu} - \mu| \leqslant 3\mathcal{E}(4\varepsilon, X) + 2\sigma_X \sqrt{\frac{\log(4/\delta)}{n}},$$

where $\varepsilon = 8\eta + 12 \cdot \frac{\log(4/\delta)}{n}$, and

$$\mathcal{E}(\varepsilon, X) = \max\left\{\mathbb{E}\left[\left|\overline{X}\right| \mathbf{1}_{\overline{X} \leqslant Q_{\varepsilon/2}(\overline{X})}\right], \mathbb{E}\left[\left|\overline{X}\right| \mathbf{1}_{\overline{X} \geqslant Q_{1-\varepsilon/2}(\overline{X})}\right]\right\},$$

with $\overline{X} = X - \mu$ and $Q_p(\overline{X}) = \sup\{M \in \mathbb{R} : \mathbb{P}\{\overline{X} \geqslant M\} \geqslant 1 - p\}$. Additionally, with probability at least $1 - 4\exp(-\varepsilon n/12)$,

$$|\widehat{\mu} - \mu| \leqslant 10\sqrt{\varepsilon}\sigma_X.$$

The theorem provides two error bounds, each offering insight into the estimator's performance under heavy-tailed and contaminated data

- **Primary Bound**: The bound $|\widehat{\mu} - \mu| \leqslant 3\mathcal{E}(4\varepsilon, X) + 2\sigma_X \sqrt{\frac{\log(4/\delta)}{n}}$ consists of two terms:
  - **Tail Expectation Term**: The term $3\mathcal{E}(4\varepsilon, X)$ captures the contribution of the distribution's tails. This term measures the expected magnitude of $\overline{X}$ in the lower and upper $\varepsilon/2$ tails, scaled by 3 to account for trimming and contamination effects. For heavy-tailed distributions, this term is small if the tails decay reasonably, but it grows for very heavy tails. The parameter $\varepsilon = 8\eta + 12 \cdot \frac{\log(4/\delta)}{n}$ balances contamination η and confidence δ, with larger η or smaller δ increasing the trimmed fraction.
  - **Variance Term**: The term $2\sigma_X \sqrt{\frac{\log(4/\delta)}{n}}$ reflects the variance's impact, decreasing with sample size n and increasing logarithmically with confidence $1/\delta$. It resembles sub-Gaussian bounds, capturing statistical error after trimming outliers. For example,

with $\sigma_X^2 = 1$, $n = 1000$, $\delta = 0.05$, we have $\log(4/0.05) \approx 4.38$, so this term is $2\sqrt{\frac{4.38}{1000}} \approx 0.133$, indicating a small error contribution for large samples.

For a numerical example, consider $n = 1000$, $\sigma_X^2 = 1$, $\eta = 0.01$, $\delta = 0.05$. Then, $\varepsilon \approx 0.133$. If X is standard normal, $Q_{0.066}(\overline{X}) \approx -1.5$, and $\mathcal{E}(4 \cdot 0.133, X) \approx \mathbb{E}\left[\left|\overline{X}\right| \mathbf{1}_{\overline{X} \leqslant -1.83}\right] \approx 0.05$, so the bound is approximately 0.283, ensuring accuracy despite 1% contamination.

- **Simpler Bound**: The bound $|\widehat{\mu} - \mu| \leqslant 10\sqrt{\varepsilon}\sigma_X$, with probability $1 - 4\exp(-\varepsilon n/12)$, simplifies the error to a single term proportional to $\sqrt{\varepsilon}\sigma_X$. Since $\varepsilon \approx 8\eta + 12 \cdot \frac{\log(4/\delta)}{n}$, this bound scales with the square root of contamination and confidence effects, offering a more interpretable form. In the example, $\sqrt{\varepsilon} \approx 0.365$, so the bound is approximatly 3.65, with probability 0.99995, indicating high confidence but a looser bound.

4.2.3 Proof of Performance Bound

We prove Theorem 32, establishing the performance bound for the univariate trimmed mean estimator under adversarial contamination. The proof follows Lugosi and Mendelson [LM21], with detailed steps to ensure clarity for readers familiar with the tools in Chapter 2. We aim to show that the estimator's error is bounded by a combination of a contamination term and a sub-Gaussian term, even when up to $2\eta n$ samples are corrupted.

Proof. The proof proceeds in four steps: first, we bound the error of the estimator on the uncorrupted sample $X_1, \ldots, X_n$; second, we show that the estimator on the uncorrupted sample satisfies the desired inequality; third, we demonstrate that adversarial corruption does not significantly increase the error; and fourth, we complete the theorem by applying the bound on $\mathcal{E}(\varepsilon, X)$.

Step 1. Bound the error of the estimator on the uncorrupted sample.

Define the event E where the following conditions hold simultaneously for the uncorrupted sample $Y_1, \ldots, Y_n$, with probability at least $1 - 4\exp(-\varepsilon N/12) \geqslant 1 - \delta/2$

(E1) $\left|\left\{i : Y_i \geqslant \mu + Q_{1-2\varepsilon}(\overline{X})\right\}\right| \geqslant \frac{3}{2}\varepsilon N$,

(E2) $\left|\left\{i : Y_i \leqslant \mu + Q_{1-\varepsilon/2}(\overline{X})\right\}\right| \geqslant \left(1 - \frac{3}{4}\varepsilon\right) N$,

(E3) $\left|\left\{i : Y_i \leqslant \mu + Q_{2\varepsilon}(\overline{X})\right\}\right| \geqslant \frac{3}{2}\varepsilon N$,

(E4) $\left|\left\{i : Y_i \geqslant \mu + Q_{\varepsilon/2}(\overline{X})\right\}\right| \geqslant \left(1 - \frac{3}{4}\varepsilon\right) N$.

These conditions ensure that the truncation levels α and β are well-positioned. For $U = \mathbf{1}_{\overline{X} \geqslant Q_{1-2\varepsilon}(\overline{X})}$, since $\mathbb{P}\left\{\overline{X} \geqslant Q_{1-2\varepsilon}(\overline{X})\right\} = 2\varepsilon$, the variance is

$$\sigma_U^2 \leqslant \mathbb{P}\left\{\overline{X} \geqslant Q_{1-2\varepsilon}(\overline{X})\right\} = 2\varepsilon.$$

Applying Bernstein's inequality, condition (E1) holds with probability of at least $1 - \exp(-\varepsilon N/12)$. Similar reasoning together with a union bound show that E hold with probability of at least $1 - 4\exp(-\varepsilon N/12) \geqslant 1 - \delta/2$.

Importantly: then event E only depents on the uncorrupted sample $Y_1, Y_2, \ldots, Y_n$. That means, any event on $X_1, X_2, \ldots, X_n$ will be independent of E.

Step 2. Show that $\frac{1}{n}\sum_{i=1}^n \phi_{\alpha,\beta}(X_i)$ satisfies an inequality of the wanted form.

On E, after corruption of at most $2\eta n$ points (noting $\eta \leqslant \varepsilon/8$), we have

$$\left|\left\{i : \widetilde{Y}_i \geqslant \mu + Q_{1-2\varepsilon}(\overline{X})\right\}\right| \geqslant \left(\frac{3}{2}\varepsilon - 2\eta\right) n \geqslant \varepsilon n,$$

and

$$\left|\left\{i : \widetilde{Y}_i \leqslant \mu + Q_{1-\varepsilon/2}(\overline{X})\right\}\right| \geqslant \left(1 - \frac{3}{4}\varepsilon - 2\eta\right) n \geqslant (1 - \varepsilon)n.$$

Therefore,

$$Q_{1-2\varepsilon}(\overline{X}) \leqslant \widetilde{Y}^*_{(1-\varepsilon)n} - \mu \leqslant Q_{1-\varepsilon/2}(\overline{X}). \tag{4.3}$$

Simililarly, on the event E, we also have

$$Q_{\varepsilon/2}(\overline{X}) \leqslant \widetilde{Y}^*_{\varepsilon n} - \mu \leqslant Q_{2\varepsilon}(\overline{X}). \tag{4.4}$$

Now, consider the uncorrupted sample $X_1, X_2, \ldots, X_n$, with noticing that

$$\alpha = \widetilde{Y}^*_{\varepsilon n}, \quad \beta = \widetilde{Y}^*_{(1-\varepsilon)n}.$$

Then, on the event E,

$$\frac{1}{n}\sum_{i=1}^{n} \phi_{\alpha,\beta}(X_i) \leqslant \frac{1}{n}\sum_{i=1}^{n} \phi_{\mu+Q_{2\varepsilon}(\overline{X}),\mu+Q_{1-\varepsilon/2}(\overline{X})}(X_i),$$

because of the two bounds (4.4) and (4.3).

Decompose this as

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} \phi_{\mu+Q_{2\varepsilon}(\overline{X}),\,\mu+Q_{1-\varepsilon/2}(\overline{X})}(X_i) \\ &= \mathbb{E}\left[\phi_{\mu+Q_{2\varepsilon}(\overline{X}),\,\mu+Q_{1-\varepsilon/2}(\overline{X})}(X)\right] \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\left(\phi_{\mu+Q_{2\varepsilon}(\overline{X}),\,\mu+Q_{1-\varepsilon/2}(\overline{X})}(X_i) - \mathbb{E}\left[\phi_{\mu+Q_{2\varepsilon}(\overline{X}),\,\mu+Q_{1-\varepsilon/2}(\overline{X})}(X)\right]\right). \end{aligned} \tag{4.5}$$

The first term of (4.5) is bounded above by

$$\mathbb{E}\phi_{\mu+Q_{2\varepsilon}(\overline{X}),\mu+Q_{1-\varepsilon/2}(\overline{X})}(X) \leqslant \mu + \mathbb{E}\left[\overline{X}\mathbf{1}_{\overline{X}\geqslant Q_{1-\varepsilon/2}(\overline{X})}\right] \leqslant \mu + \mathcal{E}(\varepsilon, X),$$

and below by

$$\mathbb{E}\phi_{\mu+Q_{2\varepsilon}(\overline{X}),\mu+Q_{1-\varepsilon/2}(\overline{X})}(X) \geqslant \mu - \mathbb{E}\left[\overline{X}\mathbf{1}_{\overline{X}\leqslant Q_{2\varepsilon}(\overline{X})}\right] \geqslant \mu - \mathcal{E}(4\varepsilon, X).$$

As a consequence, the second term of (4.5) is a sum of centered i.i.d. random variables, upper bounded by $Q_{1-\varepsilon/2}(\overline{X}) + \mathcal{E}(4\varepsilon, X)$, with variance

at most σ_X^2. By the Corollary 5 of Bernstein's inequality, conditioned on E, with probability at least $1 - \delta/4$, we have

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}\phi_{\alpha,\beta}(X_i) - (\mu + \mathcal{E}(\varepsilon, X)) \\ &\leqslant \sigma_X\sqrt{\frac{2\log(4/\delta)}{n}} + \frac{Q_{1-\varepsilon/2}(\overline{X})\log(4/\delta)}{n} + \frac{\mathcal{E}(4\varepsilon, X)\log(4/\delta)}{n}. \end{aligned} \tag{4.6}$$

Then, noticing that by Chebyshev's inequality (Theorem 2) and the definition of quantile,

$$\frac{\varepsilon}{2} = \mathbb{P}\left\{\overline{X} \geqslant Q_{1-\varepsilon/2}\left(\overline{X}\right)\right\} \leqslant \frac{\sigma_X^2}{Q_{1-\varepsilon/2}^2}, \tag{4.7}$$

or

$$Q_{1-\varepsilon/2} \leqslant \frac{\sigma_X\sqrt{2}}{\sqrt{\varepsilon}}.$$

Therfore,

$$Q_{1-\varepsilon/2}(\overline{X})\log(4/\delta)/n \leqslant \sigma_X\sqrt{\frac{\log(4/\delta)}{6n}}.$$

Moreover, since $\delta \geqslant e^{-n}/4$, we also have and

$$\mathcal{E}(4\varepsilon, X)\log(4/\delta)/n \leqslant \mathcal{E}(4\varepsilon, X).$$

Putting these two bounds to the right hand side of (4.6), we have

$$\frac{1}{n}\sum_{i=1}^{n}\phi_{\alpha,\beta}(X_i) \leqslant \mu + 2\mathcal{E}(4\varepsilon, X) + 2\sigma_X\sqrt{\frac{\log(4/\delta)}{n}}.$$

A symmetric argument for the lower tail yields, with probability at least $1 - \delta/2$,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\phi_{\alpha,\beta}(X_i) - \mu\right| \leqslant 2\mathcal{E}(4\varepsilon, X) + 2\sigma_X\sqrt{\frac{\log(4/\delta)}{n}}.$$

We achieved the goal of this step.

Step 3. Show that adversarial corruption does not significantly increase the error.

Now, we need to evaluate

$$\left|\frac{1}{n}\sum_{i=1}^{n}\phi_{\alpha,\beta}(X_i)-\frac{1}{n}\sum_{i=1}^{n}\phi_{\alpha,\beta}(\widetilde{X}_i)\right|.$$

Since $\phi_{\alpha,\beta}(X_i)\neq\phi_{\alpha,\beta}(\widetilde{X}_i)$ for at most $2\eta n$ indices, and the maximum difference is

$$\left|\phi_{\alpha,\beta}(X_i)-\phi_{\alpha,\beta}(\widetilde{X}_i)\right|\leqslant\left|Q_{\varepsilon/2}(\overline{X})\right|+\left|Q_{1-\varepsilon/2}(\overline{X})\right|,$$

And therefore,

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^{n}\phi_{\alpha,\beta}(X_i)-\frac{1}{n}\sum_{i=1}^{n}\phi_{\alpha,\beta}(\widetilde{X}_i)\right|\\&\leqslant 2\eta\left(\left|Q_{\varepsilon/2}(\overline{X})\right|+\left|Q_{1-\varepsilon/2}(\overline{X})\right|\right)\\&\leqslant\frac{\varepsilon}{2}\max\left\{\left|Q_{\varepsilon/2}(\overline{X})\right|,\left|Q_{1-\varepsilon/2}(\overline{X})\right|\right\},\end{aligned}$$

since $\eta\leqslant\varepsilon/8$.

Now, noting that

$$\frac{\varepsilon}{2}Q_{1-\varepsilon/2}(\overline{X})=\mathbb{E}\left[Q_{1-\varepsilon/2}(\overline{X})\mathbf{1}_{\overline{X}\geqslant Q_{1-\varepsilon/2}(\overline{X})}\right]\leqslant\mathbb{E}\left[\overline{X}\mathbf{1}_{\overline{X}\geqslant Q_{1-\varepsilon/2}(\overline{X})}\right],$$

on E, we obtain

$$\left|\frac{1}{n}\sum_{i=1}^{n}\phi_{\alpha,\beta}(X_i)-\frac{1}{n}\sum_{i=1}^{n}\phi_{\alpha,\beta}(\widetilde{X}_i)\right|\leqslant\mathcal{E}(\varepsilon,X).$$

Combining both bounds, we get

$$|\widehat{\mu}-\mu|\leqslant 3\mathcal{E}(4\varepsilon,X)+2\sigma_X\sqrt{\frac{\log(4/\delta)}{n}}.$$

Step 4. Complete the second part of the theorem.

The second statement of the theorem follows from the bound $\mathcal{E}(\varepsilon,X)\leqslant\sigma_X\sqrt{8\varepsilon}$, which is true since let $M=Q_{1-\varepsilon/2}\left(\overline{X}\right)$ then

$$\begin{aligned}\mathbb{E}\left[\left|\overline{X}\right|\mathbf{1}_{\overline{X}\geqslant M}\right]&\leqslant\sqrt{\mathbb{E}\left[\overline{X}^2\right]\mathbb{P}\left\{\overline{X}\geqslant M\right\}}\leqslant\sqrt{\sigma_X^2\mathbb{P}\left\{\overline{X}\geqslant M\right\}}\\&\overset{4.7}{\leqslant}\sigma_X\cdot\sqrt{\frac{\varepsilon}{2}}.\end{aligned}\tag{4.8}$$

Symmtric argument for $Q_{\varepsilon/2}\left(\overline{X}\right)$ completes the proof. □

4.3 Multivariate Case

The multivariate trimmed mean estimator extends the univariate approach from Section 4.2 to estimate the mean $\mu = \mathbb{E}[X]$ of a random vector $X \in \mathbb{R}^d$ with finite second moments, i.e., $\mathbb{E}\left[\|X\|^2\right] < \infty$, ensuring a well-defined covariance matrix $\Sigma = \mathbb{E}\left[(X-\mu)(X-\mu)^T\right]$.

The setting is similar to the previous case. Given $2n$ i.i.d. samples $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_n$, an adversary may corrupt up to ηn points in each half, producing $\widetilde{X}_1, \ldots, \widetilde{X}_n, \widetilde{Y}_1, \ldots, \widetilde{Y}_n$, with contamination parameter $\eta \in [0, 1)$. We denote $\overline{X} = X - \mu$, and let $\lambda_{\max}(\Sigma)$ be the largest eigenvalue of Σ. As in Lugosi and Mendelson [LM21], the estimator achieves sub-Gaussian performance under heavy-tailed distributions and adversarial contamination, requiring only the assumptions in Section 2.5. This section details the estimator's construction, followed by its performance bound and proof, using tools from Chapter 2.

4.3.1 Construction of the Estimator

The multivariate trimmed mean estimator generalizes the univariate approach by applying robust trimming to projections of the data along all directions in the unit sphere $S^{d-1} = \{v \in \mathbb{R}^d : \|v\| = 1\}$, then reconstructing the mean through geometric constraints. Similar to the univariate case, the algorithm uses one half of the sample to set truncation levels and the other to compute trimmed averages, intersecting the results to form the estimator. Below, we describe each step, providing intuition to make the process be easy to understand.

1. **Split the Sample**: Divide the $2n$ observations into two equal parts of size n: the first half, $\{X_1, \ldots, X_n\}$, and the second half, $\{Y_1, \ldots, Y_n\}$. An adversary may corrupt up to ηn points in each part, yielding $\left\{\widetilde{X}_1, \ldots, \widetilde{X}_n\right\}$ and $\left\{\widetilde{Y}_1, \ldots, \widetilde{Y}_n\right\}$.

2. **Set Truncation Parameter**: Given the contamination parameter $\eta \in [0, 1)$ and confidence level $\delta \in (0, 1)$, define

$$\varepsilon = \max\left(10\eta, 2560 \cdot \frac{\log(2/\delta)}{n}\right).$$

3. **Compute Directional Truncation Levels**: For each unit vector $v \in S^{d-1}$, compute the projections $\langle \widetilde{Y}_i, v\rangle$ for $i = 1, \ldots, n$, and sort them in nondecreasing order: $\langle \widetilde{Y}_1, v\rangle^* \leqslant \cdots \leqslant \langle \widetilde{Y}_n, v\rangle^*$. Define truncation levels

$$\alpha_v = \langle \widetilde{Y}_i, v\rangle^*_{(\varepsilon/2)n}, \quad \beta_v = \langle \widetilde{Y}_i, v\rangle^*_{(1-\varepsilon/2)n}.$$

Here we see that $\langle \widetilde{Y}_i, v\rangle^*_{(\varepsilon/2)n}$ is the $(\varepsilon/2)n$-th smallest projection, and $\langle \widetilde{Y}_i, v\rangle^*_{(1-\varepsilon/2)n}$ is the $(1-\varepsilon/2)n$-th largest.

4. **Define Directional Trimmed Means**: For each $v \in S^{d-1}$ and a tuning parameter $Q > 0$, compute

$$U_Q(v) = \frac{1}{n}\sum_{i=1}^{n} \phi_{\alpha_v - Q, \beta_v + Q}\left(\langle \widetilde{X}_i, v\rangle\right),$$

where the truncation function is

$$\phi_{\alpha_v - Q, \beta_v + Q}(x) = \begin{cases} \beta_v + Q & \text{if } x > \beta_v + Q, \\ x & \text{if } x \in [\alpha_v - Q, \beta_v + Q], \\ \alpha_v - Q & \text{if } x < \alpha_v - Q. \end{cases}$$

Intuition: Using the first half of the data, we project $\widetilde{X}_i$ onto v and apply the univariate truncation function, but with an interval widened by Q. This widening accounts for the complexity of handling all directions in S^{d-1}, as high-dimensional data requires broader robustness to ensure consistency across projections. If a projection $\langle \widetilde{X}_i, v\rangle$ is an outlier (e.g., due to contamination), it is capped at $\beta_v + Q$ or $\alpha_v - Q$, preventing it from skewing the average. The result, $U_Q(v)$, is a robust estimate of $\mathbb{E}[\langle X, v\rangle]$, analogous to the univariate trimmed mean.

5. **Form Geometric Constraints**: For each $v \in S^{d-1}$ and $Q > 0$, let

$$\Gamma(v, Q) = \left\{x \in \mathbb{R}^d : |\langle x, v\rangle - U_Q(v)| \leqslant 2\varepsilon Q\right\}.$$

Then, define the intersection

$$\Gamma(Q) = \bigcap_{v \in S^{d-1}} \Gamma(v, Q).$$

Intuition: Each slab $\Gamma(v, Q)$ is a strip in $\mathbb{R}^d$ containing points x whose projection onto v is close to the trimmed mean $U_Q(v)$, within a width of $2\varepsilon Q$. This width depends on ε, reflecting the contamination and confidence levels, and Q, which adjusts for dimensionality. Intersecting these slabs over all directions forms $\Gamma(Q)$, a region that constrains the estimator to points consistent with robust projections. Picture $\Gamma(Q)$ as a compact shape in $\mathbb{R}^d$, like a polygon in 2D, formed by overlapping strips, with the true mean μ likely near its center due to the robustness of $U_Q(v)$.

6. **Select the Estimator**: Let $i^* \in \mathbb{Z}$ be the smallest integer such that $\bigcap_{i \geqslant i^*} \Gamma\left(2^i\right) \neq \varnothing$. Define:

$$\widehat{\mu} \in \bigcap_{i \in \mathbb{Z}, i \geqslant i^*} \Gamma\left(2^i\right).$$

If multiple points satisfy this, choose any one.

Intuition: The tuning parameter Q is tested at powers of 2 to find the smallest i^* where the slabs intersect, ensuring $\Gamma(2^{i^*})$ is non-empty. This adaptive selection, inspired by Lepski's method [LM21], balances robustness and precision: a small Q may yield an empty intersection, while a large Q widens the slabs unnecessarily. Selecting $\widehat{\mu}$ from the tightest non-empty intersection ensures the estimator is close to μ, robust to both outliers and contamination. we can think of this as trying different-sized nets to catch the true mean, choosing the smallest net that works.

This construction extends the univariate trimmed mean by applying robust trimming to projections and using geometric intersections to reconstruct the mean. It leverages the robustness of the univariate approach

while addressing the challenges of high dimensions, making it suitable for heavy-tailed and contaminated data.

4.3.2 Performance Bound

The multivariate trimmed mean estimator, constructed in the previous subsection, provides robust mean estimation for a random vector $X \in \mathbb{R}^d$ under heavy-tailed distributions and adversarial contamination. Its error, measured by the Euclidean norm $\|\widehat{\mu} - \mu\|$, exhibits sub-Gaussian behavior, where the probability of large errors decays exponentially, as discussed in Chapter 2 (Section 2.4). This contrasts with the empirical mean, which, as shown in Section 2.5, has poor error bounds for heavy-tailed data due to Chebyshev's inequality. The performance bound, adapted from Lugosi and Mendelson [LM21], quantifies the estimator's ability to handle sample size, dimensionality, variance, and contamination, making it ideal for high-dimensional, noisy datasets.

The following theorem states the estimator's performance, using the covariance matrix $\Sigma = \mathbb{E}\left[(X-\mu)(X-\mu)^T\right]$, its trace $\operatorname{tr}(\Sigma)$, and its largest eigenvalue $\lambda_{\max}(\Sigma)$, as defined in Section 4.3. It extends the univariate bound from Section 4.2, accounting for the challenges of high-dimensional data and adversarial noise.

Theorem 33. *Let $\delta \in (0,1)$, $\eta \in [0,1)$, and consider the multivariate trimmed mean estimator $\widehat{\mu}$ defined in Section 4.3, based on $2n$ i.i.d. samples $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_n \in \mathbb{R}^d$ with mean μ and covariance matrix Σ, where up to ηn points in each half may be corrupted. There exists a universal constant $c > 0$ such that, with probability at least $1-\delta$,*

$$\|\widehat{\mu} - \mu\| \leqslant c\left(\sqrt{\frac{\operatorname{tr}(\Sigma)}{n}} + \sqrt{\frac{\lambda_{\max}(\Sigma)\log(1/\delta)}{n}} + \sqrt{\lambda_{\max}(\Sigma)\eta}\right).$$

The error bound consists of three terms, where the first two terms are similar to the MoM estimator in Theorem 29.

- **Average Variance Term**: The term $\sqrt{\frac{\operatorname{tr}(\Sigma)}{n}}$ captures the average variance across all d dimensions, where $\operatorname{tr}(\Sigma) = \sum_{i=1}^{d} \mathbb{E}\left[(X_i - \mu_i)^2\right]$ sums the variances of X's coordinates. It decreases with larger n, reflecting that more samples reduce variability. For example, in a 3D dataset ($d = 3$) with $n = 1000$ and $\operatorname{tr}(\Sigma) = 3$ (unit variance per dimension), this term is approximately $\sqrt{\frac{3}{1000}} \approx 0.055$, indicating a small error contribution when the sample size is large.

- **Directional Variance Term**: The term $\sqrt{\frac{\lambda_{\max}(\Sigma)\log(1/\delta)}{n}}$ accounts for the worst-case variance in any direction, where $\lambda_{\max}(\Sigma)$ is the largest eigenvalue of Σ. The $\log(1/\delta)$ factor ensures uniform control over all directions in S^{d-1}, with smaller δ (higher confidence) increasing the bound logarithmically. In the example, if $\lambda_{\max}(\Sigma) = 1$ and $\delta = 0.05$, then $\log(1/0.05) \approx 3$, so this term is approximately $\sqrt{\frac{1 \cdot 3}{1000}} \approx 0.055$, comparable to the average variance but sensitive to confidence.

- **Contamination Term**: The term $\sqrt{\lambda_{\max}(\Sigma)\eta}$ measures the impact of adversarial contamination, where η is the fraction of corrupted points in each half. Independent of n, it scales with $\sqrt{\eta}$, showing robustness to small contamination levels. If $\eta = 0.01$ (1% contamination) and $\lambda_{\max}(\Sigma) = 1$, this term is $\sqrt{1 \cdot 0.01} = 0.1$, slightly larger than the other terms, emphasizing contamination's effect in noisy settings.

The bound resembles the multivariate MoM estimator's performance (Theorem 29)

$$\|\widehat{\mu}_{\mathrm{MoM}} - \mu\| \leqslant c' \left(\sqrt{\frac{\operatorname{tr}(\Sigma)}{n}} + \sqrt{\frac{\lambda_{\max}(\Sigma)\log(1/\delta)}{n}} \right).$$

The trimmed mean's additional $\sqrt{\lambda_{\max}(\Sigma)\eta}$ term reflects its robustness to contamination, unlike the MoM, which assumes clean data (Chapter 3). When $\eta = 0$, the bounds are equivalent up to constants, confirming the trimmed mean's competitiveness in contamination-free settings. This makes it valuable for applications like sensor data processing, where errors may corrupt data, or financial analysis, where extreme values are common.

The bound guarantees that $\widehat{\mu}$ is close to μ with high probability, even for heavy-tailed data (requiring only finite second moments) and up to ηn corrupted points per half. The projection-based trimming and slab intersections, as constructed, mitigate outliers and noise across all directions, yielding a tight error bound that scales well with n, d, and η. For a 10D dataset with $d = 10$, $n = 1000$, $\operatorname{tr}(\Sigma) = 10$, $\lambda_{\max}(\Sigma) = 1$, $\eta = 0.01$, and $\delta = 0.05$, the bound is approximately $c \cdot (0.1 + 0.055 + 0.1) \approx 0.255c$, demonstrating precision despite contamination and high dimensionality.

The proof of Theorem 33, using concentration inequalities and empirical process techniques from Chapter 2 (e.g., Bernstein's inequality, Talagrand's inequality, dual Sudakov inequality), will be presented in the next subsection. This bound highlights the trimmed mean's robustness, making it a powerful tool for high-dimensional, noisy data.

4.3.3 Proof of Performance Bound

In this subsection, we prove Theorem 33, which establishes the sub-Gaussian performance of the multivariate trimmed mean estimator under heavy-tailed distributions and adversarial contamination. The estimator, constructed in Section 4.3, projects the data onto all directions in the unit sphere $S^{d-1} = \{v \in \mathbb{R}^d : \|v\| = 1\}$, applies univariate trimming, and intersects the resulting slabs to estimate the mean μ. Proving the error bound

$$\|\widehat{\mu} - \mu\| \leqslant c \left(\sqrt{\frac{\operatorname{tr}(\Sigma)}{n}} + \sqrt{\frac{\lambda_{\max}(\Sigma) \log(2/\delta)}{n}} + \sqrt{\lambda_{\max}(\Sigma)\eta} \right)$$

requires controlling the estimator's behavior uniformly across all directions, accounting for both statistical fluctuations and up to ηn corrupted points per sample half. The proof, adapted from Lugosi and Mendelson [LM21], leverages concentration inequalities and empirical process techniques from Chapter 2, including Bernstein's inequality (Theorem 4), Talagrand's inequality (Theorem 6), symmetrization (Theorem 16), and the contraction lemma (Theorem 18).

The proof proceeds in two main parts. First, we establish a proposition showing that for an appropriately chosen tuning parameter Q, the slab intersection $\Gamma(Q)$ is non-empty and contains points close to μ. Second, we extend this result to the estimator $\widehat{\mu}$, which is selected from the smallest non-empty intersection $\Gamma(2^{i^*})$. The proposition relies on two lemmas: one bounding the truncation levels α_v and β_v, and another controlling the deviation of the trimmed projection estimates $U_Q(v)$.

Proof of Theorem 33. We aim to show that, with probability at least $1-\delta$, the estimator $\widehat{\mu}$ satisfies:

$$\|\widehat{\mu} - \mu\| \leqslant c\left(\sqrt{\frac{\operatorname{tr}(\Sigma)}{n}} + \sqrt{\frac{\lambda_{\max}(\Sigma)\log(2/\delta)}{n}} + \sqrt{\lambda_{\max}(\Sigma)\eta}\right),$$

where $c > 0$ is a universal constant, $\Sigma = \mathbb{E}\left[(X-\mu)(X-\mu)^T\right]$, $\operatorname{tr}(\Sigma)$ is its trace, $\lambda_{\max}(\Sigma)$ is its largest eigenvalue, and $\eta \in [0,1)$ is the contamination fraction.

Step 1: Establish the Key Proposition.

We begin with a proposition that identifies a tuning parameter Q for which the slab intersection $\Gamma(Q) = \bigcap_{v \in S^{d-1}} \Gamma(v, Q)$, where $\Gamma(v, Q) = \left[x \in \mathbb{R}^d : |\langle x, v\rangle - U_Q(v)| \leqslant 2\varepsilon Q\right]$, is non-empty and contains points close to μ. This is the core technical result, adapted from Proposition 1 in Lugosi and Mendelson [LM21].

Proposition 34. *Define the truncation parameter*

$$\varepsilon = \max\left(10\eta, 2560 \cdot \frac{\log(2/\delta)}{n}\right),$$

and the tuning parameter

$$Q_0 = \max\left(\frac{256}{\varepsilon}\sqrt{\frac{\operatorname{tr}(\Sigma)}{n}}, 16\sqrt{\frac{\lambda_{\max}(\Sigma)}{\varepsilon}}\right).$$

For any $Q \in [2Q_0, 4Q_0]$, with probability at least $1 - 2\exp(-\varepsilon n/2560) \geqslant 1-\delta$, the set $\Gamma(Q) \neq \varnothing$, and for every $z \in \Gamma(Q)$,

$$\|z - \mu\| \leqslant 4\varepsilon Q_0.$$

The proposition ensures that for a carefully chosen Q, the slabs $\Gamma(v, Q)$, which constrain the estimator's projections to be near the trimmed means $U_Q(v)$, intersect to form a non-empty region $\Gamma(Q)$ containing points close to μ. The parameter Q_0 balances the statistical error ($\sqrt{\mathrm{tr}\,(\Sigma)/n}$) and contamination effects ($\sqrt{\lambda_{\max}(\Sigma)\eta}$), while ε determines the trimming fraction, accounting for both η and δ. The high-probability guarantee comes from controlling errors uniformly over S^{d-1}, a challenge we address with empirical process techniques.

Step 2: Prove the Proposition via Two Lemmas.

The proof of Proposition 34 relies on two lemmas that control the truncation levels and projection estimates, ensuring that $\Gamma(Q)$ is well-behaved. We present these as substeps, following Lugosi and Mendelson [LM21]'s Lemmas 1 and 2.

Substep 2.1: Bound Truncation Levels.

We need to ensure that the truncation levels $\alpha_v = \langle \widetilde{Y}_i, v\rangle^*_{(\varepsilon/2)n}$ and $\beta_v = \langle \widetilde{Y}_i, v\rangle^*_{(1-\varepsilon/2)n}$ do not deviate too far from $\langle \mu, v\rangle$, even with contamination.

Lemma 35. *For each $i \in [1, \dots, n]$ and $v \in S^{d-1}$, define $\overline{Y}_i(v) = \langle Y_i - \mu, v\rangle$. With probability at least $1 - \exp(-\varepsilon n/2560) \geqslant 1 - \delta/2$,*

$$\sup_{v\in S^{d-1}} \left|\left\{i : \overline{Y}_i(v) \geqslant Q_0\right\}\right| \leqslant \frac{\varepsilon}{8}n, \quad \textit{and} \quad \sup_{v\in S^{d-1}} \left|\left\{i : \overline{Y}_i(v) \leqslant -Q_0\right\}\right| \leqslant \frac{\varepsilon}{8}n.$$

Similar to the event E in the proof of univariate case, this lemma ensures that, in any direction v, at most $\varepsilon n/8$ uncorrupted points have projections $\langle Y_i - \mu, v\rangle$ exceeding Q_0 in magnitude. This bounds the tail behavior uniformly, allowing the truncation levels α_v, β_v to remain close to $\langle \mu, v\rangle$, even after ηn corruptions.

Proof. We prove the first inequality, the second is similar. We need to bound the proportion of projections $\overline{Y}_i(v) = \langle Y_i - \mu, v\rangle$ exceeding Q_0, uniformly over $v \in S^{d-1}$, which is controlled by

$$\sup_{v\in S^{d-1}} \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\left\{\overline{Y}_i(v)\geqslant Q_0\right\}}. \tag{4.9}$$

Define a Lipschitz function $\chi : \mathbb{R} \to \mathbb{R}$:

$$\chi(x) = \begin{cases} 0 & \text{if } x \leqslant Q_0/2, \\ \frac{2x}{Q_0} - 1 & \text{if } x \in (Q_0/2, Q_0], \\ 1 & \text{if } x > Q_0, \end{cases}$$

with Lipschitz constant $2/Q_0$, satisfying

$$\mathbf{1}_{\left\{\overline{Y}_i(v) \geqslant Q_0\right\}} \leqslant \chi(\overline{Y}_i(v)) \leqslant \mathbf{1}_{\left\{\overline{Y}_i(v) \geqslant Q_0/2\right\}}.$$

Thus, (4.9) is bounded as

$$\sup_{v \in S^{d-1}} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\left\{\overline{Y}_i(v) \geqslant Q_0\right\}} \leqslant \sup_{v \in S^{d-1}} \frac{1}{n} \sum_{i=1}^{n} \chi(\overline{Y}_i(v)).$$

We bound the expectation of the left hand side.

$$\begin{aligned} &\mathbb{E} \sup_{v \in S^{d-1}} \frac{1}{n} \sum_{i=1}^{n} \chi(\overline{Y}_i(v)) \\ &\leqslant \mathbb{E}\left[\chi(\overline{Y}_i(v))\right] + \mathbb{E} \sup_{v \in S^{d-1}} \frac{1}{n} \sum_{i=1}^{n} \left(\chi(\overline{Y}_i(v)) - \mathbb{E}\left[\chi(\overline{Y}_i(v))\right]\right). \end{aligned} \tag{4.10}$$

For the first term, by the definition of $\chi(\overline{Y}_i(v))$ we have

$$\mathbb{E}\left[\chi(\overline{Y}_i(v))\right] \leqslant \mathbb{P}\left\{\overline{Y}_i(v) \geqslant Q_0/2\right\}.$$

Using Chebyshev's inequality (Theorem 2), with noticing that $Q_0 \geqslant 16\sqrt{\frac{\lambda_{\max}(\Sigma)}{\varepsilon}}$,

$$\begin{aligned} \mathbb{P}\left\{\overline{Y}_i(v) \geqslant Q_0/2\right\} = \mathbb{P}\left\{\langle X, v\rangle \geqslant Q_0/2\right\} &\leqslant \frac{4\mathbb{E}\left[\langle X, v\rangle^2\right]}{Q_0^2} \\ &\leqslant \frac{4\lambda_{\max}(\Sigma)}{Q_0^2} \leqslant \frac{4\lambda_{\max}(\Sigma)}{\left(16\sqrt{\frac{\lambda_{\max}(\Sigma)}{\varepsilon}}\right)^2} = \frac{\varepsilon}{64}, \end{aligned}$$

completes bounding for the first term of (4.10).

For the deviation term of (4.10), apply the symmetrization inequality (Theorem 16), we have

$$\mathbb{E}\sup_{v\in S^{d-1}}\frac{1}{n}\sum_{i=1}^{n}\left(\chi(\overline{Y}_i(v))-\mathbb{E}\left[\chi(\overline{Y}_i(v))\right]\right)\leqslant 2\mathbb{E}\sup_{v\in S^{d-1}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\chi(\overline{Y}_i(v)),$$

where ε_i are Rademacher variables. Since χ is Lipschitz with constant $2/Q_0$, the contraction lemma (Theorem 18) with reasoning similar to the proof of Theorem 30 gives

$$\begin{aligned}\mathbb{E}\sup_{v\in S^{d-1}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\chi(\overline{Y}_i(v)) &\leqslant \frac{2}{Q_0}\mathbb{E}\sup_{v\in S^{d-1}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\overline{Y}_i(v)\\ &\leqslant \frac{2}{Q_0}\mathbb{E}\sup_{v\in S^{d-1}}\frac{1}{n}\sum_{i=1}^{n}\left|\varepsilon_i\overline{Y}_i(v)\right|\\ &\leqslant \frac{2}{Q_0}\sqrt{\frac{\operatorname{tr}(\Sigma)}{n}}\leqslant\frac{\varepsilon}{128},\end{aligned}\tag{4.11}$$

using $Q_0\geqslant\frac{256}{\varepsilon}\sqrt{\frac{\operatorname{tr}(\Sigma)}{n}}$. Thus,

$$\mathbb{E}\sup_{v\in S^{d-1}}\frac{1}{n}\sum_{i=1}^{n}\chi(\overline{Y}_i(v))\leqslant\frac{\varepsilon}{64}+\frac{\varepsilon}{128}=\frac{3\varepsilon}{128}\leqslant\frac{\varepsilon}{32}.$$

Now, apply Talagrand's inequality (Theorem 6) to

$$Z=\sup_{v\in S^{d-1}}\frac{1}{n}\sum_{i=1}^{n}\chi(\overline{Y}_i(v)),$$

we have

$$\mathbb{P}\left\{Z\geqslant\frac{\varepsilon}{16}+\sqrt{\frac{x}{n}\cdot\frac{\varepsilon}{64}}+\frac{10x}{n}\right\}\leqslant e^{-x}.$$

Set $x=\varepsilon n/2560$. The bound becomes:

$$\frac{\varepsilon}{16}+\sqrt{\frac{\varepsilon n/2560}{n}\cdot\frac{\varepsilon}{64}}+\frac{10\cdot\varepsilon/2560}{n}\leqslant\frac{\varepsilon}{16}+\frac{\sqrt{\varepsilon^2/163840}}{\sqrt{n}}+\frac{\varepsilon}{256n}\leqslant\frac{\varepsilon}{8},$$

since the additional terms are small for large n. Thus, $\mathbb{P}\{Z\geqslant\varepsilon/8\}\leqslant\exp(-\varepsilon n/2560)$, proving the lemma since (4.9) is bounded by Z. □

Substep 2.2: Control Projection Deviations.

We show that the trimmed projection estimates $U_Q(v)$ are uniformly close to $\langle \mu, v \rangle$.

Lemma 36. *For $Q \in [2Q_0, 4Q_0]$, conditioned on the event E where Lemma 35 holds, with probability at least $1 - 2\exp(-\varepsilon n/2560) \geqslant 1 - \delta/2$,*

$$\sup_{v \in S^{d-1}} |U_Q(v) - \langle \mu, v \rangle| \leqslant 2\varepsilon Q.$$

Proof. We prove that

$$\sup_{v \in S^{d-1}} (U_Q(v) - \langle \mu, v \rangle) \leqslant 2\varepsilon Q$$

holds with the wanted probability, the other case will follow by similar arguments.

On event E, Lemma 35 implies that at most $\varepsilon n/8$ uncorrupted points have $\langle Y_i - \mu, v \rangle \geqslant Q_0$ or $\leqslant -Q_0$. Since $\eta n \leqslant \varepsilon n/10$ corrupted points may shift the quantiles, the truncation levels satisfy

$$\alpha_v - \langle \mu, v \rangle \geqslant -Q_0 \quad \text{and} \quad \beta_v - \langle \mu, v \rangle \leqslant Q_0.$$

Thus, the truncation interval is

$$\langle \mu, v \rangle - Q_0 \geqslant \alpha_v - Q \geqslant -Q_0 - Q \geqslant \langle \mu, v \rangle - 5Q_0, \tag{4.12}$$

and

$$\langle \mu, v \rangle + Q_0 \leqslant \beta_v + Q \leqslant \langle \mu, v \rangle + Q_0 + Q \leqslant \langle \mu, v \rangle + 5Q_0 \tag{4.13}$$

since $Q \leqslant 4Q_0$. That means, the range of $\phi_{\alpha_v - Q, \beta_v + Q}$ is at most $10Q_0$. And thus, the difference between corrupted and uncorrupted samples is bounded as

$$\left| \frac{1}{n} \sum_{i=1}^{n} \left(\phi_{\alpha_v - Q, \beta_v + Q} \left(\langle \widetilde{X}_i, v \rangle \right) - \phi_{\alpha_v - Q, \beta_v + Q} \left(\langle X_i, v \rangle \right) \right) \right| \leqslant \eta \cdot 10Q_0 \leqslant \varepsilon Q,$$

since $\eta \leqslant \varepsilon/10$.

However, by (4.12) and (4.13), we have

$$\begin{aligned} U_Q(v) &= \frac{1}{n}\sum_{i=1}^{n} \phi_{\alpha_v - Q, \beta_v + Q}\left(\langle \widetilde{X}_i, v\rangle\right) \\ &\leqslant \frac{1}{n}\sum_{i=1}^{n} \phi_{\langle \mu, v\rangle - Q_0, \langle \mu, v\rangle + 5Q_0}\left(\langle X_i, v\rangle\right). \end{aligned}$$

Notice that E only depends on $Y_1, Y_2, \ldots, Y_n$, the right hand side of the above inequality is independent of E.

Define

$$\begin{aligned} \overline{U}_Q(v) &= \frac{1}{n}\sum_{i=1}^{n} \phi_{\langle \mu, v\rangle - Q_0, \langle \mu, v\rangle + 5Q_0}\left(\langle X_i, v\rangle\right) - \langle \mu, v\rangle \\ &= \frac{1}{n}\sum_{i=1}^{n} \phi_{-Q_0, 5Q_0}\left(\langle X_i - \mu, v\rangle\right). \end{aligned}$$

To prove the lemma, it suffices to show that

$$\sup_{v \in S^{d-1}} \overline{U}_Q(v) \leqslant \varepsilon Q. \tag{4.14}$$

Let's decompose

$$\sup_{v \in S^{d-1}} \overline{U}_Q(v) \leqslant \sup_{v \in S^{d-1}} \left(\overline{U}_Q(v) - \mathbb{E}\overline{U}_Q(v)\right) + \sup_{v \in S^{d-1}} \mathbb{E}\overline{U}_Q(v). \tag{4.15}$$

For the second term of (4.15), we have

$$\begin{aligned} &\mathbb{E}\phi_{-Q_0, 5Q_0}\left(\langle X - \mu, v\rangle\right) \\ =&\mathbb{E}\left[\phi_{-Q_0, 5Q_0}\left(\langle X - \mu, v\rangle\right) - \langle X - \mu, v\rangle\right] \\ \leqslant&\mathbb{E}\left|-Q_0 - \langle X - \mu, v\rangle\right| \mathbf{1}_{\langle X - \mu, v\rangle \leqslant -Q_0} + \mathbb{E}\left|5Q_0 - \langle X - \mu, v\rangle\right| \mathbf{1}_{\langle X - \mu, v\rangle \geqslant 5Q_0}. \end{aligned}$$

Then, using similar arguments as (4.8) and (4.11), with noticing that $2Q_0 \leqslant Q$, we can bound $\mathbb{E}\overline{U}_Q(v)$ as

$$\mathbb{E}\overline{U}_Q(v) \leqslant \frac{\varepsilon Q}{64}. \tag{4.16}$$

The deviation in is a little bit more complicated to bound. Let's define

$$\overline{W}_Q(v) = \frac{1}{n}\sum_{i=1}^{n} \phi_{-3Q,3Q}\left(\langle X_i - \mu, v\rangle\right).$$

We have

$$\begin{aligned}
&\sup_{v\in S^{d-1}} \left(\overline{U}_Q(v) - \mathbb{E}\left[\overline{U}_Q(v)\right]\right) \\
\leqslant\ &\sup_{v\in S^{d-1}} \left|\overline{U}_Q(v) - \overline{W}_Q(v)\right| + \sup_{v\in S^{d-1}} \left|\overline{W}_Q(v) - \mathbb{E}\left[\overline{W}_Q(v)\right]\right| \\
&+ \sup_{v\in S^{d-1}} \left|\mathbb{E}\left[\overline{W}_Q(v)\right] - \mathbb{E}\left[\overline{U}_Q(v)\right]\right| \\
:=&(a) + (b) + (c). \qquad (4.17)
\end{aligned}$$

To bound (a), notice that $Q \in [2Q_0, 4Q_0]$ so $[-Q_0, 5Q_0] \subset [-3Q, 3Q]$. This means that $\phi_{-3Q,3Q}(x) \neq \phi_{-Q_0,5Q_0}(x)$ only if $x \notin [-Q_0, 5Q_0]$. And in that scenario,

$$|\phi_{-3Q,3Q}(x) - \phi_{-Q_0,5Q_0}(x)| \leqslant 3Q. \qquad (4.18)$$

By Lemma 35, we have

$$|i : \langle X_i - \mu, v\rangle \notin [-Q_0, 5Q_0]| \leqslant \frac{\varepsilon n}{4}.$$

And then, from (4.18), we get

$$(a) \leqslant \frac{3\varepsilon Q}{4}.$$

Using similar arguments to bound (c), we get

$$(c) \leqslant 3Q\mathbb{P}\left\{|\langle X - \mu, v\rangle| > 5Q_0\right\} \leqslant \frac{3Q\lambda_{\max}(\Sigma)}{25Q_0^2} \leqslant \frac{3\varepsilon Q}{64}.$$

Now, we only need to bound (b). First, using contraction argument (Theorem 18) as we used in the proof for Theorem 30, we have

$$\mathbb{E}\left[\sup_{v\in S^{d-1}} \left|\mathbb{E}\left[\overline{W}_Q(v)\right] - \mathbb{E}\left[\overline{U}_Q(v)\right]\right|\right] \leqslant 2\sqrt{\frac{\mathrm{Tr}\left(\Sigma\right)}{n}}.$$

With that in mind, note that $\phi_{-3Q,3Q}$ is 1-Lipschitz, satisfying

$$\left|\phi_{-3Q,3Q}\left(\left|\langle X-\mu, v\rangle\right|\right)\right| \leqslant 3Q,$$

and

$$\mathbb{E}\left|\phi_{-3Q,3Q}\left(\left|\langle X-\mu, v\rangle\right|\right)\right|^2 \leqslant \mathbb{E}\left|\langle X-\mu, v\rangle\right|^2 \leqslant \lambda_{\max}(\Sigma),$$

Talagrand's inequality gives

$$\sup_{v\in S^{d-1}}\left|\overline{W}_Q(v)-\mathbb{E}\left[\overline{W}_Q(v)\right]\right| \leqslant 4\sqrt{\frac{\operatorname{Tr}(\Sigma)}{n}}+2\sqrt{\lambda_{\max}(\Sigma)\frac{x}{n}}+20Q\frac{x}{n}$$

with probability of at least $1-2\exp(-x)$. Set $x=\varepsilon n/2560$, and notice that $Q \geqslant 2Q_0$, we have

$$\sup_{v\in S^{d-1}}\left|\overline{W}_Q(v)-\mathbb{E}\left[\overline{W}_Q(v)\right]\right| \leqslant \frac{\varepsilon Q}{64}.$$

Putting all the bounds for (a), (b), (c) to (4.17), together with (4.16), we see that (4.14) holds as desired. □

This lemma ensures that the trimmed means $U_Q(v)$ are within $2\varepsilon Q$ of $\langle\mu, v\rangle$ in all directions, accounting for contamination and statistical error. By bounding the truncation interval and using concentration, we control deviations uniformly, ensuring the slabs $\Gamma(v, Q)$ are centered near the true mean.

Substep 2.3: Complete Proposition 34.

With Lemmas 35 and 36, we prove the proposition. On the event E (with probability of at least $1-\delta$), for all $v \in S^{d-1}$, $|U_Q(v)-\langle\mu, v\rangle| \leqslant 2\varepsilon Q$. Thus, $\mu \in \Gamma(v, Q)$ for all v, implying $\Gamma(Q) \neq \varnothing$.

Now, let $x_1, x_2 \in \Gamma(Q)$ then for every $v \in S^{d-1}$, we have

$$\left|\langle x_1-x_2, v\rangle\right| \leqslant \left|\langle x_1, v\rangle - U_Q(v)\right| + \left|\langle x_2, v\rangle - U_Q(v)\right| \leqslant 4\varepsilon Q.$$

Thus, $\|x_1-x_2\| \leqslant 4\varepsilon Q$.

Step 3: Extend to the Estimator $\widehat{\mu}$.

Choose an integer i_0 such that $Q = 2^{i_0}$ lies within the interval $[2Q_0, 4Q_0)$, and define E^* as the favorable event where both Lemma 35 and Lemma 36 hold, ensuring

$$\sup_{v \in S^{d-1}} |U_Q(v) - \langle \mu, v \rangle| \leqslant 2\varepsilon Q.$$

To clarify, recall that

$$U_Q(v) = \frac{1}{n} \sum_{i=1}^{n} \phi_{\alpha_v - Q, \beta_v + Q} \left(\langle \widetilde{X}_i, v \rangle \right).$$

Event E^* occurs with probability at least $1 - \delta$, and on E, any point within $\Gamma\left(2^{i_0}\right)$ is at most a distance of $4\varepsilon Q_0$ from the mean μ. Therefore, it is sufficient to demonstrate that, on event E^*, the sets $\Gamma\left(2^i\right)$ for $i \geqslant i_0$ form a nested sequence. By the definition of i^*, we have

$$\widehat{\mu} \in \bigcap_{i \geqslant i^*} \Gamma\left(2^i\right) \subset \Gamma\left(2^{i_0}\right),$$

which implies $\|\widehat{\mu} - \mu\| \leqslant 4\varepsilon Q_0$.

To confirm that $\Gamma\left(2^{i_0}\right) \subset \Gamma(2^{i_0+1})$, we need to verify that for all v in S^{d-1}, $|\langle x, v \rangle - U_{2Q}(v)| \leqslant 4\varepsilon Q$. For any $x \in \Gamma\left(2^{i_0}\right) \subset \Gamma(Q)$ with $Q \in [2Q_0, 4Q_0)$, we have

$$\begin{aligned} |\langle x, v \rangle - U_{2Q}(v)| &\leqslant |\langle x, v \rangle - U_Q(v)| + |U_Q(v) - U_{2Q}(v)| \\ &\leqslant 2\varepsilon Q + |U_Q(v) - U_{2Q}(v)|. \end{aligned}$$

Hence, it is enough to prove that $|U_Q(v) - U_{2Q}(v)| \leqslant 2\varepsilon Q$.

Observe that on event E^*, there are at most $\varepsilon n/4$ points $\widetilde{X}_i$ where $\langle \widetilde{X}_i, v \rangle$ falls outside the interval defined by $\alpha_v - 2^{i_0}$ and $\beta_v + 2^{i_0}$. Consequently, the number of points where $U_Q(v) \neq U_{2Q}(v)$ is at most $\varepsilon n/4$, leading to a difference of at most $(2Q\varepsilon n/4)/n = \varepsilon Q/2$.

Using induction, this reasoning extends to show that on event E, $\Gamma\left(2^i\right) \subset \Gamma(2^{i+1})$ holds for all $i \geqslant i_0$, thereby completing the proof of Theorem 33.

□

With the proof of Theorem 33 complete, we have concluded the core content of this thesis, which focuses on the median-of-means and trimmed mean estimators as robust methods for mean estimation under heavy-tailed distributions and adversarial contamination. In the chapters that follow, we will explore the computational challenges and comparative performance of these estimators, before summarizing our findings and discussing future research directions.

Chapter 5

Computational Considerations and Comparison of Estimators

Having established the theoretical performance bounds for the median-of-means (MoM) and trimmed mean estimators in Chapters 3 and 4, we now turn to practical aspects of these estimators. This chapter examines their computational complexity and compares their performance, robustness, and applicability. Section 5.1 analyzes the theoretical computational challenges of implementing both estimators, particularly in high dimensions. Section 5.2 contrasts their error bounds and robustness to heavy-tailed distributions and adversarial contamination. Finally, Section 5.3 discusses the trade-offs between computational efficiency and statistical performance, highlighting scenarios where each estimator excels.

5.1 Computational Complexity

The computational complexity of an estimator is a critical factor in its practical applicability, especially in high-dimensional settings where d can

be large. Both the MoM and trimmed mean estimators, while theoretically robust, face distinct computational challenges, as noted by Lugosi and Mendelson in [LM21]. Below, we explore these challenges for each estimator, focusing on their multivariate implementations.

5.1.1 Median-of-Means Estimator

The multivariate MoM estimator, introduced in Chapter 3, extends the univariate MoM by replacing the median with a multivariate notion of median, defined as the point minimizing the radius of a set T_a (see Section 2 of Lugosi and Mendelson [LM19b]). The algorithm proceeds as follows: partition the n samples into k blocks of size $m = n/k$, compute the block means $Z_j = \frac{1}{m}\sum_{i\in B_j} X_i$, and define

$$T_a = \left\{x \in \mathbb{R}^d \mid \exists J \subset [k], |J| > k/2 : \forall j \in J, \|Z_j - x\| \leqslant \|Z_j - a\|\right\}.$$

The estimator $\widehat{\mu}_n$ is chosen as $\arg\min_{a\in\mathbb{R}^d} \text{radius}(T_a)$, where $\text{radius}(T_a) = \sup_{x\in T_a} \|x - a\|$.

The computational complexity of this estimator arises from the optimization problem. Lugosi and Mendelson [LM19b] note that computing T_a involves evaluating distances for all block means, requiring $O(n)$ operations per evaluation of a. The optimization over $a \in \mathbb{R}^d$ is non-trivial, as T_a is an intersection of unions of closed balls, but its continuity ensures a minimum exists. A naive approach to approximate $\widehat{\mu}_n$ involves discretizing $\mathbb{R}^d$, but this is exponential in d.

A more practical approach proposed by Lugosi and Mendelson [LM19b] uses a coordinate descent algorithm. Start with a line in $\mathbb{R}^d$, discretize the segment containing the convex hull of $Z_1, \ldots, Z_k$ with mesh $O(r)$, where r is the error bound from Theorem 1 in Lugosi and Mendelson [LM19b]

$$r = \max\left(960\sqrt{\frac{\text{Tr}(\Sigma)}{n}}, 240\sqrt{\frac{\lambda_{\max}(\Sigma)\log(2/\delta)}{n}}\right).$$

Perform pairwise comparisons using the MoM estimate to find a point on the line that "defeats" others at distance $2r$, repeat on an orthogonal line through the winner, and continue for d steps. This algorithm runs in time quadratic in $1/r$ and linear in d, but only guarantees $\|\widehat{\mu}_n - \mu\|_\infty \leqslant Cr$ in the ℓ_∞-norm. To achieve Euclidean norm guarantees, random directions are needed, but this requires exponentially many directions ($O(2^d)$), making it computationally infeasible in high dimensions.

Alternatively, one can start with the geometric median of the block means, which has a bound of $C\sqrt{\frac{\mathrm{Tr}(\Sigma)\log(1/\delta)}{n}}$ (see Minsker [Min15]), and search within a ball of radius $r\sqrt{\log(1/\delta)}$. However, this exhaustive search takes time $O(\log^d(1/\delta))$, again exponential in d. Thus, while the MoM estimator is computationally simpler than the trimmed mean, achieving optimal Euclidean error bounds remains challenging in high dimensions.

5.1.2 Trimmed Mean Estimator

The multivariate trimmed mean estimator, detailed in Chapter 4.3, applies univariate trimming to projections in all directions $v \in S^{d-1}$, intersecting slabs $\Gamma(v, Q) = \left\{x \in \mathbb{R}^d : |\langle x, v\rangle - U_Q(v)| \leqslant 2\varepsilon Q\right\}$ to form $\Gamma(Q)$. The estimator $\widehat{\mu}$ is selected from $\bigcap_{i \geqslant i^*} \Gamma(2^i)$, where i^* is the smallest integer such that the intersection is non-empty.

Lugosi and Mendelson [LM21] highlight that this procedure is computationally infeasible in its naive form, primarily due to the need to compute projections over all directions in S^{d-1}. For each direction v, computing $U_Q(v)$ involves sorting the projections $\langle \widetilde{Y}_i, v\rangle$ (to find α_v, β_v) and applying the truncation function to $\langle \widetilde{X}_i, v\rangle$, costing $O(n \log n)$ per direction for $2n$ points. However, S^{d-1} is infinite, and even discretizing it with a fine mesh (e.g., an ε-net) requires $O((1/\varepsilon)^{d-1})$ directions, leading to exponential complexity in d. The final step of intersecting slabs and selecting $\widehat{\mu}$ adds further complexity, as $\Gamma(Q)$ is defined over all directions.

Lugosi and Mendelson [LM21] note that while efficient sub-Gaussian estimators exist for i.i.d. data (e.g., Hopkins [Hop20]), these fail under

contamination. The trimmed mean's robustness comes at the cost of this computational burden, posing an open problem for developing efficient versions with similar statistical guarantees.

5.2 Performance and Robustness Comparison

We now compare the MoM and trimmed mean estimators in terms of their performance bounds and robustness, leveraging the results from Chapters 3 and 4. The MoM estimator achieves sub-Gaussian performance under minimal assumptions, while the trimmed mean additionally handles adversarial contamination, making it more robust but computationally demanding.

5.2.1 Univariate Case

In the univariate case ($d = 1$), the MoM estimator, presented in Theorem 28, achieves, with probability at least $1 - \delta$

$$|\widehat{\mu}_n - \mu| \leqslant \sigma \sqrt{\frac{32 \log(1/\delta)}{n}},$$

for $k = \lceil 8 \log(1/\delta) \rceil$ blocks. This bound, derived from Lugosi and Mendelson [LM19b], holds under the minimal assumption of finite variance (σ^2), offering sub-Gaussian performance without contamination.

The univariate trimmed mean, from Theorem 32, satisfies, with probability at least $1 - \delta$

$$|\widehat{\mu} - \mu| \leqslant 3\mathcal{E}(4\varepsilon, X) + 2\sigma_X \sqrt{\frac{\log(4/\delta)}{n}},$$

where $\varepsilon = 8\eta + 12 \cdot \frac{\log(4/\delta)}{n}$, and

$$\mathcal{E}(\varepsilon, X) = \max \left\{ \mathbb{E}\left[\left| \overline{X} \right| \mathbf{1}_{\overline{X} \leqslant Q_{\varepsilon/2}(\overline{X})} \right], \mathbb{E}\left[\left| \overline{X} \right| \mathbf{1}_{\overline{X} \geqslant Q_{1-\varepsilon/2}(\overline{X})} \right] \right\}.$$

A simpler bound, with probability at least $1 - 4\exp(-\varepsilon n/12)$, is

$$|\widehat{\mu} - \mu| \leqslant 10\sqrt{\varepsilon}\sigma_X.$$

This bound, from Lugosi and Mendelson [LM21], holds under finite variance and handles up to $2\eta n$ corrupted points. When $\eta = 0$, the bound becomes $O\left(\sigma_X\sqrt{\frac{\log(4/\delta)}{n}}\right)$, matching the MoM's sub-Gaussian rate up to constants. For $\eta > 0$, the $\sqrt{\varepsilon}\sigma_X \sim \sqrt{\eta}\sigma_X$ term reflects contamination's impact, which is minimax optimal, as shown in Section 2 of Lugosi and Mendelson [LM21].

The trimmed mean's ability to handle contamination gives it a significant advantage in noisy settings, while the MoM excels in clean, heavy-tailed scenarios due to its simplicity and matching sub-Gaussian performance.

5.2.2 Multivariate Case

In the multivariate case, the MoM estimator (Theorem 29) satisfies, with probability at least $1 - \delta$

$$\|\widehat{\mu}_{\text{MoM}} - \mu\| \leqslant c\left(\sqrt{\frac{\text{Tr}(\Sigma)}{n}} + \sqrt{\frac{\lambda_{\max}(\Sigma)\log(2/\delta)}{n}}\right),$$

under finite second moments, as per Lugosi and Mendelson [LM19b]. This sub-Gaussian bound is optimal for clean data but assumes no contamination.

The multivariate trimmed mean (Theorem 33) achieves, with probability at least $1 - \delta$

$$\|\widehat{\mu} - \mu\| \leqslant c\left(\sqrt{\frac{\text{Tr}(\Sigma)}{n}} + \sqrt{\frac{\lambda_{\max}(\Sigma)\log(2/\delta)}{n}} + \sqrt{\lambda_{\max}(\Sigma)\eta}\right),$$

handling up to ηn corrupted points per half of the sample. When $\eta = 0$, the bound matches the MoM's, confirming its sub-Gaussian performance. The additional $\sqrt{\lambda_{\max}(\Sigma)\eta}$ term, shown to be necessary by Lugosi and

Mendelson [LM21], reflects the cost of contamination, but allows the trimmed mean to maintain robustness where the MoM fails.

For sub-Gaussian distributions, Lugosi and Mendelson [LM21] note that the $\sqrt{\lambda_{\max}(\Sigma)\eta}$ term can improve to $\eta\sqrt{\log(1/\eta)}\sqrt{\lambda_{\max}(\Sigma)}$, tightening the bound under stronger assumptions, a property the MoM cannot achieve without modification.

5.2.3 Robustness to Contamination

The MoM estimator assumes clean data, making it vulnerable to adversarial contamination. As discussed in Lugosi and Mendelson [LM21], and our example in Section 4.1, even a small fraction of corrupted points can significantly distort the block means, skewing the median and leading to large errors. In contrast, the trimmed mean explicitly mitigates contamination by trimming extreme projections in each direction, ensuring robustness even with up to ηn corrupted points per half, as shown in its construction (Section 4.3).

5.3 Practical Implications and Trade-offs

The MoM and trimmed mean estimators offer distinct advantages, reflecting a trade-off between computational efficiency and robustness. The MoM is computationally simpler, with linear complexity in n and manageable scaling in low dimensions, making it suitable for clean, heavy-tailed data, such as financial returns without errors. However, its lack of contamination robustness limits its use in noisy settings.

The trimmed mean, while computationally expensive due to projections over S^{d-1}, excels in scenarios with adversarial contamination, such as sensor data with faulty readings or datasets with malicious errors. Its sub-Gaussian performance and ability to handle up to ηn corrupted points per half make it ideal for robust applications, though its exponential complexity in d poses challenges in high dimensions.

In practice, the choice between these estimators depends on the data's characteristics and computational constraints. For clean, high-dimensional data, the MoM offers a practical solution with optimal sub-Gaussian performance. For noisy, lower-dimensional data, the trimmed mean provides superior robustness, though future work on efficient implementations, as suggested by Lugosi and Mendelson [LM21], could broaden its applicability.

Chapter 6

Conclusion

This thesis has explored robust mean estimation through the lens of two powerful estimators: the median-of-means (MoM) and the trimmed mean. We have analyzed their theoretical performance, computational challenges, and practical implications, focusing on their ability to handle heavy-tailed distributions and adversarial contamination. This final chapter summarizes our key findings, reflects on the broader significance of robust mean estimation, and proposes directions for future research. Section 6.1 recaps the main results, Section 6.2 discusses the importance of these methods, and Section 6.3 outlines potential avenues for further exploration.

6.1 Summary of Findings

The primary goal of this thesis was to investigate robust mean estimation for random variables and vectors under heavy-tailed distributions and adversarial contamination, focusing on the MoM and trimmed mean estimators. Chapter 3 introduced the MoM estimator, which achieves sub-Gaussian performance under minimal assumptions. In the univariate case (Theorem 28), the MoM estimator satisfies, with probability at least

$1-\delta$:

$$|\widehat{\mu}_n - \mu| \leqslant \sigma\sqrt{\frac{32\log(1/\delta)}{n}},$$

requiring only finite variance (σ^2). In the multivariate case (Theorem 29), it extends to:

$$\|\widehat{\mu}_{\mathrm{MoM}} - \mu\| \leqslant c\left(\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}} + \sqrt{\frac{\lambda_{\max}(\Sigma)\log(2/\delta)}{n}}\right),$$

under finite second moments, matching the optimal sub-Gaussian rate for clean data, as established by Lugosi and Mendelson [LM19b].

Chapter 4.2 and Chapter 4.3 explored the trimmed mean estimator, which builds on the MoM's sub-Gaussian performance while additionally handling adversarial contamination. In the univariate case (Theorem 32), the trimmed mean achieves, with probability at least $1-\delta$:

$$|\widehat{\mu} - \mu| \leqslant 3\mathcal{E}(4\varepsilon, X) + 2\sigma_X\sqrt{\frac{\log(4/\delta)}{n}},$$

where $\varepsilon = 8\eta + 12\frac{\log(4/\delta)}{n}$, and a simpler bound of $10\sqrt{\varepsilon}\sigma_X$. In the multivariate case (Theorem 33), it satisfies:

$$\|\widehat{\mu} - \mu\| \leqslant c\left(\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}} + \sqrt{\frac{\lambda_{\max}(\Sigma)\log(2/\delta)}{n}} + \sqrt{\lambda_{\max}(\Sigma)\eta}\right),$$

handling up to ηn corrupted points per sample half, as shown by Lugosi and Mendelson [LM21]. When $\eta = 0$, the bound reduces to the MoM's, but the additional $\sqrt{\lambda_{\max}(\Sigma)\eta}$ term ensures robustness to contamination, a feature the MoM lacks.

Chapter 5 examined the practical aspects of these estimators. The MoM is computationally simpler, with linear complexity in n, but achieving Euclidean norm guarantees in high dimensions requires exponential time in d, as noted by Lugosi and Mendelson [LM19b]. The trimmed mean, while optimal in performance, faces significant computational challenges due to projections over S^{d-1}, leading to exponential complexity in d, as highlighted by Lugosi and Mendelson [LM21]. The comparison revealed that the MoM excels in clean, heavy-tailed scenarios, while the trimmed mean is superior in noisy settings, though at a computational cost.

6.2 Significance of Robust Mean Estimation

Robust mean estimation is a cornerstone of statistical analysis, particularly in applications where data may be heavy-tailed or contaminated. Traditional estimators like the empirical mean fail in such scenarios, as they are sensitive to outliers, leading to poor performance (Section 2.5). The MoM and trimmed mean estimators address this challenge by achieving sub-Gaussian performance under minimal assumptions, making them invaluable tools in fields like finance, machine learning, and sensor data processing.

In finance, datasets often exhibit heavy-tailed behavior due to extreme market events. The MoM estimator provides a reliable mean estimate without requiring distributional assumptions beyond finite variance, enabling accurate risk assessment. In machine learning, robust mean estimation is crucial for feature preprocessing, where outliers can skew model training. The trimmed mean's ability to handle adversarial contamination is particularly relevant in sensor networks, where faulty readings or malicious attacks can corrupt data, as discussed in Chapter 5.

The theoretical contributions of Lugosi and Mendelson [LM19b] and Lugosi and Mendelson [LM21] underscore the optimality of these estimators. The MoM achieves sub-Gaussian rates with minimal assumptions, while the trimmed mean extends this to contaminated settings, matching the best possible bounds up to constants. This robustness ensures reliable inference in real-world scenarios, where data integrity cannot be guaranteed, highlighting the practical and theoretical importance of these methods.

6.3 Future Research Directions

While the MoM and trimmed mean estimators offer robust solutions, several open questions remain. A key challenge is the computational complexity of the multivariate trimmed mean. As noted in Chapter 5,

its reliance on projections over S^{d-1} leads to exponential complexity in d, limiting its applicability in high-dimensional settings. Future research could explore efficient approximations, such as sampling a subset of directions or using dimensionality reduction techniques, to achieve sub-Gaussian performance with polynomial-time algorithms, as suggested by Lugosi and Mendelson [LM21].

Another direction is to extend these estimators to other contamination models, such as Huber's model, where contamination is i.i.d. rather than adversarial. Lugosi and Mendelson [LM21] note that for sub-Gaussian distributions under Huber's model, the trimmed mean's contamination term can improve to $\eta\sqrt{\log(1/\eta)}$, suggesting potential for tighter bounds under specific contamination structures. Investigating hybrid estimators that combine MoM's simplicity with the trimmed mean's robustness could also yield practical solutions, potentially leveraging the geometric median approach discussed in Minsker [Min15].

Finally, the dependence on the confidence parameter δ in both estimators, as highlighted by Devroye et al. [Dev+16], warrants further exploration. Developing estimators that perform uniformly across a range of confidence levels without prior knowledge of δ could enhance their practical utility, building on the ideas in Lugosi and Mendelson [LM19b]. These directions promise to advance the field of robust mean estimation, bridging theoretical optimality with practical implementation.

Bibliography

[Ber24] Sergei N. Bernstein. "On a Modification of Chebyshev's Inequality and of the Law of Large Numbers". In: *Matematicheskii Sbornik* 31 (1924), pp. 220–230.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford: Oxford University Press, 2013. DOI: 10.1093/acprof:oso/9780199535255.001.0001.

[Blu+89] Anselm Blumer et al. "Learnability and the Vapnik-Chervonenkis Dimension". In: *Journal of the ACM* 36.4 (1989), pp. 929–965. DOI: 10.1145/76359.76371.

[Cat12] Olivier Catoni. "Challenging the Empirical Mean and Empirical Variance: A Deviation Study". In: *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* 48.4 (2012), pp. 1148–1185. DOI: 10.1214/11-AIHP454.

[Che67] Pafnuty L. Chebyshev. "Des valeurs moyennes". In: *Journal de Mathématiques Pures et Appliquées* 12 (1867), pp. 177–184.

[Dev+16] Luc Devroye et al. "Sub-Gaussian Mean Estimators". In: *The Annals of Statistics* 44.6 (2016), pp. 2695–2725. DOI: 10.1214/16-AOS1440.

[Hoe63] Wassily Hoeffding. "Probability Inequalities for Sums of Bounded Random Variables". In: *Journal of the American Statistical Association* 58.301 (1963), pp. 13–30. DOI: 10.1080/01621459.1963.10500830.

[Hop20] Samuel B. Hopkins. "Sub-Gaussian Mean Estimation in Polynomial Time". In: *The Annals of Statistics* 48.6 (2020), pp. 3468–3487. DOI: 10.1214/19-AOS1932.

[Hub64] Peter J. Huber. "Robust Estimation of a Location Parameter". In: *The Annals of Mathematical Statistics* 35.1 (1964), pp. 73–101. DOI: 10.1214/aoms/1177703732.

[Hub81] Peter J. Huber. *Robust Statistics*. 2nd ed. New York: John Wiley & Sons, 1981. DOI: 10.1002/9780470434697.

[LM19a] Gábor Lugosi and Shahar Mendelson. "Mean Estimation and Regression under Heavy-Tailed Distributions—A Survey". In: *Foundations of Computational Mathematics* 19 (2019), pp. 1145–1190. DOI: 10.1007/s10208-019-09427-x.

[LM19b] Gábor Lugosi and Shahar Mendelson. "Sub-Gaussian Estimators of the Mean of a Random Vector". In: *The Annals of Statistics* 47.2 (2019), pp. 783–794. DOI: 10.1214/17-AOS1639.

[LM21] Gábor Lugosi and Shahar Mendelson. "Robust Multivariate Mean Estimation: The Optimality of Trimmed Mean". In: *The Annals of Statistics* 49.1 (2021), pp. 393–410. DOI: 10.1214/20-AOS1975.

[LT91] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 1991. DOI: 10.1007/978-3-642-20212-4.

[Min15] Stanislav Minsker. "Geometric Median and Robust Estimation in Banach Spaces". In: *Bernoulli* 21.4 (2015), pp. 2308–2335. DOI: 10.3150/14-BEJ645.

[Tal96] Michel Talagrand. "New Concentration Inequalities in Product Spaces". In: *Inventiones Mathematicae* 126.3 (1996), pp. 505–563. DOI: 10.1007/s002220050109.

[TIS76] B. S. Tsirelson, I. A. Ibragimov, and V. N. Sudakov. "Norms of Gaussian Sample Functions". In: *Proceedings of the Third Japan-USSR Symposium on Probability Theory* (1976), pp. 20–41.

[Tuk60] John W. Tukey. "A Survey of Sampling from Contaminated Distributions". In: *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (1960). Ed. by Ingram Olkin, pp. 448–485.

[Val79] Leslie G. Valiant. "Probably Approximately Correct Learning". In: *Proceedings of the National Conference on Artificial Intelligence* (1979), pp. 436–439.

[VC82] Vladimir N. Vapnik and Alexey Ya. Chervonenkis. "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities". In: *Theory of Probability and Its Applications* 16.2 (1982), pp. 264–280. DOI: 10.1137/1116025.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018. DOI: 10.1017/9781108231596.

[Ver20] Roman Vershynin. *Gaussian and Fourier Analysis*. Lecture notes. Accessed: 2025-04-16. 2020. URL: https://www.math.uci.edu/~rvershyn/papers/GFA-book.pdf.