Recent Developments in Robust Statistics

Yeshwanth Cherapanamjeri

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2023-259
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2023/EECS-2023-259.html

December 1, 2023

Recent Developments in Robust Statistics

By

Yeshwanth Cherapanamjeri

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Peter Bartlett, Chair
Professor Prasad Raghavendra
Professor Jelani Nelson
Professor Nikita Zhivotovskiy

Summer 2023

Recent Developments in Robust Statistics

Abstract

Recent Developments in Robust Statistics

by

Yeshwanth Cherapanamjeri

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Peter Bartlett, Chair

The design of statistical estimators robust to outliers has been a mainstay of statistical research through the past six decades. These techniques are even more prescient in the contemporary landscape where large-scale machine learning systems are deployed in increasingly noisy and adaptive environments. In this thesis, we consider the task of building such an estimator for arguably the simplest possible statistical estimation problem – that of mean estimation. There is surprisingly little understanding of the computational and statistical limits of estimation and the trade-offs incurred even for this relatively simple setting. We make progress on this problem along three complementary axes.

Our first contribution is a simple *algorithmic* framework for constructing robust estimators. Our framework allows for a significant speed-up over prior approaches for mean estimation while also allowing for easy extensibility to other statistical estimation tasks where it achieves state-of-the-art performance.

Secondly, we investigate the *statistical* boundaries of mean estimation where we demonstrate the necessary statistical degradation incurred in extremely heavy-tailed scenarios. While prior work showed that estimation could be performed as well as if one had access to *Guassian* data, we establish that this is no longer true when the data possesses heavier tails. We provide lower bounds which exhibit this degradation and an (efficient) algorithm matching them.

Lastly, we consider the *stability* of these estimators to natural transformations of the data. Inspired by the empirical mean, classical work constructed estimators equivariant to *affine* transformations. These works, however, lacked the strong quantitative performance of more recent approaches. We demonstrate that such trade-offs are in fact *necessary* by constructing novel lower bounds for *affine-equivariant* estimators. We then show that classical estimators are quantitatively deficient *even* in this restricted class and devise an estimator based on a novel notion of a high-dimensional median which matches the lower bound.

To my mother and sister.

Contents

List of Figures

Acknowledgments

In many ways, graduate school was a challenging and spiritually exhausting experience. During this time, I was incredibly fortunate to be surrounded by people whose thoughtfulness and kindness made it one of the most enjoyable and rewarding periods of my life.

Firstly, I would like to thank my advisor, Peter Bartlett. Peter is not only a brilliant researcher but also an amazingly patient and understanding mentor. I pursued many different research directions through graduate school and constantly jumped between them with very little forethought. Peter displayed superhuman levels of restraint by patiently listening to my half-baked ideas on all these topics. His ability to both understand and appreciate these diverse topics to the extent of asking deeply insightful questions is something I aspire to. It is only through this that I have a much greater appreciation of my own research motivations and a set of guiding principles around which to organize my future research career. I could not have wished for a better advisor to guide me through my Ph.D.

I was also blessed with the opportunity of collaborating with several other senior researchers through my journey. Prasad Raghavendra, who in many ways, to use a cliched phrase, was a second advisor to me and was always ready to lend an empathetic ear to my troubles. My conversations with him also led to a broadening of my intellectual perspectives for which I am extremely grateful. Sébastien Bubeck, whose infectious conviction and enthusiasm led to a fun collaboration and whose advice on navigating a turbulent academic environment was invaluable. Jelani Nelson, whose clarity of thought and aesthetic appreciation of complicated material are an inspiration. Costis Daskalakis, whose sense of adventure and ability to seek out interesting research problems very different from the status quo and incidentally inspired my first project in grad school. Bin Yu, who (repeatedly) taught me the joys and challenges of working on grounded practical methodology and whose commitment to rigorous scientific advancement I find deeply inspiring. Lastly, I have had the great pleasure of collaborating with Nikita Zhivotovskiy whose wonderful enthusiasm for classical problems in learning theory and commitment to scholarship renews one faith in the academic enterprise.

I am also deeply indebted to Prateek Jain, Praneeth Netrapalli, and Nagarajan Natarajan who mentored me for two wonderful years at Microsoft Research. Their patience and encouragement through my first few projects in theoretical machine learning motivated my decision to apply to graduate school. I would also like to express my gratitude to Ganesh Ramakrishnan and Soumen Chakrabarti, who supervised my undergraduate thesis and kindled my enthusiasm for research.

The senior graduate students and postdocs at Berkeley patiently mentored me, technically and beyond, through many challenging research projects. Nicolas Flammarion marshalled me through my first research project in graduate school and whose encouragement of my non-academic pursuits enriched my graduate school experience in more ways than one. Sam Hopkins patiently taught me the nuances of sum-of-squares and whose perspectives on heavy tailed estimation helped inspire much of the work in this thesis. Nilesh Tripuraneni's insightful perspectives on the statistical modeling of real-world scenarios (and its failures) were profoundly influential on my research. Manolis Zampetakis introduced me to the wonderful world of Econometrics with its intriguing interplay of game theory and statistics.

I am also thankful for the numerous other collaborators I've been fortunate enough to work with. Sidhanth Mohanty and Morris Yau for a fun early project on list-decodable estimation. Efe Aras, Mike Jordan and Tarun Kathuria on some projects in heavy-tailed estimation. Gauthier Gidel and Rémi Tachet des Combes on the theoretical foundations of adversarial examples. Andrew Ilyas who brought much enthusiasm to and taught me most of what I understand on the challenges of statistical estimation in Econometrics. Sandeep Silwal, David Woodruff, Fred Zhang, Richard Zhang and Samson Zhou for some fascinating problems in numerical linear algebra and adaptive sketching. Aliyah Hsu who along with Briton Park and Anobel Odisho, taught me the challenges (and opportunities) of applying machine learning methodology to real-life scientific domains. Ishaq Aden-Ali and Abhishek Shetty have been wonderful friends and colleagues to learn from through our exploration of classical topics in combinatorial learning theory. Nived Rajaraman and Ayush Sekhari have shed much light on the challenges of ensuring privacy in modern machine learning. Finally, Zihao Chen's perseverance has been inspiring in our recent project on heavy-tailed estimation taking me back full circle to the start of my Ph.D.

I also sincerely express my gratitude to the numerous research groups that have been so welcoming through my time at Berkeley. Peter's group was always full of exciting research ideas and was always open to new directions: Xiang Cheng, Niladri Chatterji, Aldo Pacchiano, Wenlong Mou, Alex Tsigler, and Juanky Perdomo were a joy to be around. The efforts of Siqi Liu, Seri Khoury, Chinmay Nirkhe and Elizabeth Yang played an immense role in making the theory group one of the most supportive intellectual environments to be part of as a young grad student. Finally, the Yu group, was a wonderful environment to appreciate the intricacies of real-life statistical research in a variety of domains.

I'd also like to thank Gireeja Ranade for a wonderful GSI experience, helping teach EECS 127. Her commitment to pedagogy and openness towards experimentation are an inspiration. I'd also like to thank Alistair Sinclair and Yun Song and Prasad Raghavendra and Luca Trevisan for my first teaching experiences in CS 70 and CS 170 respectively.

I'm especially thankful for Jean Nguyen and Shirley Salanio for their support, both in navigating a confusing academic program but also the American immigration system on numerous occasions.

I'd like to thank some senior graduate students whose mentorship was invaluable through graduate school. The extensive weekend workout sessions (with very few noticeable results) with Raaz Dwivedi will be dearly missed. Niladri Chatterji's relentless advocacy for Friday evening happy hours were a welcome reprieve during punishing deadlines. Somil Bansal's sage advice steered us through some trying times.

I would like to thank my friends without whom I would have graduated in half the time but the journey would've been far less enjoyable. I will miss them dearly. The household (and honorary members) made 1735 Cedar a second home despite moving 10,000 miles to graduate school: Ashish Kumar, with whom I shared much of the difficulties of being a fresh immigrant in a foreign country and who, despite his enthusiasm, knew when to slow down and smell the roses, Armin Askari, without whose efforts I may have never stepped out of the house and who has been a wonderful companion on many memorable trips to Europe and South America, Evonne Ng, whose endless supply of baked goods, boba tea, and binge sessions got the house through a difficult lockdown period and Kiran Shiragur, whose calm dismissal any pressing professional commitments led to many enjoyable chai-time conversations. Out side of the Cedar House, Sidhanth Mohanty led me on numerous memorable, meandering strolls through the beautiful streets of Berkeley and whose joyful enthusiasm never failed to lift ones spirits, Zihao Chen's tenacity (while sometimes materially destructive) were a constant source of inspiration, Nilesh Tripuraneni's endless supply of single-malt whisky, Buddha Board memes, and vignettes on American cultural appreciation were not only educational but also joyous and heartening, Allan Jabri's emotional sensitivity and love for dumplings, pork, and gummy bears led to many hilarious but contemplative discussions (mostly before a deadline), Abhishek Shetty's profound spiritual bond with the Devil's Advocate was the source of several spirited and memorable debates (all of which remain unresolved), Nived Rajaraman's numerous culinary talents and warm personality made for many enjoyable gatherings at 1641 Walnut, Chandan Singh's love for Mean Girls, Pizza, and FIFA enlivened many a gathering, Efe Aras' boundless energy and ruthlessly combative but endearing spirit were a godsend through several otherwise dull evenings spent working on probability theory assignments, Sushrut Karmalkar's infectious love for K-culture and remotely organized binge sessions which brought normalcy to a challenging period, Frederic Koehler's sharp wit and biting sarcastic commentary brought humor to some particularly egregious cinematic choices, Melih Elibol's experienced wisdom provided much-needed perspective on life and research, Juanky Perdomo's joyful spirit and passion for celebration brought much spiritual rejuvenation albeit at some short-term physical cost, Ghassen Jerfel's fondness for the finer things in life lifted everyone around him, Yu Sun's emphatic personality, wizened soul and repeated attempts at musical enlightenment (which largely proved fruitless) will be sorely missed, Colin Li's enthusiasm and infinite patience through our many Sunday afternoon conversations and culinary excursions through Berkeley and Oakland are fond memories, Ishaq Aden-Ali's unmatched charisma and enthusiastic appreciation for Indian food and heavy cream brought much cheer, Morris Yau's penchant for storytelling dramatically brought to life foundational events in human history, Wenlong Mou's exuberance and commitment to intellectual rigor made for many fun lunch conversations, and Tarun Kathuria's near infinite frustration on the failings of a certain large American corporation made for several entertaining conversations.

I would also like to thank friends from Microsoft Research: Himanshu Zade, who was an amazing travel companion and a gracious host during my visit to Seattle. Aditi Raghunathan was a constant source of engaging stories and entertainment. Mitali Bafna and Surbhi Goel made sure traveling to conferences wouldn't be a lonely experience.

I have been fortunate that many of my friends from my undergraduate institution IIT-Bombay moved to the Bay Area: Divyam Bansal, Navin Chandak, Mohit Gupta, Raghav Gupta, Ayush Kanodia, Guna Prasaad, and Vipul Venkataraman. Most of them chose a very different path than I did but have always made the effort to include me in their celebrations.

It is heartening to see how their lives have turned out.

Finally, my family, to whom this thesis is dedicated. My mother who despite immense hardships afforded me all the freedom to make my own choices and supported them through them. It is only now, as I grow older, that I appreciate the courage, commitment, effort, and patience that it took. I can only asipre to be half as much. My sister, who posses a will and wisdom far beyond her years, offered clear advice and unconditional love and support through the most turbelent periods of my life. To them, I owe the world.

Chapter 1

Introduction

Procedures for statistical estimation and analysis play a central role in the modern sciences and the functioning of large scale industries. Indeed, with the recent introduction and wide dissemination of computing technology, we have witnessed an explosion both in the amount of data available for such enterprises but also in the range of statistical methodology that aims to make use of such data. While this expansion opens up a wealth of new opportunities for technological advancement, it also presents significant challenges. One unfortunate consequence is that maintaining data *quality* quickly becomes impractical due to the scale and breadth of domains from which it is collected. Such data often contains extreme amounts of noise making accurate statistical inference challenging.

In this thesis, we specifically focus on the task of learning with *outliers*. Informally, these are points in a dataset whose behavior deviates from what is typically expected. Outliers are encountered in a range of domains from quantitative finance to operations research and network monitoring. The source of such data is also similarly varied. These points while typically comprising a small fraction of available data often have a devastating impact on the performance of typical algorithms employed in statistical analysis. For an illustration of such effects, consider Fig. 1.1 which shows the impact of outliers on the popular ordinary least squares (OLS) estimator. As we can see, even a small amount of outliers have a drastic impact on the OLS estimator. This detrimental effects are present in other estimation tasks as well. We will focus on arguably the simplest statistical estimation problem form of this type: mean estimation. We investigate the statistical and algorithmic limits of estimation in this setting and observe that in some cases, outliers do not have a significant impact on the recovery error while in others they do. The nature of this impact will be clear in the following chapters. Through the rest of the chapter, we will formally describe the problem and prior work in Section 1.1 before presenting our contributions in Section 1.3.

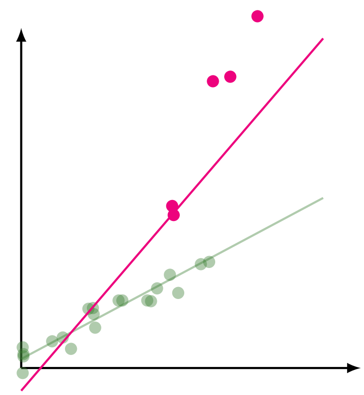


Figure 1.1: The results of using ordinary least squares on data with outliers. The green points represent the typical non-outlier data points while the red ones denote outlier data points which deviate from typical behavior. The green line is the desired least squares estimate obtained from running OLS on only on the green points while using the whole dataset results in the drastically different red line.

1.1 Problem Definition

The high-dimensional mean estimation problem without outliers is described below where $\|x\|$ denotes the Euclidean norm of a vector $x \in \mathbb{R}^d$.

Problem 1.1.1. Given n independent and identically distributed (iid) random vectors $\boldsymbol{X} = \{X_i\}_{i=1}^n$ drawn from a distribution D over $\mathbb{R}^d$ with mean μ, design an estimator, $\widehat{\mu}$, satisfying:

$$\mathbf{Pr}\left\{\|\widehat{\mu}(\boldsymbol{X}) - \mu\| \leqslant r_\delta\right\} \geqslant 1 - \delta$$

which minimizes r_δ for a target failure probability δ.

Note that the above problem statement imposes no constraints on the distribution (beyond possessing a mean) due to which no non-trivial recovery guarantees are possible. In our results, we will impose restrictions on the *moments* (for instance, the variance) of the distribution enabling stronger recovery guarantees. We defer a formal description of such assumptions to subsequent chapters.

More relevant to the previous discussion is that the problem statement also does not provide a formal description of how *outliers* are generated. Here, we will focus on two outlier models: the adversarial and heavy-tailed models which have been intensely investigated over the past 60 years and much is known of their statistical and computational properties.

Adverarial Corruption Model: This outlier model, which traces back to early work by Huber [35], allows an adversary to inspect the dataset, $\boldsymbol{X}$, and arbitrarily change an η fraction of them for some $\eta \in [0, 1]$. Here, the performance of an estimation procedure is measured in terms of the fraction of corruption, η, it may reliably tolerate without incurring arbitrarily poor performance. More recent work, however, has focused on obtaining quantitative finite sample guarantees. We defer an in-depth discussion to Section 1.2.

Heavy-tailed Corruption Model: In this setting, outliers occur naturally as part of the data by loosening the assumptions on the data generating distribution D. For instance, when milder assumptions are made about D such as merely the existence of a variance as opposed stronger assumptions like sub-Gaussianity, outliers are more likely to be present in the dataset due to the increased likelihood of tail events. The performance of a robust estimator is measured in a *statistical* sense through the dependence of r_δ on the number of datapoints, n, the dimension, d, and the failure probability, δ. This setting has been extensively studied in more recent work [52, 36, 2, 48]. We provide more context for these developments in Section 1.2.

As we will see, the algorithmic contributions presented in this thesis apply to *both* corruption models while the lower bounds are proved for each setting individually.

1.2 Prior Work

As alluded to in Section 1.1, there is much work on each of the outlier models previously discussed. We will start with the Adversarial Corruption Model.

Adversarial Corruption Model

We present classical work on the topic before proceeding to more recent developments.

Classical Work. The adversarial corruption model may be traced back to the early work of Huber [35] in response to a question raised by Tukey [60]. This work considered the one-dimensional setting and noted the extreme brittleness of the empirical mean to outliers in the data while also observing that alternative estimators such as the median and the Winsorized mean are more robust. In addition, the asymptotic normality of some of these estimators was established. An extension to higher dimensions was first formulated by Tukey [61] who proposed the Tukey median which generalizes the median in higher dimensions. In this and subsequent early work [49, 34, 57, 23, 54, 55, 56, 63, 14, 25, 45, 24, 40, 43, 42], the performance of these estimators was evaluated in terms of its breakdown point (see [22]):

$$\gamma(\boldsymbol{X}) = \sup \left\{ \eta : \sup \left\{ \left\| \widehat{\mu}(\boldsymbol{Y}) - \frac{1}{n} \sum_{x \in \boldsymbol{X}} x \right\| : |\boldsymbol{Y}| = |\boldsymbol{X}| \text{ and } |\boldsymbol{Y} \cap \boldsymbol{X}| \geqslant (1-\eta)|\boldsymbol{X}| \right\} < \infty \right\}$$

which measures the largest amount of corruption that can estimator can tolerate before its error can be made arbitrarily bad. Note that the only requirement of an estimator to have high breakdown point is that it achieves *finite* error. Thus, this notion is necessarily coarse

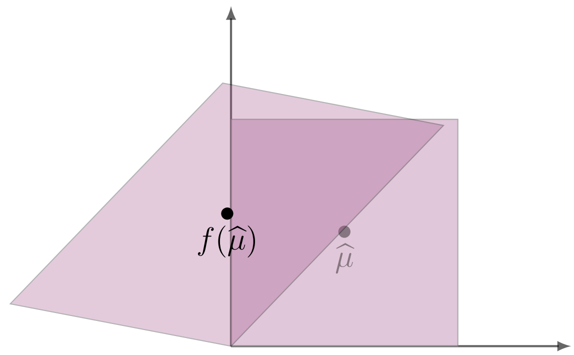


Figure 1.2: An illustration of the affine equivariant requirement. Here, the result of an estimator $\widehat{\mu}$ run on the transformation of the square to the tilted parallelogram is required to coincide with the transformation of the estimate obtained when run on the square itself.

in that it does not allow for a *quantitative* evaluation of these methods which has been the focus of more recent work.

Somewhat complementary to more recent developments, prior work also focused extensively on the *stability* properties of these estimators where the estimator is required to be stable to natural transformations of the data. Inspired by the empirical mean, one such property that received significant attention was the affine equivariance of the estimators being considered. Formally, this is the property that an estimator, $\widehat{\mu}$, is equivariant with respect to any invertible affine transformation, f; i.e $f(\widehat{\mu}(\boldsymbol{X})) = \widehat{\mu}(f(\boldsymbol{X}))$. This property is illustrated in Fig. 1.2. Indeed, several estimators based on the aforementioned Tukey median [61], the Stahel-Donoho depth [57, 23], simplicial volume [54], minimum volume ellipsoid [56], and the simplicial depth [41] are affine-equivariant and attempt to simultaneously achieve high breakdown point with affine-equivariance. In addition, the robustness [49, 63, 14, 25, 45, 24, 40, 43, 42] and consistency properties [64, 65] of these affine-equivariant estimators have been well studied.

Recent Developments. On the other hand, recent work in this setting has developed along two complementary axes. Firstly, there is increased emphasis on the *computational* aspects of these estimators and secondly, the *quantitative* properties of these estimators have received greater attention. On the computational side, these novel estimators are computable in polynomial (and subsequently, near-*linear*) time while quantitatively achieving *optimal* recovery guarantees in terms of the corruption fraction η. The first efficient estimator with near-optimal recovery guarantees (in terms of the corruption fraction η) was proposed in a breakthrough result of Diakonikolas, Kamath, Kane, Li, Moitra, and Stewart [18]. Since then, the statistical and computational complexity has been substantially improved in follow-up works [9, 21] resulting in estimators with near-optimal statistical and computational performance. For instance, in the setting of finite variance, these estimators run in near-*linear* time and achieve recovery error of $O(\sqrt{\eta})$ with d/η samples with high probability both of which are known to be *optimal*. These ideas have since been extended to numerous other settings which remain out of the scope of this thesis and we direct the interested reader to the excellent survey by Diakonikolas and Kane [19] for an expanded discussion on the topic.

Note that while these estimators are computationally and statistically efficient, in contrast to prior work, they sacrifice the stability properties possessed by the estimators discussed previously. Next, we move on to the heavy-tailed corruption model.

Heavy-tailed Corruption Model

As previously discussed, an alternative statistical model for outliers is the heavy-tailed corruption model. In this setting, minimal assumptions are made about the data generating distribution (for instance, the covariance of the data-generating distribution exists as opposed to stronger ones such as Gaussianity) and hence, outliers occur naturally as part of the data. This in contrast to an adversary maliciously corrupting the datapoints in the adversarial setting. Here, estimators such as the empirical mean remain *consistent* but suffer from poor *statistical* performance and the emphasis is on designing estimators which avoid this degradation. The performance of these estimators is evaluated based on the dependence of r_δ on the number of data points n, dimension d, and the failure probability δ. For instance, the empirical mean, achieves the following disappointing rate in the finite variance setting:

$$r_\delta = O\left(\sqrt{\frac{d}{n\delta}}\right)$$

via Chebyshev's inequality which is unfortunately tight for the empirical mean.

In one dimension, optimal estimators based on the median-of-means framework were devised (and independently discovered) in a series of classical works [52, 36, 2]. In the one-dimensional setting, these estimators achieved the following:

$$r_\delta = O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right)$$

which is known to be optimal. On the other hand, the *high-dimensional* setting remained open till the pioneering work of Lugosi and Mendelson [48] whose estimator achieves the optimal sub-Gaussian rate. That is, they proposed an estimator which satisfies:

$$r_\delta = O\left(\sqrt{\frac{d + \log(1/\delta)}{n}}\right).$$

Surprisingly, this is the rate obtained by using the empirical mean on *Gaussian* data which is also known to be statistically optimal. This is despite making *no* higher-order assumptions about the data distribution. Unfortunately, this estimator is not known to be computable efficiently. A computationally efficient version by Hopkins [30] with the same guarantees followed shortly after along with alternative approaches [47] achieving the same guarantees. Since then, these ideas have been improved and extended to numerous other settings leading to estimators with strong statistical and computational performance [10, 16, 39, 51]. Interestingly, some recent works have also focused on the strong connections between the heavy-tailed and adversarially robust settings yielding estimators simultaneously robust to both corruption models [16, 47, 33, 20]. An alternative line work has also incorporated privacy guarantees into these estimators [44, 31, 37, 32].

Interestingly, some recent work in this line has sought to restore the affine-equivariant properties of these robust estimators emphasized in classical work. We draw attention to three recently developed estimators: the work of Depersin and Lecue [15], the setting considered by Duchi, Haque, and Kuditipudi [26] and Brown, Hopkins, and Smith [7] which in turn build upon approaches by Brown, Gaboardi, Smith, Ullman, and Zakynthinou [6], and the recent result of Lugosi and Mendelson [46]. Depersin and Lecué [15] consider the Stahel-Donoho estimator and show that it achieves sub-Gaussian statistical performance. On the other hand, in [6], the authors construct affine-equivariant estimators with sub-Gaussian error and strong privacy guarantees with subsequent work [26, 7] achieving computational efficiency. Finally, sub-Gaussian estimators with direction-dependent accuracy are developed in [46]. Unfortunately, all these estimators require stronger assumptions on the distribution (beyond a minimal assumption of the existence of a variance) ranging from the estimability of the covariance matrix [15] to higher order moment assumptions [46, 6, 7]. Furthermore, these bounds in [46] scale with the expected *Euclidean* deviation of a sample from its mean which when evaluated in an isotropic transformation of the data could be arbitrarily large.

1.3 Our Contributions

In the context of the discussion in Section 1.2, we now describe our contributions to the computational and statistical understanding of mean estimation. The work described in this thesis marks developments along three central facets:

Algorithmic Framework [10]. Our first contribution describes an efficient algorithmic framework for heavy-tailed estimation. As noted previously, the estimator proposed by Hopkins [30] is computationally efficient. However, its technical complexity leads both to poor theoretical runtimes and difficulties in extending it to other settings limiting its practical application. We propose a simple algorithmic framework which significantly reduces the runtime of this estimator while also being easily extensible to other estimation problems [11]. This work is described in Chapter 2.

Statistical Frontiers [12]. Next, we consider the statistical limits of robust estimation. Lugosi and Mendelson [48] showed that when the *variance* exists, the mean is estimable as well as would be possible if one had *Gaussian* data. However, in application domains such as quantitative finance and operations research, even this assumption may not hold true.

We formally investigate the impact that these large noise environments have on statistical performance in Chapter 3 where we show that the optimal sub-Gaussian rate is no longer possible. We provide a tight characterization of the optimal rates in this setting with a computationally efficient estimator (building on Chapter 2) and present novel lower bounds witnessing the rate.

Necessary Compromises [8]. Finally, in recent work, we attempt to restore the stability properties of classical estimators to the current wave of quantitatively optimal estimators. Modern estimators while possessing strong quantitative guarantees lack the strong stability guarantees enjoyed by classical estimators. Meanwhile, classical estimators do not possess strong quantitative performance guarantees. Strikingly, we show that there exists a *necessary* trade-off between these two desiderata in *both* the heavy-tailed and adversarial corruption scenarios. We show that *any* affine-equivariant estimator can only achieve the following rate:

$$r_\delta = \widetilde{\Omega}\left(\sqrt{\frac{d\log(1/\delta)}{n}} + \sqrt{d\eta}\right)$$

a drastic degradation from the *sub-Gaussian* rate previously encountered. Furthermore, the dependence on the corruption factor also degrades by a factor of $\sqrt{d}$. We develop a novel high-dimensional median which achieves this rate and prove statistical lower bounds for the specific class of affine-equivariant estimators. Our estimator addresses the quantitative deficiencies of classical work while also enjoying their natural stability properties. This work is presented in Chapter 4.

Chapter 2

Algorithmic Framework

In this chapter, we present a simple algorithmic framework for heavy-tailed estimation, specialized to the problem of heavy-tailed mean estimation[1]. In this setting, an assumption of *finite variance* is imposed on the data generating distribution:

Assumption 2.0.1. The distribution, P, satisfies:

$$\mathop{\mathbb{E}}_{X \sim P}\left[(X-\mu)(X-\mu)^\top\right] = \Sigma.$$

Note that no additional assumptions are imposed on P and specifically, avoid those on its higher order moments which allows for modeling of heavy-tailed behavior in the data. As discussed in Chapter 1, Lugosi and Mendelson [48] devised an estimator which achieves the optimal *sub-Gaussian* rate of:

$$r_\delta = O\left(\sqrt{\frac{\mathrm{Tr}(\Sigma) + \|\Sigma\| \log(1/\delta)}{n}}\right)$$

while Hopkins [30] proposed the first computationally efficient variant. However, the estimator in [48] is not known to be efficiently computable while that in [30] is technically complicated and hence, incurs exorbitantly large runtimes while also being challenging to extend to other settings. Our framework allows for simpler constructions of efficient estimators and is extensible to other estimation problems.

Through the remainder of the chapter, we overview the one-dimensional setting and provde intuition for the median-of-means framework in Section 2.1, we then discuss some high-dimensional extensions in Section 2.2 before presenting a simplified (but computationally inefficient) version of our estimator in Section 2.3. We formally describe our estimator and establish its runtime and accuracy guarantees in Section 2.4 and finally, Section 2.5 contains concentration results used in our proof.

[1]However, the framework has been employed to construct efficient algorithms for other estimation problems such as linear regression and covariance estimation [11]. Furthermore, an observation of Depersin and Lecué [16] implies that this algorithm is also *adversarially* robust.

2.1 One-dimensional Setting

Here, we will present a proof of the result for the one-dimensional setting which will help illustrate the median-of-means framework for heavy-tailed estimation. Note that Assumption 2.0.1 simplifies to the following:

$$\mathop{\mathbb{E}}_{X \sim P} \left[(X - \mu)^2 \right] = \sigma^2$$

and the optimal achievable rate is characterized in the following theorem:

Theorem 2.1.1 ([52, 36, 2])**.** *Let $\boldsymbol{X} = X_1, \dots, X_n$ be iid random variables with mean μ and variance σ^2. There exist absolute constants $C, c > 0$ and an estimator which, when given inputs $\boldsymbol{X}$ and a target confidence δ satisfying $\log(1/\delta) < cn$, returns a point x^* with:*

$$|x^* - \mu| \leqslant C\sigma \sqrt{\frac{\log(1/\delta)}{n}},$$

with probability at least $1 - \delta$.

Proof. This proof and all of the results presented in this thesis utilize the median-of-means framework illustrated in Fig. 2.1. The data is first split into k equally sized batches, the empirical mean is computed within each batch, and the k estimates thus obtained are combined through an appropriate aggregation function, f. Here, we will simply use the one-dimensional median.

We have by a simple application of Chebyshev's inequality:

$$\forall i \in [k] : \mathbf{Pr} \left\{ |\widehat{\mu}_i - \mu| \leqslant 4\sigma \sqrt{\frac{k}{n}} \right\} \geqslant \frac{9}{10}.$$

Now setting $k = C \log(1/\delta)$, an application of Hoeffding's inequality (Theorem A.1.1), now yields:

$$\mathbf{Pr} \left\{ \sum_{i=1}^{k} \mathbf{1} \left\{ |\widehat{\mu}_i - \mu| \leqslant 4\sigma \sqrt{\frac{k}{n}} \right\} \geqslant \frac{3k}{4} \right\} \geqslant 1 - \delta.$$

We condition on the above event and the theorem follows as on the above event:

$$|\widehat{\mu} - \mu| \leqslant 4\sigma \sqrt{\frac{k}{n}}.$$

□

We observe that the parameter k represents a trade-off between accuracy and reliability with larger values of k leading to more reliable but less accurate estimates and vice versa. The proof which is relatively simple in the one-dimensional case is challenging to extend to the high-dimensional setting due to the lack of a obvious notion of a high-dimensional median. We discuss several candidates in the subsequent section.

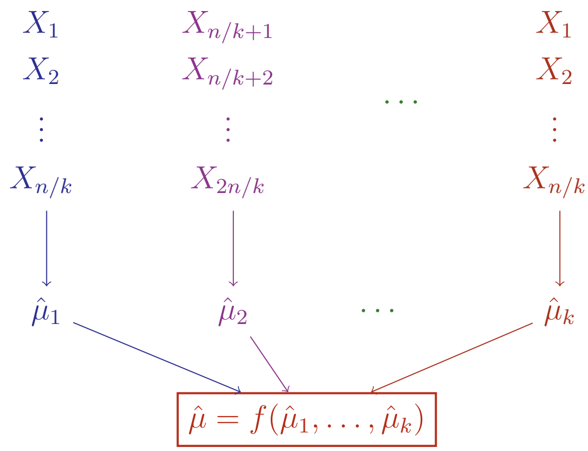


Figure 2.1: The median-of-means framework for robust estimation. The data is first split into k equally sized batches, the empirical mean is computed in each batch, and the estimates are finally combined with an appropriate aggregation function, f. f is typically chosen to correspond to some notion of a median.

2.2 High-dimensional Setting

Unfortunately, generalizing the standard one-dimensional median to higher dimensions is not straightforward. Here, we will describe a few proposals and the general principles underlying them. This generality will allow for easy comparison of these different estimators in later chapters. One of the first high-dimensional generalizations to find use in robust statistics was the Tukey median [61]. Amongst its other appealing properties, the Tukey median is also *affine-equivariant*. The starting point in describing the Tukey median is the concept of a *depth function*. This function measures how *close to the center* a point is with respect to a set of points. For instance, the depth function corresponding to the Tukey median is defined as follows for $\boldsymbol{Y} = \{y_i\}_{i=1}^n \subset \mathbb{R}$:

$$D_\tau^1(y; \boldsymbol{Y}) = \min\left(|\{i : y_i \geqslant y\}|, |\{i : y_i \leqslant y\}|\right).$$

The Tukey Median of a set of points $\boldsymbol{X} = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ is now defined below:

$$\widehat{\mu}_\tau(\boldsymbol{X}) = \arg\max D_\tau^d(x; \boldsymbol{X}) \text{ where } D_\tau^d(x; \boldsymbol{X}) = \min_{\|u\|=1} D_\tau^1\left(\langle u, x\rangle; \{\langle u, x\rangle\}_{i=1}^n\right).$$

When $d = 1$, note that this reduces to the standard one-dimensional Median. In addition, the breakdown properties of the Tukey Median and other affine-equivariant estimators have been closely investigated by Maronna [49] and Huber [34]. These works concluded that the *breakdown point* [22] these well-known affine equivariant estimates is at most $1/(d+1)$ without any additional assumptions on the point set such as symmetricity. This somewhat disappointing discovery led to the search for estimators with improved breakdown properties.

One of the first such approaches was the Stahel-Donoho estimator independently discovered by Stahel [57] and Donoho [23]. Here, one utilizes an alternative notion of *outlyingness* where $\mathrm{Med}(\boldsymbol{Y})$ denotes the median of $\boldsymbol{Y}$:

$$D^1_{\mathrm{SD}}(y; \boldsymbol{Y}) = \frac{|y - \mathrm{Med}(\boldsymbol{Y})|}{\mathrm{MAD}(\boldsymbol{Y})} \text{ where } \mathrm{MAD}(\boldsymbol{Y}) = \mathrm{Med}\left(\{|y_i - \mathrm{Med}(\boldsymbol{Y})|\}_{i=1}^n\right).$$

The Stahel-Donoho estimate is a point with *minimum* outlyingness:

$$\widehat{\mu}_{\mathrm{SD}}(\boldsymbol{X}) = \arg\min D^d_{\mathrm{SD}}(x; \boldsymbol{X}) \text{ where } D^d_{\mathrm{SD}}(x; \boldsymbol{X}) = \max_{\|u\|=1} D^1_{\mathrm{SD}}\left(\langle u, x\rangle; \{\langle u, x_i\rangle\}_{i=1}^n\right)$$

This estimator is known to have a breakdown point approaching 1/2. However, all these approaches suffer from the following drawbacks:

1. There are no quantitative bounds on their performance.

2. Furthermore, attainable bounds depend on the non-degeneracy of the dataset with error bounds growing arbitrarily large as the dataset approaches degeneracy.

Since the Stahel-Donoho estimator, numerous alternative approaches with differing notions of depth have been proposed: these include estimators based on the simplicial volume [54], S-estimation [55], the minimum volume ellipsoid [56] and the simplicial depth [41]. In addition, the robustness [49, 63, 14, 25, 45, 24, 40, 43, 42] and consistency properties [64, 65] of these estimators have been studied. However, despite this interest, there exist no *quantitative* accuracy guarantees in terms of the number of data points n, dimension d, failure probability δ, and corruption fraction η exist for these estimators.

The appropriate generalization, utilized in the statistically optimal estimator constructed by Lugosi and Mendelson [48] which also achieves the optimal breakdown point of 1/2, is based on a notion of outlyingness closely related to the Stahel-Donoho estimator with the main difference being the lack of the scale based normalization in the denominator:

$$D^1_{\mathrm{LM}}(y; \boldsymbol{Y}) = \frac{|y - \mathrm{Med}(\boldsymbol{Y})|}{\cancel{\mathrm{MAD}(Y)}}.$$

And the corresponding high-dimensional median is defined as follows:

$$\widehat{\mu}_\tau(\boldsymbol{X}) = \arg\min D^d_{\mathrm{LM}}(y; \boldsymbol{X}) \text{ where } D^d_{\mathrm{LM}}(x; \boldsymbol{X}) = \min_{\|u\|=1} D^1_{\mathrm{LM}}\left(\langle u, x\rangle; \{\langle u, x_i\rangle\}_{i=1}^n\right).$$

While the corresponding median is no longer affine-equivariant, this subtle change now allows for an estimator with quantitatively *optimal* performance. Furthermore, as we will see, a suitable approximation of this median is *efficiently computable* as opposed to the alternative notions discussed here.

Finally, we briefly describe the geometric insight underlying the analysis of Lugosi and Mendelson. They establish that for all directions, v, projections of *most* of the bucketed means, $\{\widehat{\mu}_i\}_{i\in[k]}$, lie close to the projections of the true mean μ along v. However, the precise set that satisfy this may differ with the direction considered. This property is illustrated in Fig. 2.2. Formally, they establish the following lemma:

Lemma 2.2.1 ([48]). *There exist absolute constants $c, C_1, C_2 > 0$ such that the following holds. Let $\boldsymbol{X} = X_1, \dots, X_n$ be n iid random vectors with mean μ and covariance Σ. For $\delta \in (0,1)$ with $\log(1/\delta) < cn$, $k = C_1 \log(1/\delta)$ and bucketed means $\widehat{\mu}_1, \dots, \widehat{\mu}_k$ produced from $\boldsymbol{X}$, we have:*

$$\forall \|v\| = 1 : \sum_{i=1}^{n} \mathbf{1}\left\{ |\langle v, \widehat{\mu}_i \rangle - \langle v, \mu \rangle| \leqslant C_2 \sqrt{\frac{\mathrm{Tr}(\Sigma) + \|\Sigma\| \log(1/\delta)}{n}} \right\} \geqslant 0.95k$$

with probability at least $1 - \delta$.

We will not prove this result here but it will be implied by a stronger result that will be required in subsequent analysis. For now, observe that this implies a point exists with at outlyingness at most:

$$C \sqrt{\frac{\mathrm{Tr}(\Sigma) + \|\Sigma\| \log(1/\delta)}{n}}$$

and furthermore, a simple analysis shows that *any* such point must be close to the true mean μ. The approach of Hopkins [30] uses a semi-definite relaxation of the Lugos-Mendelson median; i.e. it relaxes the problem of directly finding the median point. We take an alternative approach while leads to a simpler algorithm with much smaller runtimes. In subsequent sections, we abstract out the geometric concentration property in Lemma 2.2.1 and describe how we use it to construct *efficient* algorithms.

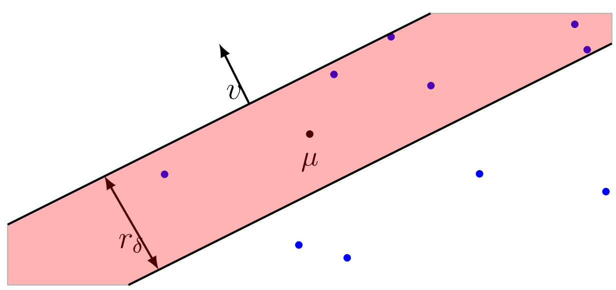


Figure 2.2: Illustration of the geometric property established in the analysis of Lugosi and Mendelson [48]. Formally, for every unit vector v, at least $0.9k$ are within a distance of r_δ of μ when projected onto v. Note, however, that the precise subset that satisfy this may differ across v.

2.3 Testing-to-estimation Warm Up

We present in this section a simple descent based algorithm. This algorithm is computationally intractable but is simple to analyze and much of the intuition behind its analysis transfers to the computationally efficient version as well. The main driving principle behind the framework is that for many robust estimation problems, *testing* whether a given candidate point x is close to the mean is often significantly easier than directly finding an accurate estimate. The key insight of our approach is that the *solutions* to these testing problems also contain information about how to improve the current estimate. While they do not immediately yield an optimal solution, a small number of iterations of this procedure suffice to establish our optimal guarantees. Furthermore, the simplicity of the testing procedure leads to substantial improvements to computational efficiency over prior work.

Intuition

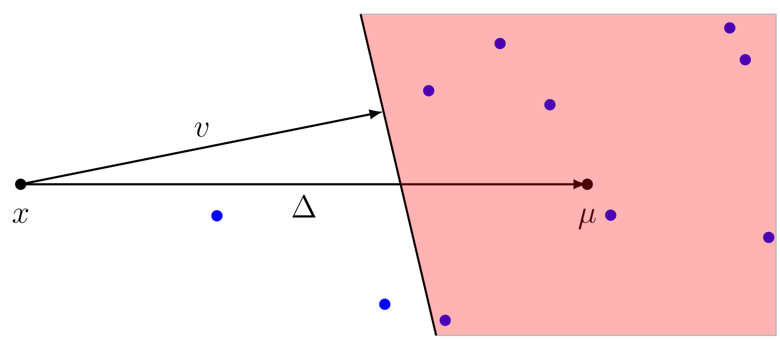


Figure 2.3: The direction v solution to **MTE** is well aligned with the vector joining the current estimate x to the true mean μ.

We provide some intuition for our procedure specialized to mean estimation and present the testing problem utilized here. Drawing inspiration from Lugosi and Mendelson [48], who show that along any direction, most of the bucketed means, henceforth referred to as Z_i, are close to the mean, μ. Thus, to test whether a point, x, is far from the mean, it is sufficient to check whether there exists a direction where most of the Z_i are far away from x along that direction. This is formally expressed in the following polynomial optimization problem:

$$\begin{gathered}\max \sum_{i=1}^{k} b_i \\ b_i^2 = b_i \\ \|v\|^2 = 1 \\ b_i \langle v, Z_i - x \rangle \geqslant b_i^2 r \quad \forall i \in [k]\end{gathered} \tag{\textbf{MTE}}$$

This polynomial problem over the set of variables $b_1, \ldots, b_k$ and $v_1, \ldots, v_d$ is parameterized by $r > 0$, the current estimate $x \in \mathbb{R}^d$ and the bucketed means $\boldsymbol{Z} \in \mathbb{R}^{k \times d}$. Its polynomial constraints are encoding the number of Z_i beyond a distance r from x when projected along a direction v. Intuitively, this program tries to find a direction v so as to maximize the number of Z_i beyond a distance r from x along that direction. Observe from [48] that for an appropriate choice of r, along all directions v, a large fraction of the Z_i are close to the mean.

Formally, for all directions v, $|\{i : |\langle Z_i - \mu, v\rangle| \leqslant r\}| \geqslant 0.9k$ (see Lemma 2.2.1). Therefore this optimization problem has a large value when x is far from the mean and Δ, the unit vector along $\mu - x$ (see Fig. 2.3), can be used to certify this.

Strikingly, the direction v returned by the solution of the above problem also contains information about the location of the mean when r is chosen appropriately, which enables improvement of the quality of the current estimate. As illustrated in Fig. 2.3, the direction returned by this optimization problem is strongly correlated with the vector joining the current point x to the mean μ.

Therefore, moving a small distance along the vector v should intuitively take us closer to the mean. Given solutions to the polynomial optimization problem **MTE**, we may iteratively improve our estimate until no further change is necessary.

Algorithm 1 Mean Estimation

Input: Data Points $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, Target Confidence δ
$x^\dagger \leftarrow$ Initial Mean Estimate$(\boldsymbol{X})$, $T \leftarrow C\log(n)$, $k \leftarrow C\log(1/\delta)$
Split data into k bins, $\mathcal{B}_i$ consisting of $\{X_{(i-1)\frac{n}{k}+j}\}_{j=1}^{n/k}$
$Z_i \leftarrow$ Mean$(\mathcal{B}_i)$ $\forall\, i \in [k]$ and $\boldsymbol{Z} \leftarrow (Z_1, \ldots, Z_k)$
$x^* =$ Gradient Descent$(\boldsymbol{Z}, x^\dagger, T)$
Return: x^*

Algorithm 2 Gradient Descent

Input: Bucket Means $\boldsymbol{Z} \in \mathbb{R}^{k \times d}$, Initialization $x^\dagger$, Number of Iterations T
$x^*, x_0 \leftarrow x^\dagger$ and $D^*, D_0 \leftarrow \infty$
for $t = 0 : T$ **do**
  $D_t \leftarrow$ Distance Estimation$(\boldsymbol{Z}, x_t)$
  $g_t \leftarrow$ Gradient Estimation$(\boldsymbol{Z}, x_t)$
  if $D_t < D^*$ **then**
    $x^* \leftarrow x_t$
    $D^* \leftarrow D_t$
  end if
  $x_{t+1} \leftarrow x_t + \frac{1}{20} D_t g_t$
end for
Return: x^*

Algorithm 3 Distance Estimation

Input: Data Points $\boldsymbol{Z} \in \mathbb{R}^{k \times d}$, Current point x
$D^* = \max\{r > 0 : \textbf{MTE}(x, r, \boldsymbol{Z}) \geqslant 0.9k\}$
Return: D^*

Algorithm 4 Gradient Estimation

Input: Data Points $\boldsymbol{Z} \in \mathbb{R}^{k \times d}$, Current point x
$D^* =$ Distance Estimation$(\boldsymbol{Z}, x)$
$(b, g) =$ **MTE**$(x, D^*, \boldsymbol{Z})$
Return: g

Algorithm 5 Initial Mean Estimate

Input: Set of data points $\boldsymbol{X} = \{X_i\}_{i=1}^n$
$\hat{\mu} \leftarrow \arg\min_{X_i \in \boldsymbol{X}} \min \left\{ r > 0 : \sum_{j=1}^n \mathbf{1}\{\|X_j - X_i\| \leqslant r\} \geqslant 0.6n \right\}$
Return: $\hat{\mu}$

Algorithm

In this section we put the intuition provided previously into practice and propose a procedure that estimates the mean in the ideal situation where **MTE** can be exactly solved (the method is formally described in Algorithm 1):

1. First, following the median of means framework, the samples X_i are divided into k buckets and the mean of the samples within each bucket is computed as $Z_i = \frac{k}{n} \sum_{j=(i-1)n/k}^{in/k+1} X_j$.
2. Second, the estimate of the mean is iteratively updated using a descent-based approach, using the solution to **MTE**. As mentioned in Section 2.3, we need to run **MTE** with an appropriate choice of r for the solution v to be correlated with the direction $x - \mu$. In the Distance Estimation step of our algorithm, we estimate a suitable choice of r (see Algorithm 3). This value of r is subsequently used in the Gradient Estimation step, to obtain an appropriate descent direction g (see Algorithm 4).

From this point on, we refer to the solution of **MTE** as $(b, v) =$ **MTE**$(x, r, \boldsymbol{Z})$.

Analysis warm-up

In this simplified setting, we provide an analysis of our method and show that it obtains the optimal sub-Gaussian rate. This is formally expressed in the following theorem.

Theorem 2.3.1. *There exist constants $c, C > 0$ such that the following hold. Let $\boldsymbol{X} = (X_1, \ldots, X_n) \in \mathbb{R}^{n \times d}$ be n i.i.d. random vectors with mean μ and covariance Σ. Then Algorithms 1 and 2 when instantiated with Algorithms 3 and 4 and run with inputs $\boldsymbol{X}$ and target confidence δ with $\log(1/\delta) \leqslant cn$ returns x^* satisfying:*

$$\|x^* - \mu\| \leqslant C \sqrt{\frac{\mathrm{Tr}(\Sigma) + \|\Sigma\| \log(1/\delta)}{n}},$$

with probability at least $1 - \delta$.

The main step of the proof is in the analysis of the gradient descent algorithm, Algorithm 2. Algorithm 1 pre-processes the dataset, $\boldsymbol{X}$, to produce the bucketed estimates, $\boldsymbol{Z}$, an initialization $x^\dagger$ and iteration count T for Algorithm 2. The guarantees of Algorithm 5 (see Lemma A.2.1 for a simple proof) ensure that the initialization is within $O(\sqrt{\mathrm{Tr}(\Sigma)})$ of the true mean. Hence, the bulk of the proof is in the analysis of Algorithm 2, the main steps of which we outline below:

1. **Distance Estimation:** We show that when the current estimate x is far from μ, Algorithm 3 accurately estimates the distance of x to μ. See Lemma 2.3.4.

2. **Gradient Estimation:** Next, we show that when x is far away from the mean μ, the vector g obtained by solving **MTE** in Algorithm 4 is well aligned with the vector joining the current point x to the mean μ. See Lemma 2.3.5.

3. **Gradient Descent:** Combining the previous two steps, we prove that we eventually converge to a good approximation to the mean.

In the proofs for the correctness of Algorithm 2, we make use of the following assumptions[2] which formalize the insight of Lugosi and Mendelson [48].

Assumption 2.3.2. Let $\boldsymbol{Z} = (Z_1, \dots, Z_k)$ satisfy for some $\widetilde{\mu} \in \mathbb{R}^d$ and $r^* > 0$:

$$\forall v \in \mathbb{R}^d, \|v\| = 1 : |\{i : \langle Z_i - \widetilde{\mu}, v \rangle \geqslant r^*\}| \leqslant 0.05k.$$

Furthermore, we assume that the initialization $x^\dagger$ satisfies for $D > 0$:

$$\|x^\dagger - \widetilde{\mu}\| \leqslant D.$$

We now present our main technical theorem on the correctness of Algorithm 2.

Theorem 2.3.3. *There exist constants $C_1, C_2 > 0$ such that the following holds. Let $\boldsymbol{Z} = (Z_1, \dots, Z_k)$ and $\widetilde{\mu}$ satisfy Assumption 2.3.2 for some $r^*, D > 0$ and suppose $T \geqslant C_1 \log(D/r^*)$. Then, Algorithm 2 when instantiated with Algorithms 3 and 4 and when input $\boldsymbol{Z}$, $x^\dagger$, and T, outputs x^* satisfying:*

$$\|x^* - \widetilde{\mu}\| \leqslant C_2 r^*.$$

Before establishing Theorem 2.3.3, we first prove that the **Distance Estimation** (Algorithm 3) and **Gradient Estimation** (Algorithm 4) steps are correct. We start with Algorithm 3.

[2] Note that we analyze Algorithm 2 in slightly greater generality in anticipation of its eventual application in subsequent chapters.

Lemma 2.3.4. *Let Assumption 2.3.2 hold for* $\boldsymbol{Z}$ *for some* $\widetilde{\mu} \in \mathbb{R}^d, r^* > 0$*. Now, for any* $x \in \mathbb{R}^d$*, Algorithm 3 on input* $\boldsymbol{Z}$ *and* x *returns a distance estimate* D^* *which satisfies:*

$$|D^* - \|x - \widetilde{\mu}\|| \leqslant r^*.$$

Proof. We first prove the lower bound $\|x - \widetilde{\mu}\| - r^* \leqslant D^*$. We may assume that $\|x - \widetilde{\mu}\| > r^*$, as the alternate case is trivially true. For $r = \|x - \widetilde{\mu}\| - r^*$, we can simply pick the vector $v = \Delta$ where Δ is the unit vector along $\widetilde{\mu} - x$. Under Assumption 2.3.2, we have that for at least $0.95k$ points:

$$\langle Z_i - x, v\rangle = \langle Z_i - \widetilde{\mu}, v\rangle + \langle \widetilde{\mu} - x, v\rangle \geqslant \|x - \widetilde{\mu}\| - r^* = r.$$

This implies the lower bound holds when $\|x - \widetilde{\mu}\| > r^*$.

For the upper bound $D^* \leqslant \|x - \widetilde{\mu}\| + r^*$, suppose, for the sake of contradiction, there is a value of $r > \|x - \widetilde{\mu}\| + r^*$ for which the optimal value of **MTE**$(x, r, \boldsymbol{Z})$ is greater than $0.9k$. Let v be the solution of **MTE**$(x, r, \boldsymbol{Z})$. This means that for $0.9k$ of the Z_i, we have:

$$\langle Z_i - \widetilde{\mu}, v\rangle = \langle Z_i - x, v\rangle + \langle x - \widetilde{\mu}, v\rangle \geqslant r - \|x - \widetilde{\mu}\| > r^*.$$

This contradicts Assumption 2.3.2, proving the upper bound. □

Next, we move on to the **Gradient Estimation** step (Algorithm 4).

Lemma 2.3.5. *Let Assumption 2.3.2 hold for* $\boldsymbol{Z}$ *for some* $\widetilde{\mu} \in \mathbb{R}^d, r^* > 0$*. Now, let* $x \in \mathbb{R}^d$ *satisfying:*

$$\|x - \widetilde{\mu}\| \geqslant 4r^*. \tag{2.1}$$

Then, letting Δ *denote the unit vector along* $\widetilde{\mu} - x$*, Algorithm 4 on input* $\boldsymbol{Z}$ *and* x*, returns a gradient estimate* g *satisfying:*

$$\langle g, \Delta\rangle \geqslant \frac{1}{2}.$$

Proof. We have, from the definition of D^* in Algorithms 3 and 4, that for $0.9k$ of the Z_i, $\langle Z_i - x, g\rangle \geqslant D^*$. We also have, from Assumption 2.3.2, that $\langle Z_i - \widetilde{\mu}, g\rangle \leqslant r^*$ for $0.95k$ of the Z_i. Let Z_j satisfy both those inequalities. Therefore, for Z_j, the lower bound from Lemma 2.3.4 implies

$$\|\widetilde{\mu} - x\| - r^* \leqslant D^* \leqslant \langle Z_j - x, g\rangle = \langle Z_j - \widetilde{\mu}, g\rangle + \langle \widetilde{\mu} - x, g\rangle \leqslant r^* + \|\widetilde{\mu} - x\|\langle \Delta, g\rangle.$$

By rearranging the above inequality and using the assumption on $\|\widetilde{\mu} - x\|$ in Eq. (2.1), we get the required conclusion. □

We now use Lemmas 2.3.4 and 2.3.5 to establish Theorem 2.3.3.

Proof of Theorem 2.3.3. Let $\widetilde{r} = 4r^*$. To start with, define $\mathcal{G} = \{x : \|x - \widetilde{\mu}\| \leqslant \widetilde{r}\}$. We now consider two cases:

Case 1: None of the iterates x_t lie in $\mathcal{G}$. In this case, note that by Lemma 2.3.4 and the definition of $\widetilde{r}$, we have:

$$\frac{3}{4}\|x_t - \widetilde{\mu}\| \leqslant D_t \leqslant \frac{5}{4}\|x_t - \widetilde{\mu}\|. \tag{2.2}$$

Moreover, we have by the definition of the update rule of x_t in Algorithm 1:

$$\|x_{t+1} - \widetilde{\mu}\|^2 = \|x_t - \widetilde{\mu}\|^2 + \frac{1}{10}D_t\langle x_t - \widetilde{\mu}, g_t\rangle + \frac{D_t^2}{400} \leqslant \|x_t - \widetilde{\mu}\|^2 - \frac{D_t\|x_t - \widetilde{\mu}\|}{20} + \frac{D_t^2}{400}$$
$$\leqslant \|x_t - \widetilde{\mu}\|^2 - \frac{3}{80}\|x_t - \widetilde{\mu}\|^2 + \frac{1}{320}\|x_t - \widetilde{\mu}\|^2 \leqslant \frac{39}{40}\|x_t - \widetilde{\mu}\|^2,$$

where we use Lemma 2.3.5 for the first inequality and the inequalities in Eq. (2.2) for the second. An iterative application of the above inequality establishes Theorem 2.3.3 in this case.

Case 2: At least one of the iterates x_t lies in $\mathcal{G}$. We have from Lemma 2.3.4:

$$D_t \leqslant 5r^*.$$

At the completion of the algorithm, we have from another application of Lemma 2.3.4:

$$\|x^* - \widetilde{\mu}\| - r^* \leqslant D^* \leqslant D_t \leqslant 5r^*.$$

Re-arranging the above inequality proves Theorem 2.3.3 in this case as well.

The above two cases conclude the proof of the theorem. □

Finally, Theorem 2.3.1 follows from conditioning on events in Lemmas 2.2.1 and A.2.1 and a subsequent application of Theorem 2.3.3 for the correctness of Algorithm 2 with our setting of T and $x^\dagger$ in Algorithm 1. □

While Theorem 2.3.1 guarantees a sub-Gaussian rate, the algorithm is not efficient due to the non-convexity of **MTE**. In the next section, we consider a semi-definite relaxation which is efficiently solvable while also providing the same optimal guarantees.

2.4 Testing-to-estimation Efficient Variant

In this section, we define a semi-definite programming relaxation of the polynomial optimization problem **MTE**. We then design new Distance Estimation and Gradient Estimation algorithms that use the tractable solutions to the relaxation instead of the original polynomial optimization problem. We then use these solutions to update our mean estimate along the same lines as Section 2.3, albeit with some added technical difficulty.

The Semi-Definite Relaxation of MTE

Here, we propose a semidefinite programming relaxation of **MTE**, a variant of the Threshold-SDP from [30]. We first define a semidefinite matrix $X \in \mathbb{R}^{(k+d+1)\times(k+d+1)}$ symbolically indexed by 1, the variables b_i and v_j and denote by the vector $v_{b_i} \coloneqq (X_{b_i,v_1}, \ldots, X_{b_i,v_d})$:

$$\begin{gathered} \max \sum_{i=1}^{k} X_{1,b_i} \\ X_{1,b_i} = X_{b_i,b_i} \\ X_{1,1} = 1 \\ \sum_{j=1}^{d} X_{v_j,v_j} = 1 \\ \langle v_{b_i}, Z_i - x \rangle \geqslant X_{b_i,b_i} r \ \forall i \in [k] \\ X \succcurlyeq 0 \end{gathered} \tag{\textbf{MT}}$$

Similar to the polynomial optimization **MTE**, this optimization problem is also parameterized by a vector $x \in \mathbb{R}^d$, $r > 0$ and a dataset $\boldsymbol{Z}$. We refer to solutions of this program as $(X, m) = \mathbf{MT}(x, r, \boldsymbol{Z})$ with m denoting the optimal value and X the optimal solution.

Algorithm 6 Distance Estimation

Input: Data Points $\boldsymbol{Z} \in \mathbb{R}^{k\times d}$, Current point x
$D^* = \max\{r > 0 : \mathbf{MT}(x, r, \boldsymbol{Z}) \geqslant 0.9k\}$
Return: D^*

Algorithm 7 Gradient Estimation

Input: Data Points $\boldsymbol{Z} \in \mathbb{R}^{k\times d}$, Current point x
$D^* =$ Distance Estimation$(\boldsymbol{Z}, x)$
$(X, m) = \mathbf{MT}(x, D^*, \boldsymbol{Z})$
$X_v =$ Submatrix of X corresponding to the indices v_i
$g =$ Top singular vector of X_v
$\mathcal{H} = \{i : \langle Z_i - x, g \rangle \geqslant 0\}$
if $|\mathcal{H}| \geqslant 0.9k$ **then**
  Return: g
else
  Return: $-g$
end if

The main contribution of our paper is in showing that the solutions to the relaxation **MT** can be used to improve the estimate similarly to those of **MTE**. We redefine the Distance and Gradient Estimation steps in Algorithms 1 and 2 using **MT** in Algorithms 6 and 7.

Algorithm

To efficiently estimate the mean, we instantiate Algorithms 1 and 2 to use solutions of **MT** instead of **MTE**. The new Distance Estimation and Gradient Estimation procedures are desribed in Algorithms 6 and 7.

As opposed to the polynomial optimization problem, solutions to the relaxation may not necessarily return a single vector v but rather a semidefinite matrix which corresponds to the relaxation of v. This matrix may not uniquely determine a descent direction. We, therefore, round the solution to a provably good descent direction which we use to iteratively improve our estimate. It is noteworthy that the singular value decomposition does not provide a sign direction. Thankfully the correct orientation is easily determined from the data points.

To analyze the runtime of Algorithms 1 and 2 with Algorithms 6 and 7, we first note that the semidefinite relaxation has $O(k^2+d^2)$ variables. However, by projecting all the data down to a subspace containing the k bucket means, we may effectively reduce the number of variables to $O(k^2)$ with an $O(k^2 d)$ time pre-processing step. Therefore, we are now left with $O(k^2)$ variables. The runtime of interior point methods for solving semidefinite programs with $O(k^2)$ variables and $O(k)$ constraints is $O(k^{3.5})$ [1]. Furthermore, a single call of the Distance Estimation procedure can be efficiently implemented using $\widetilde{O}(1)$ rounds of binary search on the parameter r. Therefore, the total cost of a single call to Algorithm 6 is $\widetilde{O}(k^{3.5})$. Similarly, the total cost of a call to Algorithm 7 is $\widetilde{O}(k^{3.5})$. Since the cost of each iteration is dominated by a single call of Algorithms 6 and 7, the total cost per iteration is $\widetilde{O}(k^{3.5})$. Since, we only run $\widetilde{O}(1)$ iterations, the total cost of the Algorithms 1 and 2 instantiated with Algorithms 6 and 7 is $\widetilde{O}(k^{3.5}+k^2 d+nd)$.

Analysis

We proceed primarily as in the previous section, but with the added technical difficulties arising from the use of the semi-definite relaxation. Here, we establish the following efficient analogue of Theorem 2.3.1:

Theorem 2.4.1. *There exist absolute constants $C, c > 0$ such that the following hold. Let $\boldsymbol{X} = (X_1, \dots, X_n) \in \mathbb{R}^{n\times d}$ be n i.i.d. random vectors with mean μ and covariance Σ. Then Algorithms 1 and 2 when instantiated with Algorithms 6 and 7 and run with inputs $\boldsymbol{X}$ and target confidence $\delta \in (0, 1/2)$ with $\log(1/\delta) \leqslant cn$ returns x^* satisfying:*

$$\|x^* - \mu\| \leqslant C\sqrt{\frac{\mathrm{Tr}(\Sigma) + \|\Sigma\| \log(1/\delta)}{n}},$$

with probability at least $1-\delta$. Furthermore, the procedure has runtime $\widetilde{O}((\log(1/\delta))^{3.5} + d(\log(1/\delta))^2 + nd)$.

As before, we have three main steps in analyzing Algorithm 2

1. **Distance Estimation:** We show that the Distance Estimation step in Algorithm 6 provides an accurate estimate of the distance of the current point from the mean. See Section 2.4.

2. **Gradient Estimation:** Next, we show that when x is far away from the mean μ, the vector g output by Algorithm 7 is well aligned with the vector joining the current point x to the mean μ. See Section 2.4.

3. **Gradient Descent:** Combining the previous two steps, we prove that we eventually converge to a good approximation to the mean. See Section 2.4.

We now present the analogue of Assumption 2.3.2 used to analyze the relaxed variant:

Assumption 2.4.2. Let $\boldsymbol{Z} = (Z_1, \dots, Z_k)$ satisfy for some $\widetilde{\mu} \in \mathbb{R}^d$, $r^* > 0$:

$$\max_{X \in \mathcal{S}_r} \sum_{i=1}^{k} X_{b_i,b_i} \leqslant \frac{k}{20}.$$

for all $r \geqslant r^*$ where $\mathcal{S}_r$ denotes the set of feasible solutions for $\mathbf{MT}(\widetilde{\mu}, r, \boldsymbol{Z})$. Furthermore, we assume that the initialization $x^\dagger$ satisfies for $D > 0$:

$$\|x^\dagger - \widetilde{\mu}\| \leqslant D.$$

The above assumption is a strengthening of Assumption 2.3.2 for the case where we use **MT** instead of **MTE**. We use the following fact at several points in the subsequent analysis:

Remark 2.4.3. Note that Assumption 2.4.2 implies Assumption 2.3.2.

We prove that Assumption 2.4.2 holds with high probability in Section 2.5 (Lemma 2.5.1). The analysis uses standard techniques from empirical process theory and follows similar analyses from [48, 30]. Here, we restrict ourselves to the analysis of the gradient descent step where we establish the following analogue of Theorem 2.3.3.

Theorem 2.4.4. *There exists absolute constants $C_1, C_2 > 0$ such that the following holds. Let $\boldsymbol{Z} = (Z_1, \dots, Z_k)$ and $\widetilde{\mu}$ satisfy Assumption 2.4.2 for some $r^*, D > 0$ and suppose $T \geqslant C_1 \log(D/r^*)$. Then, Algorithm 2 when instantiated with Algorithms 6 and 7 and when input $\boldsymbol{Z}$, $x^\dagger$, and T, outputs x^* satisfying:*

$$\|x^* - \widetilde{\mu}\| \leqslant C_2 r^*.$$

We now proceed to establish that the Distance and Gradient Estimation steps in Algorithms 6 and 7 function as expected in the next two subsections before proving Theorem 2.4.4. We start with Distance Estimation.

Distance Estimation Step

Here, we analyze Algorithm 6. We show that an accurate estimate of the distance of the current point from the mean can be found. We begin with a lemma which shows that a feasible solution for $\mathbf{MT}(x, r, \boldsymbol{Z})$ can be converted to a feasible solution for $\mathbf{MT}(\widetilde{\mu}, r^*, \boldsymbol{Z})$ with a reduction in optimal value.

Lemma 2.4.5. *Let us assume Assumption 2.4.2. Let $X \in \mathbb{R}^{(k+d+1)\times(k+d+1)}$ be a positive semi-definite matrix, symbolically indexed by 1 and the variables b_i and v_j. Moreover, suppose that X satisfies:*

$$X_{1,1} = 1, \quad X_{b_i,b_i} = X_{1,b_i}, \quad \sum_{j=1}^{d} X_{v_j,v_j} = 1, \quad \sum_{i=1}^{k} X_{b_i,b_i} \geqslant 0.9k.$$

Then, there is a set of at least $0.85k$ indices $\mathcal{T}$ such that for all $i \in \mathcal{T}$:

$$\langle Z_i - \widetilde{\mu}, v_{b_i} \rangle < X_{b_i,b_i} r^*,$$

and a set of at least $k/3$ indices $\mathcal{R}$ such that for all $j \in \mathcal{R}$, we have $X_{b_j,b_j} \geqslant 0.85$.

Proof. We prove the first claim by contradiction. Firstly, note that X is infeasible for $\mathbf{MT}(\widetilde{\mu}, r^*, \boldsymbol{Z})$ as the optimal value for $\mathbf{MT}(\widetilde{\mu}, r^*, \boldsymbol{Z})$ is less than $k/20$ (Assumption 2.4.2) and that the only constraints of $\mathbf{MT}(\widetilde{\mu}, r^*, \boldsymbol{Z})$ violated by X are constraints of the form:

$$\langle Z_i - \widetilde{\mu}, v_{b_i} \rangle < X_{b_i,b_i} r^*.$$

Now, let $\mathcal{T}$ denote the set of indices for which the above inequality is violated. We can convert X to a feasible solution for $\mathbf{MT}(\widetilde{\mu}, r^*, \boldsymbol{Z})$ by setting to 0 the rows and columns corresponding to the indices in $\mathcal{T}$. Let X' be the matrix obtained by the above operation. We have from Assumption 2.4.2:

$$0.05k \geqslant \sum_{i=1}^{k} X'_{b_i,b_i} = \sum_{i=1}^{k} X_{b_i,b_i} - \sum_{i\in\mathcal{T}} X_{b_i,b_i} \geqslant 0.9k - |\mathcal{T}|,$$

where the last inequality follows from the fact that $X_{b_i,b_i} \leqslant 1$. By rearranging the above inequality, we get the first claim of the lemma.

For the second claim, let $\mathcal{R}$ denote the set of indices j satisfying $X_{b_j,b_j} \geqslant 0.85$. We have:

$$0.9k \leqslant \sum_{j=1}^{k} X_{b_j,b_j} = \sum_{j\in\mathcal{R}} X_{b_j,b_j} + \sum_{j\notin\mathcal{R}} X_{b_j,b_j} \leqslant |\mathcal{R}| + 0.85k - 0.85|\mathcal{R}| \implies \frac{k}{3} \leqslant |\mathcal{R}|.$$

This establishes the second claim of the lemma. $\square$

The following lemma shows the correctness of Algorithm 6 when the distance between $\widetilde{\mu}$ and a point x is small.

Lemma 2.4.6. *Assume Assumption 2.4.2. Suppose $x \in \mathbb{R}^d$ satisfies $\|x - \widetilde{\mu}\| \leqslant 20r^*$. Then, Algorithm 6 on input $\boldsymbol{Z}$ and x, returns a value D^* satisfying*

$$D^* \leqslant 25r^*.$$

Proof. Let $r' = 25r^*$. Suppose that the optimal value of $\mathbf{MT}(x, r', \boldsymbol{Z})$ is greater than $0.9k$ and let an optimal solution be X. Let $\mathcal{R}$ and $\mathcal{T}$ denote the two sets from Lemma 2.4.5 and $j \in \mathcal{R} \cap \mathcal{T}$. We have:

$$0.85r' \leqslant \langle Z_j - x, v_{b_j} \rangle = \langle Z_j - \widetilde{\mu}, v_{b_j} \rangle + \langle \widetilde{\mu} - x, v_{b_j} \rangle < r^* + \|x - \widetilde{\mu}\|,$$

where the first inequality follows from the fact that $j \in \mathcal{R}$ and the fact that X is feasible for $\mathbf{MT}(x, r', \boldsymbol{Z})$ and the last inequality follows from the inclusion of j in $\mathcal{T}$ and Cauchy-Schwarz.

By plugging in the bounds on r', we get:

$$\|x - \widetilde{\mu}\| > 0.85r' - r^* > 20r^*$$

which is a contradiction and proves the lemma. □

The next lemma analyzes the case where the candidate point x is far from $\widetilde{\mu}$ and concludes the analysis of Algorithm 6.

Lemma 2.4.7. *Assume Assumption 2.4.2. Suppose $x \in \mathbb{R}^d$ satisfies $\|x - \widetilde{\mu}\| \geqslant 20r^*$. Then, Algorithm 6 on input $\boldsymbol{Z}$ and x, returns a value D^* satisfying*

$$0.95\|x - \widetilde{\mu}\| \leqslant D^* \leqslant 1.25\|x - \widetilde{\mu}\|.$$

Proof. Define Δ to be the unit vector in the direction of $x - \widetilde{\mu}$. From Assumption 2.3.2 (which is implied by Assumption 2.4.2), the number of Z_i satisfying $\langle Z_i - \widetilde{\mu}, \Delta \rangle \geqslant r^*$ is less than $k/20$. Therefore, we have that for at least $0.95k$ points:

$$\langle Z_i - x, -\Delta \rangle = \langle x - \widetilde{\mu} + \widetilde{\mu} - Z_i, \Delta \rangle = \|x - \widetilde{\mu}\| - r^* \geqslant 0.95\|x - \widetilde{\mu}\|.$$

With the monotonicity (Lemma A.2.2) of $\mathbf{MT}(x, r, \boldsymbol{Z})$ in r, this implies the lower bound.

For the upper bound, we show that the optimal value of $\mathbf{MT}(x, 1.25\|x - \widetilde{\mu}\|, \boldsymbol{Z})$ is less than $0.9k$. For the sake of contradiction, assume the contrary and let X be a feasible solution that achieves $0.9k$. Let $\mathcal{R}$ and $\mathcal{T}$ be the two sets from Lemma 2.4.5 and $j \in \mathcal{R} \cap \mathcal{T}$. We have for j:

$$\begin{aligned} 0.85(1.25\|x - \widetilde{\mu}\|) \leqslant X_{b_j,b_j} 1.25\|x - \widetilde{\mu}\| &\leqslant \langle Z_j - x, v_{b_j} \rangle = \langle Z_j - \widetilde{\mu}, v_{b_j} \rangle + \langle \widetilde{\mu} - x, v_{b_j} \rangle \\ &< X_{b_j,b_j} r^* + \|\widetilde{\mu} - x\| \end{aligned}$$

where the first inequality follows from $j \in \mathcal{R}$ and the last from $j \in \mathcal{T}$ and Cauchy-Schwarz. By re-arranging the above inequality, we get:

$$X_{b_j,b_j} > (1.0625\|x - \widetilde{\mu}\| - \|x - \widetilde{\mu}\|)(r^*)^{-1} > 1,$$

which is a contradiction. Therefore, we get from the monotonicity of $\mathbf{MT}(x, r, \boldsymbol{Z})$ (see Lemma A.2.2), that $D^* \leqslant 1.25\|x - \widetilde{\mu}\|$, concluding the proof of the lemma. □

Gradient Estimation Step

Next, we analyze the Gradient Estimation step of the algorithm. We show that an approximate gradient can be found as long as x is not too close to the mean $\widetilde{\mu}$. The following lemma shows that we obtain a non-trivial estimate of the gradient in Algorithm 7.

Lemma 2.4.8. *Assume Assumption 2.4.2. Suppose $x \in \mathbb{R}^d$ satisfies $\|x - \widetilde{\mu}\| \geqslant 20r^*$ and let Δ be the unit vector along $\widetilde{\mu} - x$. Algorithm 7 returns a g satisfying:*

$$\langle g, \Delta \rangle \geqslant \frac{1}{15}.$$

Proof. In the running of Algorithm 7, let X denote the solution of $\mathbf{MT}(x, D^*, \boldsymbol{Z})$. We begin by factorizing the solution X into $UU^\top$ with the rows of U denoted by u_1, $u_{b_1}, \ldots, u_{b_k}$ and $u_{v_1}, \ldots, u_{v_d}$. We also define the matrix $U_v = (u_{v_1}, \ldots, u_{v_d})$ in $\mathbb{R}^{(k+d+1)\times d}$. From the constraints in **MT**, we have:

$$X_{b_i,b_i} = \|u_{b_i}\|^2 \leqslant 1 \implies \|u_{b_i}\| \leqslant 1, \quad \sum_{j=1}^{d} X_{v_j,v_j} = \sum_{j=1}^{d} \|u_{v_j}\|^2 = \|U_v\|_F^2 = 1 \implies \|U_v\|_F = 1.$$

Let $\mathcal{R}$ and $\mathcal{T}$ denote the sets from Lemma 2.4.5 and $j \in \mathcal{T} \cap \mathcal{R}$. By noting that $v_{b_j} = u_{b_j}^\top U_v$, we have for j:

$$0.85D^* \leqslant \langle Z_j - \widetilde{\mu}, v_{b_j} \rangle + \langle \widetilde{\mu} - x, v_{b_j} \rangle \leqslant X_{b_j,b_j} r^* + u_{b_j}^\top U_v(\widetilde{\mu} - x),$$

where the first inequality follows from $j \in \mathcal{R}$ and the second from $j \in \mathcal{T}$. We get by rearranging the above equation and using our bound on D^* from Lemma 2.4.7:

$$0.80\|\widetilde{\mu} - x\| \leqslant 0.85D^* \leqslant X_{b_j,b_j} r^* + u_{b_j}^\top U_v(\widetilde{\mu} - x). \tag{2.3}$$

By rearranging Eq. (2.3), using Cauchy-Schwarz, $\|u_{b_i}\| \leqslant 1$ and the assumption on $\|x - \widetilde{\mu}\|$:

$$\|U_v(\widetilde{\mu} - x)\| \geqslant u_{b_j}^\top U_v(\widetilde{\mu} - x) \geqslant 0.75\|\widetilde{\mu} - x\|.$$

We finally get that:

$$\|U_v \Delta\| \geqslant 0.75.$$

Now, we have:

$$1 = \|U_v\|_F^2 = \|U_v \mathcal{P}_\Delta\|_F^2 + \|U_v \mathcal{P}_\Delta^\perp\|_F^2 \geqslant \|U_v \mathcal{P}_\Delta^\perp\|_F^2 + (0.75)^2 \implies \|U_v \mathcal{P}_\Delta^\perp\|_F \leqslant 0.67.$$

Let y be the top singular vector of X_v. Note that $X_v = U_v^\top U_v$ and y is also the top right singular vector of U_v. We have that:

$$0.75 \leqslant \|U_v y\| \leqslant \|U_v \mathcal{P}_\Delta y\| + \|U_v \mathcal{P}_\Delta^\perp y\| \leqslant \|\mathcal{P}_\Delta y\| + \|U_v \mathcal{P}_\Delta^\perp\|_F \leqslant \|\mathcal{P}_\Delta y\| + 0.67.$$

Hence, we have:

$$|\langle y, \Delta \rangle| \geqslant \frac{1}{15}.$$

Note that the algorithm returns either y or $-y$. Firstly, consider the case where $\langle y, \Delta \rangle > 0$. From Assumption 2.3.2 (implied by Assumption 2.4.2), we have for at least $0.95k$ points:

$$\langle Z_i - \widetilde{\mu}, y \rangle \leqslant r^*.$$

Therefore, we have for these $0.95k$ points:

$$\langle Z_i - x, y \rangle = \langle Z_i - \widetilde{\mu}, y \rangle + \langle \widetilde{\mu} - x, y \rangle \geqslant -r^* + \frac{20r^*}{15} > 0.$$

Therefore, when $\langle y, \Delta \rangle > 0$, we return y which satisfies $\langle \widetilde{\mu} - x, y \rangle > 0$. This implies the lemma in this case. The alternative where $\langle y, \Delta \rangle < 0$ is similar with $-y$ used instead of y. This concludes the proof of the lemma. □

Gradient Descent Step

Finally, we establish Theorem 2.4.4, the analogue of Theorem 2.3.3 for the relaxation. The proof follows along the lines of that of Theorem 2.3.3 with some minor modifications.

Proof of Theorem 2.4.4. Let $\mathcal{G} = \{x : \|x - \widetilde{\mu}\| \leqslant 20r^*\}$. We prove the theorem in two cases:

Case 1: None of the iterates x_t fall into $\mathcal{G}$. In this case, we have from Lemma 2.4.7:

$$0.95\|x_t - \widetilde{\mu}\| \leqslant D_t \leqslant 1.25\|x_t - \widetilde{\mu}\| \tag{2.4}$$

and we get:

$$\|x_{t+1} - \widetilde{\mu}\|^2 = \|x_t - \widetilde{\mu}\|^2 - 2\frac{D_t}{20}\langle g_t, \widetilde{\mu} - x_t \rangle + \frac{D_t^2}{400} \leqslant \|x_t - \widetilde{\mu}\|^2 - \frac{D_t\|\widetilde{\mu} - x_t\|}{150} + \frac{D_t^2}{400}$$
$$\leqslant \|x_t - \widetilde{\mu}\|^2 - D_t\left(\frac{\|\widetilde{\mu} - x_t\|}{150} - \frac{D_t}{400}\right) \leqslant \left(1 - \frac{1}{500}\right)\|x_t - \widetilde{\mu}\|^2.$$

where the first inequality follows from Lemma 2.4.8 and the last inequality follows by substituting the lower bound on D_t in the first term and the upper bound on D_t in the second term (Eq. (2.4)). An iterated application of the above inequality yields the theorem in this case.

Case 2: One of the iterates x_t falls into $\mathcal{G}$. If the algorithm returns an element from $\mathcal{G}$, the theorem is trivially true. From Lemma 2.4.6, we have for the iterate $x_t \in \mathcal{G}$:

$$D_t \leqslant 25r^*.$$

Therefore, we have at the completion of the algorithm a value $D^* \leqslant 25r^*$ together with x^* lying outside $\mathcal{G}$. Thus, we have from Lemma 2.4.7:

$$0.95\|x^* - \widetilde{\mu}\| \leqslant 25r^* \implies \|x^* - \widetilde{\mu}\| \leqslant 30r^*.$$

The previous two cases conclude the proof of the theorem. □

Wrapping up - Proof of Theorem 2.4.1

To conclude the proof of Theorem 2.4.1, note that the runtime guarantees follow from the analysis in Section 2.4. Therefore, the only remaining step is to verify that Assumption 2.4.2 holds with high probability. This follows from an application of Lemma A.2.1 to the random vectors $X_1, \dots, X_n$ and Lemma 2.5.1 to the bucketed means $\boldsymbol{Z}$. This concludes the proof of Theorem 2.4.1. □

2.5 Statistical Analysis

We show, here, that Assumption 2.4.2 holds with high probability. The main technical result of this section is the following lemma. The proof of the lemma relies on standard results from empirical process theory and is similar to previous analyses from [48, 30].

Lemma 2.5.1. *There exist absolute constants C_1, C_2 such that the following holds. Let $\delta \in (0,1)$ and $\boldsymbol{Y} = (Y_1, \dots, Y_k) \in \mathbb{R}^{k \times d}$ be k i.i.d.random vectors with mean μ and covariance Λ with $k \geqslant C_1 \log(1/\delta)$. Then, we have:*

$$\forall r \geqslant C_2 \left(\sqrt{\frac{\operatorname{Tr} \Lambda}{k}} + \sqrt{\|\Lambda\|} \right) : \max_{X \in \mathcal{S}_r} \sum_{i=1}^{k} X_{b_i, b_i} \leqslant \frac{k}{20},$$

with probability at least $1 - \delta$ where $\mathcal{S}_r$ denotes the feasible solutions of $\boldsymbol{MT}(\mu, r, \boldsymbol{Y})$.

The proof is carried out in two stages:

1. In the first, we show that the random variable in the conclusion of the lemma satisfies the bounded differences condition and hence, concentrates around its expectation.

2. Second, we bound the *expectation* of the variable and show that it is small.

We establish the bounded differences condition below.

Lemma 2.5.2. *Let $\boldsymbol{Y} = (Y_1, \dots, Y_k)$ be any set of k vectors in $\mathbb{R}^d$, $r > 0$, and $x \in \mathbb{R}^d$. Now, let $\boldsymbol{Y}' = (Y_1, \dots, Y_i', \dots, Y_k)$ be the same set of k vectors with the i^{th} vector replaced by $Y_i' \in \mathbb{R}^d$. If m and m' are the optimal values of $\boldsymbol{MT}(x, r, \boldsymbol{Y})$ and $\boldsymbol{MT}(x, r, \boldsymbol{Y}')$, we have:*

$$|m - m'| \leqslant 1$$

Proof. Firstly, assume that X is a feasible solution to $\mathbf{MT}(x, r, \boldsymbol{Y})$. Now, define X' as:

$$X'_{i,j} = \begin{cases} X_{i,j} & \text{if } i, j \neq b_i \\ 0 & \text{otherwise} \end{cases}$$

That is X' is equal to X except with the row and column corresponding to b_i being set to 0. We see that X' forms a feasible solution to $\mathbf{MT}(x, r, \boldsymbol{Y}')$. Therefore, we have that:

$$\sum_{j=1}^{k} X_{b_j,b_j} = \sum_{j=1,j\neq i}^{k} X'_{b_j,b_j} + X_{b_i,b_i} \leqslant \sum_{j=1,j\neq i}^{k} X'_{b_j,b_j} + 1 \leqslant m' + 1$$

where the bound $X_{b_i,b_i} \leqslant 1$ follows from the fact that the 2×2 sub-matrix of X formed by the rows and columns indexed by 1 and b_i is positive semidefinite and the constraint that $X_{b_i,b_i} = X_{1,b_i}$. Since the above series of equalities holds for all feasible solutions X of **MT**$(x, r, \boldsymbol{Y})$, we get:

$$m \leqslant m' + 1.$$

Through a similar argument, we also conclude that $m' \leqslant m + 1$. Putting the above two inequalities together, we get the desired conclusion. □

For the next few lemmas, we are concerned with the case where $x = \mu$ and we verify that the expectation is small. As a first step, we define the 2-to-1 norm of a matrix M.

Definition 2.5.3. The 2-to-1 norm of $M \in \mathbb{R}^{n\times d}$ is defined as

$$\|M\|_{2\to1} = \max_{\substack{\|v\|=1 \\ \sigma_i\in\{\pm1\}}} \sigma^\top M v = \max_{\|v\|=1} \|Mv\|_1$$

We consider the classical semidefinite programming relaxation of the 2-to-1 norm. To start with, we will define a matrix $X \in \mathbb{R}^{(n+d+1)\times(n+d+1)}$ with the rows and columns indexed by 1 and the elements σ_i and v_j. The semidefinite programming relaxation is defined as follows:

$$\begin{aligned} \max \sum_{i,j} M_{i,j} X_{\sigma_i,v_j} \\ X_{1,1} = 1 \\ \sum_{j=1}^{d} X_{v_j,v_j} = 1 \\ X_{\sigma_i,\sigma_i} = 1 \\ X \succcurlyeq 0 \end{aligned} \tag{TOR}$$

We now state a theorem of Nesterov as stated in [30]:

Theorem 2.5.4. *([53]) There is a constant $K_{2\to1} = \sqrt{\pi/2} \leqslant 2$ such that the optimal value, m, of the semidefinite programming relaxation TOR satisfies:*

$$m \leqslant K_{2\to1}\|M\|_{2\to1}.$$

In the next step, we will bound the expected 2-to-1 norm of $\boldsymbol{Z}$.

Lemma 2.5.5. *Let $\boldsymbol{Y} = (Y_1, \dots, Y_n) \in \mathbb{R}^{n \times d}$ be a set of n i.i.d. random vectors such that $\mathbb{E}[Y_i] = 0$ and $\mathbb{E}[Y_i Y_i^\top] = \Lambda$. Then, we have:*

$$\mathbb{E}\|\boldsymbol{Y}\|_{2\to 1} \leqslant 4\sqrt{n \operatorname{Tr}\Lambda} + n \max_{\|v\|=1} \mathbb{E}[|\langle v, Y\rangle|].$$

Proof. Denoting by Y and Y_i' random vectors that are independently and identically distributed as Y_i and by σ_i independent Rademacher random variables, we have:

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{Y}\|_{2\to 1}] &= \mathbb{E}\left[\max_{\|v\|=1} \sum_{i=1}^n |\langle Y_i, v\rangle|\right] = \mathbb{E}\left[\max_{\|v\|=1} \sum_{i=1}^n |\langle Y_i, v\rangle| + \mathbb{E}|\langle v, Y_i\rangle| - \mathbb{E}|\langle v, Y_i\rangle|\right] \\
&\leqslant \mathbb{E}\left[\max_{\|v\|=1} \sum_{i=1}^n |\langle Y_i, v\rangle| - \mathbb{E}|\langle Y_i', v\rangle|\right] + n \max_{\|v\|=1} \mathbb{E}[|\langle v, Y\rangle|] \\
&\leqslant \mathbb{E}\left[\max_{\|v\|=1} \sum_{i=1}^n \sigma_i(|\langle Y_i, v\rangle| - |\langle Y_i', v\rangle|)\right] + n \max_{\|v\|=1} \mathbb{E}\left[|\langle v, Y\rangle|\right].
\end{aligned}$$

For the first term, we get via a standard symmetrization argument:

$$\begin{aligned}
\mathbb{E}\left[\max_{\|v\|=1} \sum_{i=1}^n \sigma_i(|\langle Y_i, v\rangle| - |\langle Y_i', v\rangle|)\right] &\leqslant \mathbb{E}\left[\max_{\|v\|=1} \sum_{i=1}^n \sigma_i |\langle Y_i, v\rangle|\right] + \mathbb{E}\left[\max_{\|v\|=1} \sum_{i=1}^n -\sigma_i |\langle Y_i', v\rangle|\right] \\
&= 2\mathbb{E}\left[\max_{\|v\|=1} \sum_{i=1}^n \sigma_i |\langle v, Y_i\rangle|\right] \leqslant 4\mathbb{E}\left[\max_{\|v\|=1} \sum_{i=1}^n \sigma_i \langle v, Y_i\rangle\right] \\
&= 4\mathbb{E}\left[\left\|\sum_{i=1}^n \sigma_i Y_i\right\|\right] \leqslant 4\left(\mathbb{E}\left[\left\|\sum_{i=1}^n \sigma_i Y_i\right\|^2\right]\right)^{1/2} \\
&= 4\left(\mathbb{E} \sum_{1\leqslant i,j\leqslant n} \sigma_i \sigma_j \langle Y_i, Y_j\rangle\right)^{1/2} = 4\sqrt{n \operatorname{Tr}\Lambda},
\end{aligned}$$

where the second inequality follows from the Ledoux-Talagrand Contraction Theorem (Corollary A.1.9 of Theorem A.1.8). □

We now bound the expected value of $\mathbf{MT}(\mu, r, \boldsymbol{Y})$ by relating it to $\|\boldsymbol{Y}\|_{2\to 1}$.

Lemma 2.5.6. *Let $r > 0$ and $\boldsymbol{Y} = (Y_1, \dots, Y_k) \in \mathbb{R}^{k \times d}$ be k i.i.d. random vectors with mean μ and covariance Λ. Denoting by $\mathcal{S}$ the feasible solutions for $\boldsymbol{MT}(\mu, r, \boldsymbol{Y})$, we have:*

$$\mathbb{E} \max_{x \in \mathcal{S}} \sum_{i=1}^k X_{1,b_i} \leqslant \frac{1}{r}\left(5\sqrt{k \operatorname{Tr}\Lambda} + k \max_{\|v\|=1} \mathbb{E}[|\langle v, Y\rangle|]\right).$$

Proof. Firstly, let X be a feasible solution for **MT**$(\mu, r, \boldsymbol{Y})$. We construct a new, symmetric matrix W which is indexed by σ_i and v_j as opposed to b_i and v_j for X:

$$W_{\sigma_i,\sigma_j} = 4X_{b_i,b_j} - 2X_{1,b_i} - 2X_{1,b_j} + 1, \quad W_{v_i,v_j} = X_{v_i,v_j}, \quad W_{1,1} = 1,$$
$$W_{1,v_i} = X_{1,v_i}, \quad W_{1,\sigma_i} = 2X_{1,b_i} - 1, \quad W_{v_i,\sigma_j} = 2X_{v_i,b_j} - X_{1,v_i}.$$

We prove that W is a feasible solution to the SDP relaxation TOR of $\boldsymbol{Y} - \mu$. We see that:

$$W_{\sigma_i,\sigma_i} = 1 \text{ and } \sum_{i=1}^{d} W_{v_i,v_i} = 1.$$

Then, we simply need to verify that W is PSD. Let $w \in \mathbb{R}^{k+d+1}$ indexed by 1, σ_i and v_j. We construct from w a new vector w', indexed by 1, b_i and v_j and defined as follows:

$$w'_1 = w_1 - \sum_{i=1}^{k} w_{\sigma_i}, \quad w'_{b_i} = 2w_{\sigma_i}, \quad w'_{v_j} = w_{v_j}.$$

With w' defined as above, we have the following equality:

$$w^\top W w = (w')^\top X w' \geqslant 0.$$

Since the above condition holds for all $w \in \mathbb{R}^{k+d+1}$, we get that $W \succcurlyeq 0$. Therefore, we conclude that W is a feasible solution to the SDP relaxation TOR of $\boldsymbol{Y} - \mu$.

We bound the expected value of **MT**$(\mu, r, \boldsymbol{Y})$ as follows, denoting by v_{b_i} the vector $(X_{b_i,v_1}, \ldots, X_{b_i,v_d})$ and by v the vector $(X_{1,v_1}, \ldots, X_{1,v_d})$:

$$\begin{aligned}
\mathbb{E} \max_{X \in \mathcal{S}} \sum_{i=1}^{k} X_{1,b_i} &= \mathbb{E} \max_{X \in \mathcal{S}} \sum_{i=1}^{k} X_{b_i,b_i} \leqslant \frac{1}{r} \mathbb{E} \max_{X \in \mathcal{S}} \sum_{i=1}^{k} \langle v_{b_i}, Y_i - \mu \rangle \\
&= \frac{1}{2r} \mathbb{E} \max_{X \in \mathcal{S}} \Big[\sum_{i=1}^{k} \langle 2v_{b_i} - v, Y_i - \mu \rangle + \sum_{i=1}^{k} \langle v, Y_i - \mu \rangle \Big] \\
&\leqslant \frac{1}{2r} \left(\mathbb{E} \max_{X \in \mathcal{S}} \sum_{i=1}^{k} \langle 2v_{b_i} - v, Y_i - \mu \rangle + \mathbb{E} \max_{X \in \mathcal{S}} \sum_{i=1}^{k} \langle v, Y_i - \mu \rangle \right).
\end{aligned}$$

Noting that X is PSD and specifically, the 2×2 submatrix indexed by v_i and b_j, we have:

$$X^2_{v_i,b_j} \leqslant X_{v_i,v_i} X_{b_j,b_j} \leqslant X_{v_i,v_i} \implies \|v_{b_j}\|^2 = \sum_{i=1}^{d} X^2_{v_i,b_j} \leqslant \sum_{i=1}^{d} X_{v_i,v_i} = 1.$$

Therefore, we get for the second term in the above equation:

$$\mathbb{E} \max_{X \in \mathcal{S}} \sum_{i=1}^{k} \langle v, Y_i - \mu \rangle \leqslant \mathbb{E} \left\| \sum_{i=1}^{k} Y_i - \mu \right\| \leqslant \left(\mathbb{E} \left\| \sum_{i=1}^{k} Y_i - \mu \right\|^2 \right)^{1/2} = (k \operatorname{Tr} \Lambda)^{1/2}.$$

We bound the first term using the following series of inequalities where W is constructed from X as described above:

$$\begin{aligned}\mathbb{E}\max_{X\in\mathcal{S}}\sum_{i=1}^{k}\langle 2v_{b_i}-v, Y_i-\mu\rangle &= \mathbb{E}\max_{X\in\mathcal{S}}\sum_{i=1}^{k}\sum_{j=1}^{d}(Y_i-\mu)_j W_{\sigma_i,v_j}\\ &= \mathbb{E}\max_{X\in\mathcal{S}}\sum_{i=1}^{k}\sum_{j=1}^{d}(\boldsymbol{Y}_{i,j}-\mu_j)W_{\sigma_i,v_j} \leqslant 2\mathbb{E}\|\boldsymbol{Y}-\mathbf{1}\mu^\top\|_{2\to 1},\end{aligned}$$

where the inequality follows from Theorem 2.5.4. With Lemma 2.5.5, the previous two bounds conclude the proof of the lemma. □

We are now able to prove Lemma 2.5.1.

Proof of Lemma 2.5.1. From Lemma 2.5.6 and the fact that:

$$\max_{\|v\|=1}\mathbb{E}\left[|\langle v, Y\rangle|\right] \leqslant \max_{\|v\|=1}\sqrt{\mathbb{E}\left[\langle v, Y\rangle^2\right]} \leqslant \sqrt{\|\Lambda\|}$$

for a mean-zero random vector Y with covariance Λ, we get:

$$\mathbb{E}\max_{X\in\mathcal{S}}\sum_{i=1}^{k}X_{b_i,b_i} \leqslant \frac{k}{40}.$$

Now from Lemma 2.5.2 and an application of the bounded difference inequality (Theorem A.1.2), with probability at least $1-\delta$:

$$\max_{X\in\mathcal{S}}\sum_{i=1}^{k}X_{b_i,b_i} \leqslant \frac{k}{20}$$

concluding the proof of the lemma. □

Chapter 3

Statistical Frontiers

In the previous chapter, we considered the problem of heavy-tailed mean estimation in the setting of bounded variance. We described a simple general algorithmic framework and instantiated it for mean estimation to construct an efficient algorithm which obtains the optimal *sub-Gaussian* rate. That is, the rate that one would have obtained if one had access to *Gaussian* data. Strikingly, no penalty is paid for the lack of more stringent requirements on the distribution. For example, there are no restrictions on the higher-order moments of the distribution which allow for strong concentration properties for simple estimators like the empirical mean which was saw is substantially sub-optimal both in terms of its dependence on the failure probability, δ, and in its multiplicative interaction with the dimension.

In this chapter, we will investigate the impact of noise on the best achievable *statistical* performance of an estimator in settings where even the *variance* of the distribution doesn't exist. These scenarios are ubiquitous in important application domains such as quantitative finance and operations research. Here, as before, the empirical mean is brittle to noise. However, its vulnerability is further exacerbated in these heavier-tailed settings. While the empirical mean is far from optimal, we will see that the best achievable rate for *any* estimator degrades sharply in this setting with the sub-Gaussian rate no longer possible. We will characterize the effect of this noise by establishing *statistical lower bounds* and design an efficient algorithm whose performance matches the lower bound. In fact, we will largely rely on the algorithmic framework developed in Chapter 2.

Formally, we will assume that P satisfies for some *known* $\alpha \in [0, 1]$:

$$\forall \|v\| = 1 : \mathop{\mathbb{E}}_{X \sim P} \left[|\langle v, X - \mu \rangle|^{1+\alpha} \right] \leqslant 1. \tag{MC}$$

Note that when $\alpha = 0$, this captures distributions for which the *largest* moment that exists is the population *mean* while $\alpha = 1$ corresponds to the finite *variance* setting. For intermediate values of α, this condition allows for smooth interpolation between these two extremes. Our main algorithmic result is an estimator whose guarantees are detailed in the following theorem.

Theorem 3.0.1. *There exist absolute constants C, c such that the following holds. Let $\boldsymbol{X} = X_1, \ldots, X_n$ be iid random vectors with mean μ, satisfying the weak moment assumption MC for some known $\alpha \in [0, 1]$. There is a polynomial-time algorithm which, when given inputs $\boldsymbol{X}$ and a target confidence δ with $\log(1/\delta) < cn$, returns a point x^* satisfying:*

$$\|x^* - \mu\| \leqslant C\left(\sqrt{\frac{d}{n}} + \left(\frac{d}{n}\right)^{\frac{\alpha}{1+\alpha}} + \left(\frac{\log(1/\delta)}{n}\right)^{\frac{\alpha}{1+\alpha}}\right)$$

with probability at least $1 - \delta$.

Complementary to the upper bound, we present the following matching *lower* bound which shows that the performance of the estimator is *optimal*.

Theorem 3.0.2. *There exist an absolute constant C such that the following holds. Let $n, d > C$ and $\delta \in \left(e^{-\frac{n}{4}}, \frac{1}{4}\right)$. Then, there exists a set of distributions over $\mathbb{R}^d$, $\mathcal{F}$ such that each $D \in \mathcal{F}$ satisfies MC and the following holds for any estimator $\widehat{\mu}$:*

$$\mathbb{P}_{D\in\mathcal{F}}\left\{\|\widehat{\mu}(\boldsymbol{X}) - \mu(D)\| \geqslant \frac{1}{24}\cdot\max\left(\left(\frac{d}{n}\right)^{\frac{\alpha}{1+\alpha}}, \sqrt{\frac{d}{n}}, \left(\frac{\log(2/\delta)}{n}\right)^{\frac{\alpha}{1+\alpha}}\right)\right\} \geqslant \delta,$$

where $\boldsymbol{X} = X_1, \ldots, X_n$ are generated iid from D and $\mu(D)$ denotes the mean of D.

Together, Theorems 3.0.1 and 3.0.2 have the following implications:

- In the setting where δ is a *constant*, our upper and lower bounds simplify to $O(\sqrt{d/n} + (d/n)^{\alpha/(1+\alpha)})$. Interestingly, Theorem 3.0.1 and Theorem 3.0.2 reveal the existence of a phase transition in the estimation rate when $n \asymp d$—the estimation rate is dominated by $\sqrt{d/n}$ when $n \lesssim d$ and $(d/n)^{\alpha/(1+\alpha)}$ when $n \gtrsim d$ where performance is degraded by the weak moment assumption.

- While it is established in [17] that it is impossible to obtain subgaussian rates in this setting even in one dimension, our results reveal a decoupling between the terms depending on the failure probability and the dimension that parallels the behavior observed in the finite-variance setting (where $\alpha = 1$).

- Finally, our results also extend to the more general problem of mean estimation under adversarial corruption. We recover the mean up to an error of $O(\eta^{\alpha/(1+\alpha)})$ which is information-theoretically optimal (Theorem A.3.1). Furthermore, our sample complexity of (d/η) from Theorem 3.0.1 is optimal as a consequence of Theorem 3.0.2.

Theorem 3.0.1 is established with a simple two-stage estimation procedure. In the first step, $\boldsymbol{X}$ is truncated to discard samples that are too far from the true mean by using a coarse initial estimate as a proxy. The second step utilizes the remaining samples in the testing-to-estimation framework of Chapter 2 to construct an efficient descent based algorithm.

The main technical challenge is in verifying the assumptions needed by the gradient descent procedure (Assumption 2.4.2 and Theorem 2.4.4) in this heavier tailed scenario. Concretely, the analysis in Chapter 2 makes critical use of the decomposition of the variance of sums of independent random variables which does not hold here. This allows tight control of the *second* moments of $\sum_{i=1}^{m} X_i$ and $\|X - \mu\|$, crucial to the previous analysis. Despite the lack of such decompositions for weak moments, we establish tight control over the appropriate quantities allowing us to establish our optimal recovery guarantees.

Similarly, the presence of weak moments also complicates the task of establishing a matching lower bound with tight dependence on the dimension d. The main difficulty is in proving the optimality of the dimension-dependent term, $(d/n)^{\alpha/(1+\alpha)}$. For the specific case where $\alpha = 1$, the lower bound may be proved within the estimation-to-testing framework for proving minimax rates (see, for example, [62, Chapter 15]) by utilizing a distribution over a collection of isotropic Gaussian distributions with well-separated means. However, this approach fails for the weak-moment mean estimation problem; indeed, hypercontractivity properties of Gaussian distributions ensure a bounded variance leading to a lower bound that scales as $1/\sqrt{n}$ as opposed to the slower rate $n^{-\alpha/(1+\alpha)}$. We, instead, use a collection of carefully chosen distributions with discrete supports whose means are separated by $O((d/n)^{\alpha/(1+\alpha)})$. Further challenges arise at this point—if we follow the standard path of bounding the complexity of the testing problem in terms of pairwise f-divergences between distributions in the hypothesis set, we obtain vacuous bounds. We instead directly analyze the posterior distribution obtained from the framework and show that random independent samples from the posterior tend to be well separated, yielding our tight lower bound.

The rest of the chapter is organized as follows. We describe our estimator which is essentially the one discussed in Chapter 2 with minor modifications and prove Theorem 3.0.1 in Section 3.1. The main technical contribution, here, is showing that the appropriate *statistical* concentration results still hold even in this *weak moment* setting. We then present our statistical lower bounds, Theorem 3.0.2, proving the optimality of Theorem 3.0.1 in Section 3.2.

3.1 An Efficient Estimator

In this section, we prove Theorem 3.0.1 by verifying the conditions required for the success of the gradient-descent approach from Chapter 2 (Assumption 2.4.2 and Theorem 2.4.4). As alluded to previously, this is made technically challenging due to the lack of decomposition properties enjoyed by the variance. The weaker moment conditions also require modifications to the algorithm itself which we describe subsequently.

Algorithm

Our estimator is defined in Algorithms 8 to 10. Note that, in addition to the bucketing and gradient descent steps, we have an additional pre-processing step which prunes data points provably far from the true mean (Algorithm 10) before the bucketing step. This is required to control the *variance* of the bucketed means which allows establishing Assumption 2.4.2 with the right parameters. At the same time, the truncation must not be too aggressive to significantly distort the mean of the data points used to construct the bucketed means.

Algorithm 8 Mean Estimation

Input: Data Points $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, Target Confidence δ
$x^\dagger \leftarrow$ Initial Mean Estimate($\{X_1, \dots, X_{n/2}\}$) (Algorithm 5)
$\boldsymbol{Z} \leftarrow$ Produce Bucket Estimates($\left\{X_{n/2+1}, \dots, X_n\right\}, x^\dagger, \delta$)
$T \leftarrow C \log(n)$
$x^* =$ Gradient Descent($\boldsymbol{Z}, x^\dagger, T$) (Algorithm 2)
Return: x^*

Algorithm 9 Produce Bucket Estimates

Input: Data Points $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, Mean Estimate $x^\dagger$, Target Confidence δ
$\boldsymbol{Y} \leftarrow$ Prune Data($\boldsymbol{X}, x^\dagger$)
$m \leftarrow |\boldsymbol{Y}|$
$k \leftarrow C \log(1/\delta)$
Split data points into k buckets with bucket $\mathcal{B}_i = \{Y_{(i-1)\frac{m}{k}+1}, \dots, Y_{i\frac{m}{k}}\}$
$Z_i \leftarrow$ Mean($\mathcal{B}_i$) $\forall\, i \in [k]$ and $\boldsymbol{Z} \leftarrow (Z_1, \dots, Z_k)$
Return: $\boldsymbol{Z}$

Algorithm 10 Prune Data

Input: Set of data points $\boldsymbol{X} = \{X_i\}_{i=1}^n$, Mean Estimate $x^\dagger$
$\tau \leftarrow C \max\left(n^{\frac{1}{1+\alpha}} d^{-\frac{(1-\alpha)}{2(1+\alpha)}}, \sqrt{d}\right)$
$\mathcal{C} \leftarrow \{X_i : \|X_i - x^\dagger\| \leqslant \tau\}$
Return: $\mathcal{C}$

Analysis

Here, we formally establish Theorem 3.0.1. We will do so by verifying the conditions of Theorem 2.4.4 (Assumption 2.4.2) with the correct parameters for the dataset returned by Algorithm 9. Throughout, we will assume that the estimate $x^\dagger$ used in Algorithm 8 satisfies $\|x^\dagger - \mu\| \leqslant 60\sqrt{d}$ from Lemma A.2.1. We will analyze the algorithm in two steps:

1. First, we analyze the truncation step (Algorithm 10) and establish bounds on the variances of the points returned and the distortion of the means incurred by the truncation.

2. Secondly, we analyze the bucketing step (Algorithm 9) where we bound the values of the mean testing problem **MT** similarly to Lemma 2.5.1. From Lemmas 2.5.5 and 2.5.6, this requires control of the (trace of the) variance of the points returned by the truncation step and also the *directional* moments of the bucketed means.

We now analyze the truncation step.

Analyzing Algorithm 10

We will need the following key lemma which bounds the $(1+\alpha)^{th}$ moment of the length of a random vector satisfying MC.

Lemma 3.1.1. *Let X be a zero-mean random vector satisfying MC for $\alpha \in [0,1]$. We have:*

$$\mathbb{E}[\|X\|^{1+\alpha}] \leqslant \frac{\pi}{2} \cdot d^{\frac{1+\alpha}{2}}.$$

Proof. The argument hinges on a Gaussian projection trick which introduces $g \sim \mathcal{N}(0, I)$ to rewrite the norm. From the concavity of $f(x) = |x|^{(1+\alpha)/2}$ when $x \geqslant 0$, we have:

$$\begin{aligned}\mathbb{E}[\|X\|^{1+\alpha}] &= \mathbb{E}_X\left[\left(\sqrt{\frac{\pi}{2}}\mathbb{E}_g|\langle X, g\rangle|\right)^{1+\alpha}\right] \leqslant \frac{\pi}{2}\mathbb{E}_X\mathbb{E}_g\left[|\langle X, g\rangle|^{1+\alpha}\right] \\ &= \frac{\pi}{2}\mathbb{E}_g\|g\|^{1+\alpha}\mathbb{E}_X\left[\left|\left\langle X, \frac{g}{\|g\|}\right\rangle\right|^{1+\alpha}\right] \leqslant \frac{\pi}{2}\mathbb{E}_g[\|g\|^{1+\alpha}] \leqslant \frac{\pi}{2} \cdot d^{\frac{1+\alpha}{2}}.\end{aligned}$$

□

Our next lemma bounds the deviation in the means and the blow up in the weak moments when the distribution is truncated to a general set (and not just an Euclidean ball as in Algorithm 10). Here, we cannot establish variance control as the set could potentially be unbounded. We will bound the variance in a later result.

Lemma 3.1.2. *Let ν be a mean-zero distribution over $\mathbb{R}^d$ satisfying MC for $\alpha \in [0,1]$. Furthermore, let $A \subset \mathbb{R}^d$ be such that $\nu(A) = \delta \leqslant \frac{1}{2}$. Let $\nu_S()$ be the conditional distribution of ν conditioned on the set S. Then we have for $Y \sim \nu(A^c)$:*

$$\textit{Claim 1: } \|\mu(\nu_{A^c})\| \leqslant 2\delta^{\frac{\alpha}{1+\alpha}}, \qquad \textit{Claim 2: } \forall \|v\| = 1, \quad \mathbb{E}\left[|\langle v, Y - \mu(\nu_{A^c})\rangle|^{1+\alpha}\right] \leqslant 20.$$

Proof. Letting $p_A = \mathbb{P}\{X \in A\}$, we have $\nu = p_A\nu_A + p_{A^c}\nu_{A^c}$. Then,

$$\|\mu(\nu_{A^c})\| = \max_{\|v\|=1}\langle v, \mu(\nu_{A^c})\rangle.$$

So for any $\|v\| = 1$:

$$\langle v, \mu(\nu_{A^c})\rangle = \langle v, \mu(\nu_{A^c}) - p_A\mu(\nu_A) - p_{A^c}\mu(\nu_{A^c})\rangle$$

$$= \langle v, p_A \mu(\nu_{A^c}) - p_A \mu(\nu_A) \rangle = p_A \langle v, \mu(\nu_{A^c}) - \mu(\nu_A) \rangle.$$

Since $\mu(\nu) = 0$, we have $p_A \mu(\nu_A) = -p_{A^c} \mu(\nu_{A^c})$. We now get:

$$p_A \langle v, \mu(\nu_A) - \mu(\nu_{A^c}) \rangle = p_A \left\langle v, \mu(\nu_A) + \frac{p_A}{p_{A^c}} \mu(\nu_A) \right\rangle = \left(1 + \frac{p_A}{p_{A^c}} \right) \langle v, p_A \mu(\nu_A) \rangle.$$

Finally,

$$\begin{aligned} \langle v, p_A \mu(\nu_A) \rangle &= \mathbb{E}_{X \sim \mu} \left[\mathbf{1} \left\{ X \in A \right\} \langle X, v \rangle \right] \\ &\leqslant \left(\mathbb{E} \left[\left(\mathbf{1} \left\{ X \in A \right\} \right)^{\frac{1+\alpha}{\alpha}} \right] \right)^{\frac{\alpha}{1+\alpha}} \cdot \left(\mathbb{E} \left[\left| \langle X, v \rangle \right|^{1+\alpha} \right] \right)^{\frac{1}{1+\alpha}} = p_A^{\frac{\alpha}{1+\alpha}} \end{aligned}$$

where the inequality follows by Hölder's inequality. We get the first claim as:

$$\max_{\|v\|=1} \langle v, \mu(\nu_{A^c}) \rangle = \max_{\|v\|=1} \left(1 + \frac{p_A}{p_{A^c}} \right) \langle v, p_A \mu(\nu_{A^c}) \rangle \leqslant \max_{\|v\|=1} \left(1 + \frac{p_A}{p_{A^c}} \right) p_A^{\frac{\alpha}{1+\alpha}} \leqslant 2\delta^{\frac{\alpha}{1+\alpha}},$$

where the final inequality follows from the fact that $p_{A^c} \geqslant p_A$.

For the second claim, let $Y \sim \nu_{A^c}$ and $\mu_Y = \mathbb{E}[Y]$. We decompose the term as:

$$\mathbb{E} \left[\left| \langle Y - \mu_Y, v \rangle \right|^{1+\alpha} \right] \leqslant 2^{1+\alpha} \cdot \mathbb{E} \left[\left| \langle \mu_Y, v \rangle \right|^{1+\alpha} + \left| \langle Y, v \rangle \right|^{1+\alpha} \right].$$

For the second term, we have with $Z \sim \nu_A$:

$$\mathbb{E} \left[\left| \langle Y, v \rangle \right|^{1+\alpha} \right] = p_{A^c}^{-1} \left(\mathbb{E} \left[\left| \langle X, v \rangle \right|^{1+\alpha} \right] - p_A \mathbb{E} \left[\left| \langle Z, v \rangle \right|^{1+\alpha} \right] \right) \leqslant 2.$$

Therefore, we finally have:

$$\mathbb{E} \left[\left| \langle Y - \mu_Y, v \rangle \right|^{1+\alpha} \right] \leqslant 8 + 2^{1+\alpha} \cdot 2^{1+\alpha} \cdot \delta^\alpha \leqslant 16,$$

which proves the second claim. ∎

Next, a simple lemma used to bound the *variance* of points returned by Algorithm 10.

Lemma 3.1.3. *Let $X \sim \nu$ be a mean-zero random vector satisfying the weak-moment condition for some $0 \leqslant \alpha \leqslant 1$. Then, we have for any $\tau > 0$:*

$$\mathbb{E} \left[\|X\|^2 \cdot \mathbf{1} \left\{ \|X\| \leqslant \tau \right\} \right] \leqslant \frac{\pi}{2} d^{\frac{1+\alpha}{2}} \tau^{1-\alpha}.$$

Proof. The proof of the lemma proceeds as follows:

$$\mathbb{E} \left[\|X\|^2 \cdot \mathbf{1} \left\{ \|X\| \leqslant \tau \right\} \right] \leqslant \tau^{1-\alpha} \mathbb{E} \left[\|X\|^{1+\alpha} \mathbf{1} \left\{ \|X\| \leqslant \tau \right\} \right] \leqslant \tau^{1-\alpha} \mathbb{E} \left[\|X\|^{1+\alpha} \right] \leqslant \frac{\pi}{2} d^{\frac{1+\alpha}{2}} \tau^{1-\alpha},$$

where the last inequality follows from Lemma 3.1.1. ∎

Finally, we analyze Algorithm 10 in the following lemma.

Lemma 3.1.4. *There exist absolute constants $C_1, C_2, c > 0$ such that the following holds. Let $\boldsymbol{X} = \{X_i\}_{i=1}^n$ be iid zero-mean random vectors distributed according to ν satisfying MC for $\alpha \in [0,1]$. Furthermore, let $x^\dagger$ satisfy $\|x^\dagger\| \leqslant 60\sqrt{d}$. Then, the output $\boldsymbol{Y}$ of Algorithm 10 with input $\boldsymbol{X}$ and $x^\dagger$ are iid with mean $\widetilde{\mu}$ and covariance $\widetilde{\Sigma}$. Furthermore, they satisfy:*

$$\textit{Claim 1: } \mathbb{P}\left\{|\boldsymbol{Y}| \geqslant \frac{3n}{4}\right\} \geqslant 1 - e^{-cn}, \qquad \textit{Claim 2: } \|\tilde{\mu}\| \leqslant 2\left(\frac{d}{n}\right)^{\frac{\alpha}{1+\alpha}}$$

$$\textit{Claim 3: } \forall \|v\| = 1 : \mathbb{E}\left[|\langle Y_i - \tilde{\mu}, v\rangle|^{1+\alpha}\right] \leqslant C_1, \qquad \textit{Claim 4: } \operatorname{Tr}\tilde{\Sigma} \leqslant C_2 \max\left(n^{\frac{1-\alpha}{1+\alpha}} d^{\frac{2\alpha}{(1+\alpha)}}, d\right).$$

Proof. First, consider the set $A = \{x : \|x - x^\dagger\| \leqslant \tau\}$ as defined in Algorithm 10. Note that $\{x : \|x\| \leqslant 0.75\tau\} \subseteq A$. We have by Markov's inequality and Lemma 3.1.1:

$$\mathbb{P}\{X_i \in A\} \geqslant 1 - \min\left(\frac{d}{n}, \frac{1}{25}\right).$$

By Hoeffding's inequality (Theorem A.1.1), the definition of Y_i, we have with probability at least $1 - e^{-cn}$:

$$|\boldsymbol{Y}| \geqslant \frac{3n}{4},$$

proving the first claim of the lemma. For the next two claims, note that each of the Y_i are iid according to ν_A. Again, we get from the bound on $\mathbb{P}\{X_i \in A\}$ by an application of Lemma 3.1.2, the next two claims of the lemma:

$$\text{Claim 2: } \|\tilde{\mu}\| \leqslant 2\left(\frac{d}{n}\right)^{\frac{\alpha}{1+\alpha}}, \qquad \text{Claim 3: } \forall \|v\| = 1 : \mathbb{E}\left[|\langle Y_i - \tilde{\mu}, v\rangle|^{1+\alpha}\right] \leqslant 20.$$

For the final claim, note that as $\|x^\dagger\| \leqslant 60\sqrt{d}$, we have $A \subseteq B := \{x : \|x\| \leqslant 1.25\tau\}$. Therefore, we have by the property of the mean that:

$$\begin{aligned}\operatorname{Tr}\tilde{\Sigma} = \mathbb{E}\left[\|Y_i - \tilde{\mu}\|^2\right] \leqslant \mathbb{E}\left[\|Y_i\|^2\right] &= \frac{1}{\nu(A)}\mathbb{E}\left[\|X_j\|^2 \mathbf{1}\{X_j \in A\}\right] \\ &\leqslant 2\mathbb{E}\left[\|X_j\|^2 \mathbf{1}\{X_j \in B\}\right] \leqslant C \max\left(n^{\frac{1-\alpha}{1+\alpha}} d^{\frac{2\alpha}{(1+\alpha)}}, d\right),\end{aligned}$$

where the final inequality follows from Lemma 3.1.3 and the definition of τ. □

Now, we move onto the bucketing step (Algorithm 9).

Analyzing Algorithm 9

Here, we require the following key technical result bounding the weak moment of sums of independent random variables. Note that weak moments do not satisfy the variance decomposition property where the variance of a sum of independent random variables is the sum of their variances. However, the next lemma shows that an *approximate* version of this property continues to hold.

Lemma 3.1.5. *Let $X_1, \dots, X_n$ be n mean-zero i.i.d. random variables satisfying for some $\alpha \in [0,1]$:*

$$\mathbb{E}[|X_i|^{1+\alpha}] \leqslant 1. \tag{3.1}$$

Then, we have:

$$\mathbb{E}\left[\left|\sum_{i=1}^n X_i\right|^{1+\alpha}\right] \leqslant 2n. \tag{3.2}$$

Proof. We first need the following claim.

Claim 3.1.6. Let $g(x) = \operatorname{sgn}(x)|x|^\alpha$ for some $0 < \alpha \leqslant 1$. Then we have for any $h \geqslant 0$:

$$\max_x g(x+h) - g(x) = 2g\left(\frac{h}{2}\right).$$

Proof. Consider the function $l(x) = g(x+h) - g(x)$. We see that l is differentiable everywhere except at $x = 0$ and $x = -h$. As long as $x \neq 0, -h$, we have:

$$l'(x) = g'(x+h) - g'(x) = \alpha(|x+h|^{\alpha-1} - |x|^{\alpha-1})$$

Since, we have $\alpha \leqslant 1$, $x = -\frac{h}{2}$ is a local maxima for $l(x)$. Furthermore, note that $l'(x) \geqslant 0$ for $x \in (-\infty, -\frac{h}{2}) \setminus \{-h\}$ and $l'(x) \leqslant 0$ for $x \in (-\frac{h}{2}, \infty) \setminus \{0\}$. Therefore, we get from the continuity of l that $x = -\frac{h}{2}$ is a global maxima for $l(x)$. Substituting yields the claim. □

The case where $\alpha = 0$ is trivial. When $\alpha > 0$, we start by defining:

$$S_i = \sum_{j=1}^i X_j, \qquad S_0 = 0, \qquad f(x) = |x|^{1+\alpha}, \qquad f'(x) = (1+\alpha)\operatorname{sgn}(x)|x|^\alpha.$$

Therefore, we have from an application of Claim 3.1.6:

$$\begin{aligned}
\mathbb{E}[f(S_n)] &= \mathbb{E}\left[\sum_{i=1}^n f(S_i) - f(S_{i-1})\right] = \sum_{i=1}^n \mathbb{E}[f(S_i) - f(S_{i-1})] \\
&= \sum_{i=1}^n \mathbb{E}\left[\int_{S_{i-1}}^{S_i} f'(x)dx\right] = \sum_{i=1}^n \mathbb{E}\left[X_i f'(S_{i-1}) + \int_{S_{i-1}}^{S_i} f'(x) - f'(S_{i-1})dx\right] \\
&= \sum_{i=1}^n \mathbb{E}\left[\int_{S_{i-1}}^{S_i} f'(x) - f'(S_{i-1})dx\right] \leqslant 2\sum_{i=1}^n \mathbb{E}\left[\int_0^{|X_i|} f'\left(\frac{t}{2}\right) dt\right] \\
&= 2\sum_{i=1}^n \mathbb{E}\left[\int_0^{|X_i|/2} 2f'(s)\, ds\right] = 4\sum_{i=1}^n \mathbb{E}\left[f\left(\frac{|X_i|}{2}\right)\right] \leqslant 2n.
\end{aligned}$$

□

We are now finally, ready to analyze Algorithm 9 in the following lemma. The main result of this section is the following high probability guarantee on the set of points output by Algorithm 9.

Lemma 3.1.7. *There exist absolute constants $c, C > 0$ such that the following hold. Let $\boldsymbol{X} = \{X_i\}_{i=1}^n$ be iid random vectors with mean μ, satisfying MC for $\alpha \in [0,1]$ and $\delta \in (0,1)$ be such that $\log(1/\delta) < cn$. Furthermore, suppose that $x^\dagger$ satisfies $\|x^\dagger - \mu\| \leqslant 60\sqrt{d}$. Let $\boldsymbol{Z} = \{Z_i\}_{i=1}^k$ denote the set of vectors output by Algorithm 9 with inputs $\boldsymbol{X}$, $x^\dagger$ and δ. Then, there exists a point $\widetilde{\mu}$ such that for all r satisfying:*

$$r \geqslant C \left(\sqrt{\frac{d}{n}} + \left(\frac{d}{n}\right)^{\frac{\alpha}{1+\alpha}} + \left(\frac{\log(1/\delta)}{n}\right)^{\frac{\alpha}{1+\alpha}} \right),$$

we have

$$\|\widetilde{\mu} - \mu\| \leqslant 2 \left(\frac{d}{n}\right)^{\frac{\alpha}{1+\alpha}} \text{ and } \max_{X \in \mathcal{S}} \sum_{i=1}^k X_{b_i, b_i} \leqslant \frac{k}{20},$$

with probability at least $1 - \delta/2$ where $\mathcal{S}$ denotes the set of feasible solutions of $\boldsymbol{MT}(\widetilde{\mu}, r, \boldsymbol{Z})$.

Proof. Note that it is sufficient to prove the lemma for $\mu = \mathbf{0}$. We may now assume each of the Y_i are iid random variables satisfying the conclusions of Lemma 3.1.4. Therefore, Z_i are iid random vectors with mean $\widetilde{\mu}$ and covariance $\widetilde{\Sigma}$ satisfying:

$$\|\widetilde{\mu}\| \leqslant 2 \left(\frac{d}{n}\right)^{\frac{\alpha}{1+\alpha}} \qquad \operatorname{Tr} \tilde{\Sigma} \leqslant C' \left(k \max \left\{ \frac{d}{n}, \left(\frac{d}{n}\right)^{\frac{2\alpha}{1+\alpha}} \right\} \right).$$

Furthermore, we have by an application of Lemma 3.1.5 that:

$$\forall \|v\| = 1 : \mathbb{E}\left[|\langle v, Z_i - \widetilde{\mu} \rangle|^{\frac{\alpha}{1+\alpha}} \right] \leqslant C^\dagger \left(\frac{k}{n}\right)^{\alpha}.$$

From Lemma 2.5.6 and the fact that:

$$\max_{\|v\|=1} \mathbb{E}\left[|\langle v, Z_i - \widetilde{\mu} \rangle| \right] \leqslant \max_{\|v\|=1} \left(\mathbb{E}\left[\langle v, Z_i - \widetilde{\mu} \rangle^{1+\alpha} \right] \right)^{1+\alpha} \leqslant C^\dagger \left(\frac{k}{n}\right)^{\frac{\alpha}{1+\alpha}},$$

we get:

$$\mathbb{E} \max_{X \in \mathcal{S}} \sum_{i=1}^k X_{b_i, b_i} \leqslant \frac{k}{40}.$$

Now from Lemma 2.5.2 and an application of the bounded difference inequality (Theorem A.1.2), with probability at least $1 - \delta$:

$$\max_{X \in \mathcal{S}} \sum_{i=1}^k X_{b_i, b_i} \leqslant \frac{k}{20}$$

concluding the proof of the lemma. □

Wrapping up - Proof of Theorem 3.0.1

To conclude the proof of Theorem 3.0.1, we union bound over the events in Lemmas 3.1.7 and A.2.1. The theorem now follows from Theorem 2.4.4 as Assumption 2.4.2 is satisfied for $\widetilde{\mu}$ and r from Lemma 3.1.7 and the bound on $\|\widetilde{\mu} - \mu\|$. □

3.2 A Matching Lower Bound

We will now show that the performance guarantees of Theorem 3.0.1 are *tight*. As discussed, the proof of our lower bound bypasses standard information theoretic techniques such as Fano's inequality and we instead perform an explicit analysis of the posterior distribution in the classic Bayesian estimation-to-testing framework for proving minimax lower bounds. We will first prove the bound in the easier bounded *covariance* setting ($\alpha = 1$) where *Gaussians* witness the lower bound before considering the general setting. We require both bounds as for the bounded covariance setting, the lower bound holds for any n, d while for the general setting, they are specific to $n \gtrsim d$.

The Bounded Covariance Setting

Here, MC relaxes to the following:

$$\mathop{\mathbb{E}}_{X \sim P} \left[(X - \mu)(X - \mu)^\top \right] \preccurlyeq I.$$

And we will now consider datasets generated according to the following process:

1. First, draw $\mu \sim \mathcal{N}(0, I)$.
2. Then, draw $\boldsymbol{X} = X_1, \dots, X_n$ iid from $\mathcal{N}(\mu, I)$.

Lemma 3.2.1. *Let $n, d > 50$ and μ and $\boldsymbol{X}$ be generated according to the above process. Then, we have for any estimator, $\widehat{\mu}(\cdot)$:*

$$\mathbf{Pr}_{\mu, \boldsymbol{X}} \left\{ \|\widehat{\mu}(\boldsymbol{X}) - \mu\| \geqslant \frac{1}{2} \sqrt{\frac{d}{n}} \right\} \geqslant \frac{1}{2}.$$

Proof. We first consider the posterior density of μ given $\boldsymbol{X}$. First define:

$$\overline{X} = n^{-1} \sum_{i=1}^{n} X_i \qquad \widetilde{X} = \frac{n}{n+1} \overline{X}.$$

We now have the posterior density of μ, $f(\cdot \mid \boldsymbol{X})$:

$$f(\mu \mid \boldsymbol{X}) \propto \exp\left\{ -\frac{\|\mu\|^2}{2} \right\} \exp\left\{ -\sum_{i=1}^{n} \frac{\|X_i - \mu\|^2}{2} \right\}$$

$$\propto \exp\left\{-\frac{(n+1)\|\mu - \widetilde{X}\|^2}{2}\right\}.$$

Therefore, the posterior distribution of μ is $\mathcal{N}(\widetilde{X}, I/(n+1))$; i.e. a Gaussian distribution with mean $\widetilde{X}$ and variance $I/(n+1)$. Now, for any estimator $\widehat{\mu}$, we get for any $t > 0$:

$$\mathbf{Pr}\left\{\|\widehat{\mu}(\boldsymbol{X}) - \mu\| \geqslant t \mid \boldsymbol{X}\right\} \geqslant \mathbf{Pr}\left\{\|\mathcal{P}_{\Delta}^{\perp}(\widehat{\mu}(\boldsymbol{X}) - \mu)\| \geqslant t \mid \boldsymbol{X}\right\}$$

where Δ is the unit vector along $\widehat{\mu}(\boldsymbol{X}) - \widetilde{X}$. And noting that $\|\mathcal{P}_{\Delta}^{\perp}(\widehat{\mu}(\boldsymbol{X}) - \mu)\|$ is distributed according to $\|g\|$ where g is a $(d-1)$-dimensional Gaussian random vector with mean 0 and variance $I/(n+1)$. And hence, we get from the above inequality, our bounds on n, d and the concentration of lengths of Gaussian random vectors (Lemma A.1.5):

$$\mathbf{Pr}\left\{\|\widehat{\mu}(\boldsymbol{X}) - \mu\| \geqslant \frac{1}{2}\sqrt{\frac{d}{n}}\,\middle|\, \boldsymbol{X}\right\} \geqslant \mathbf{Pr}\left\{\|g\| \geqslant \frac{1}{2}\sqrt{\frac{d}{n}}\,\middle|\, \boldsymbol{X}\right\} \geqslant \frac{1}{2}.$$

Averaging the above equation with respect to $\boldsymbol{X}$, we get:

$$\mathbf{Pr}\left\{\|\widehat{\mu}(\boldsymbol{X}) - \mu\| \geqslant \frac{1}{2}\sqrt{\frac{d}{n}}\right\} \geqslant \frac{1}{2}$$

concluding the proof. □

The General Setting

Here, we prove a lower bound for the general $\alpha \in [0, 1]$ setting. As opposed to the bounded covariance setting where Gaussians were used in the lower bound constructions, both the class of distributions and the analysis of the posterior is made more complex.

For a given dimension d, and sample size $n \geqslant 8d$, we will consider a family of distributions parameterized by size $d/2$ subsets of $[d]$. That is, we will consider a family of distributions $\mathcal{F} = \{D_S : S \subset [d] \text{ and } |S| = d/2\}$. Now, for each particular distribution D_S, we have $X \sim D_S$ as follows:

$$X = \begin{cases} 0, & \text{with probability } 1 - \frac{d}{8n} \\ n^{\frac{1}{1+\alpha}} \cdot d^{-\frac{(1-\alpha)}{2(1+\alpha)}} \cdot e_i, & \text{for } i \in S \text{ with probability } \frac{1}{4n}. \end{cases}$$

Defining, $\mu_S = \mu(D_S)$, we will first show that each D_S satisfies MC.

Lemma 3.2.2. *Let $X \sim D_S$ for some $S \subset [d]$ such that $|S| = d/2$. Then, X satisfies:*

$$\forall v : \|v\| = 1 : \mathbb{E}\left[|\langle v, X - \mu_S\rangle|^{1+\alpha}\right] \leqslant 1.$$

Proof. We first note that:

$$(\mu_S)_i = \begin{cases} 0, & \text{for } i \notin S \\ \frac{n^{-\frac{\alpha}{1+\alpha}} \cdot d^{-\frac{(1-\alpha)}{2(1+\alpha)}}}{4}, & \text{otherwise.} \end{cases}$$

Let v satisfy $\|v\| = 1$. We have by the convexity of $f(x) = |x|^{1+\alpha}$ and Jensen's inequality:

$$\mathbb{E}\left[|\langle v, X - \mu_S\rangle|^{1+\alpha}\right] \leqslant 2\,\mathbb{E}\left[|\langle v, X\rangle|^{1+\alpha} + |\langle v, \mu_S\rangle|^{1+\alpha}\right] \leqslant 4\,\mathbb{E}\left[|\langle v, X\rangle|^{1+\alpha}\right].$$

We now have by Hölder's inequality:

$$\begin{aligned} \mathbb{E}\left[|\langle v, X\rangle|^{1+\alpha}\right] &= \sum_{i\in S} \frac{1}{4n}|v_i|^{1+\alpha} \cdot nd^{-\frac{1-\alpha}{2}} = \frac{1}{4}\sum_{i\in S}|v_i|^{1+\alpha} d^{-\frac{1-\alpha}{2}} \\ &\leqslant \frac{1}{4}\left(\sum_{i\in S} v_i^2\right)^{\frac{1+\alpha}{2}} \left(\sum_{i\in S} d^{-1}\right)^{\frac{1-\alpha}{2}} \leqslant \frac{1}{4}, \end{aligned}$$

concluding the proof of the lemma. □

We now prove a lemma that establishes the lower bound when $n \gtrsim d$ in the constant probability regime. We use the following generative process for the data $\boldsymbol{X} = X_1, \dots, X_n$:

1. Randomly pick a subset S uniformly from the set $\{T \subset [d] : |T| = d/2\}$.
2. Generate $X_1, \dots, X_n$ iid from the distribution, D_S.

Lemma 3.2.3. *There exist absolute constants $C_1, C_2 > 0$ such that the following holds. Let $d \geqslant C_1$, $n \geqslant C_2 d$ and $(S, \boldsymbol{X})$ be generated according to the above process. We have, for any estimator $\widehat{\mu}$,*

$$\mathbf{Pr}_{S,\boldsymbol{X}}\left\{\|\widehat{\mu}(\boldsymbol{X}) - \mu_S\| \geqslant \frac{1}{24} \cdot \left(\frac{d}{n}\right)^{\frac{\alpha}{1+\alpha}}\right\} \geqslant \frac{1}{4}.$$

Proof. We first define the random variable $Y \coloneqq \sum_{i=1}^n \mathbf{1}\{X_i \neq 0\}$. From the definition of the distributions D_S we have:

$$\mathbb{E}[Y] = \frac{d}{8}$$

Therefore, we have that $Y \leqslant d/4$ with probability at least $1/2$, by Markov's inequality. We now define the following random set: $T \coloneqq \{i \in [d] : \exists j \in [n] \text{ such that } (X_j)_i \neq 0\}$. We see from the definition of T and Y that $|T| \leqslant Y$. We have with probability at least $1/2$ that $|T| \leqslant d/4$. Let $\boldsymbol{X}$ be an outcome for which $|T| = k \leqslant d/4$. We have by the symmetry of the distribution that:

$$\mathbf{Pr}\{S|\boldsymbol{X}\} = \begin{cases} \frac{1}{\binom{d-k}{d/2-k}}, & \text{if } T \subset S \text{ and } |S| = d/2 \\ 0, & \text{otherwise.} \end{cases}$$

For given $\boldsymbol{X}$, define $Z_i = \mathbf{1}\{i \in S\}$ for $i \notin T$ (For $i \in T$, Z_i is 1). We have for Z_i and Z_j for distinct $i, j \notin T$:

$$\mathbb{E}[Z_i|\boldsymbol{X}] = \mathbb{E}[Z_j|\boldsymbol{X}] = \frac{d-2k}{2(d-k)}.$$

Furthermore, we have:

$$\begin{aligned}
\operatorname{Cov}(Z_i, Z_j|\boldsymbol{X}) &= \frac{(d-2k)(d-2k-2)(d-k) - (d-2k)^2(d-k-1)}{4(d-k)^2(d-k-1)} \\
&= \frac{(d-2k)((d-2k)(d-k) - 2(d-k) - (d-2k)(d-k) + (d-2k))}{4(d-k)^2(d-k-1)} \\
&= \frac{-d(d-2k)}{4(d-k)^2(d-k-1)} < 0.
\end{aligned}$$

Now, consider some $R \subset [d]$ such that $|R| = d/2$ and $T \subset R$. Let $Q = R \setminus T$. For Q, we have $|Q| = d/2 - k$. We have for S:

$$|S \cap R| = k + \sum_{i \in Q} Z_i.$$

This means that:

$$\operatorname{Var}(|S \cap R| \mid \boldsymbol{X}) = \operatorname{Var}\left(\sum_{i \in Q} Z_i \mid \boldsymbol{X}\right) \leqslant \sum_{i \in Q}\left(\frac{d-2k}{2(d-k)}\right)^2 \leqslant \frac{|Q|}{4} \leqslant \frac{d}{8}.$$

Furthermore, we have that:

$$\mathbb{E}(|S \cap R| \mid \boldsymbol{X}) = k + \left(\frac{d}{2} - k\right) \cdot \frac{(d-2k)}{2(d-k)} \leqslant \frac{d}{4} + \frac{d}{4} \cdot \frac{d}{4(3d/4)} = \frac{d}{4} + \frac{d}{12} = \frac{d}{3}.$$

Therefore, we have by Chebyshev's inequality that:

$$\mathbf{Pr}\left\{|S \cap R| \geqslant \frac{5d}{12}\right\} \leqslant \frac{1}{2}.$$

Note that for any S_1, S_2 such that $|S_i| = \frac{d}{2}$ and $|S_1 \cap S_2| \leqslant \frac{5d}{12}$, we have:

$$\|\mu_{S_1} - \mu_{S_2}\| \geqslant \sqrt{2 \cdot \frac{d}{12} \cdot \left(\frac{n^{-\frac{\alpha}{1+\alpha}} \cdot d^{-\frac{1-\alpha}{2(1+\alpha)}}}{4}\right)^2} \geqslant \frac{1}{12} \cdot \left(\frac{d}{n}\right)^{\frac{\alpha}{1+\alpha}}.$$

Consider any estimator $\widehat{\mu}$. Suppose there exists R such that $T \subset R$, $|R| = d/2$ and $\|\widehat{\mu}(\boldsymbol{X}) - \mu_R\| \leqslant \frac{1}{24} \cdot \left(\frac{d}{n}\right)^{\frac{\alpha}{1+\alpha}}$. Then, we have by the triangle inequality:

$$\mathbf{Pr}\left\{\|\widehat{\mu}(\boldsymbol{X}) - \mu_S\| \geqslant \frac{1}{24} \cdot \left(\frac{d}{n}\right)^{\frac{\alpha}{1+\alpha}} \middle| \boldsymbol{X}\right\} \geqslant \frac{1}{2}.$$

In the alternate case where $\|\widehat{\mu}(\boldsymbol{X}) - \mu_R\| \geqslant \frac{1}{24} \cdot \left(\frac{d}{n}\right)^{\frac{\alpha}{1+\alpha}}$ for all such R, the same conclusion holds trivially. From these two cases, we obtain:

$$\mathbf{Pr}\left\{ \|\widehat{\mu}(\boldsymbol{X}) - \mu_S\| \geqslant \frac{1}{24} \cdot \left(\frac{d}{n}\right)^{\frac{\alpha}{1+\alpha}} \middle| \boldsymbol{X} \right\} \geqslant \frac{1}{2}.$$

Since such an $\boldsymbol{X}$ occurs with probability at least $1/2$, we arrive at our result:

$$\mathbf{Pr}\left\{ \|\widehat{\mu}(\boldsymbol{X}) - \mu_S\| \geqslant \frac{1}{24} \cdot \left(\frac{d}{n}\right)^{\frac{\alpha}{1+\alpha}} \right\} \geqslant \frac{1}{4}.$$

□

As part of our proof, we use the following one-dimensional lower bound from [17].

Theorem 3.2.4. *For any n, $\delta \in \left(2^{-\frac{n}{4}}, \frac{1}{2}\right)$, there exists a set of distributions $\mathcal{G}$ such that any $D \in \mathcal{G}$ satisfies the weak-moment condition for some $\alpha > 0$ and for any estimator $\widehat{\mu}$:*

$$\mathbf{Pr}_{D \in \mathcal{G}}\left\{ |\widehat{\mu}(\boldsymbol{X}) - \mu(D)| \geqslant \left(\frac{\log(2/\delta)}{n}\right)^{\frac{\alpha}{1+\alpha}} \right\} \geqslant \delta$$

where $\boldsymbol{X} = X_1, \ldots, X_n$ are drawn iid from D.

Finally, we have the proof of Theorem 3.0.2:

Proof of Theorem 3.0.2. When $n > 8d$, the bound follows from Lemma 3.2.3 and Theorem 3.2.4. When $n \leqslant 8d$, the bound follows from the bounded-covariance ($\alpha = 1$) setting in Lemma 3.2.1. □

Chapter 4

Necessary Compromises

In this chapter, we study the statistical performance of *stable* estimators and derive information theoretic lower bounds on their performance. In Chapters 2 and 3, we constructed an efficient algorithmic framework for robust estimation and observed that the statistical performance of these estimators is *optimal* even in extremely noisy settings where the sub-Gassian rate is no longer *possible*. However, these estimators lack the natural affine-equivariant properties of previous estimators such as the Tukey Median [61] and the Stahel-Donoho estimator [57, 23]. On the other hand, these classical estimators lack the strong quantitative guarantees of more recent work. They either lack quantitative guarantees entirely or are sub-optimal.

We investigate this behavior under the two outlier models described in Chapter 1: the heavy-tailed and adversarial contamination models. We find that in both these settings, statistical degradation is *necessary* for affine-equivariant estimators with optimal rates degrading by a factor of $\sqrt{d}$. However, classical estimators are sub-optimal even within this restricted class. To remedy this, we design a novel affine-equivariant estimator with near-*optimal* statistical performance and robustness. Our estimator is based on a novel notion of a high-dimensional median which may be of independent interest.

Formally, we study the robust mean estimation problem where we are given n independent and identically distributed (i.i.d) data points $\boldsymbol{X} = \{X_i\}_{i=1}^n \subset \mathbb{R}^d$ drawn from a distribution, D, with mean μ and variance Σ along with a target failure probability δ. Furthermore, an arbitrarily chosen η fraction of the data points may be corrupted in a possibly adversarial way. The goal, now, is to design an estimator $\widehat{\mu}$ with the smallest r_δ satisfying:

$$\mathbb{P}\left\{\|\widehat{\mu}(\boldsymbol{X}) - \mu\|_\Sigma \leqslant r_\delta\right\} \geqslant 1 - \delta \text{ where } \|x\|_\Sigma \coloneqq \sqrt{x^\top \Sigma^{-1} x}.$$

The above notion of error, commonly referred to as the Mahalanobis Distance, is a natural affine-equivariant metric. Equivalently, the Mahalanobis distance may be viewed as measuring the *Euclidean* distance under the affine transformation that renders the distribution isotropic. Hence, for affine-equivariant estimators, our results characterize the optimal achievable *Euclidean* error for distributions with $\Sigma \preccurlyeq I$[1]. We present our results in the

[1]Note that without any restriction on Σ, no uniform error bound is possible.

Mahalanobis norm as our bounds hold for *any* estimator whose performance is measured in this norm. Furthermore, note that, as before, we make no other assumptions on the data distribution beyond the existence of a mean and variance, hence, allowing for heavy-tailed scenarios where higher moments might not even *exist* and Σ might not even be *estimable* from the given samples.

As a point of comparison, recall from Chapter 1, that in the Euclidean setting where error is measured in the *Euclidean* norm, we have the following characterization of the optimal rate:

$$r_\delta = O\left(\sqrt{\frac{\mathrm{Tr}(\Sigma) + \|\Sigma\| \log(1/\delta)}{n}}\right).$$

When $\eta = 0$, this rate, referred to as the *sub-gaussian* rate, is known to be *optimal* for Gaussians and hence, cannot be improved upon in general. Note that when $\Sigma \preccurlyeq I$, the above rate simplifies to:

$$r_\delta = O\left(\sqrt{\frac{d + \log(1/\delta)}{n}} + \sqrt{\eta}\right).$$

However, for the *Mahalanobis* norm, all known estimators require stronger assumptions to establish quantitative guarantees. Often, these results require the additional property that a multiplicative approximation to Σ is estimable from the samples.

Our upper bound remedying these difficulties is presented in the following theorem:

Theorem 4.0.1. *There exist absolute constants $C_1, C_2 > 0$ such that the following hold. Let $n, d \in \mathbb{N}, \delta \in (0,1)$ and $\eta \in [0, 1/(6d)]$. Suppose $\boldsymbol{X} = \{X_i\}_{i=1}^n$ are generated iid from a distribution D with mean μ and covariance Σ. Then, there exists an affine-equivariant estimator, $\widehat{\mu}$, which when given any η-corrupted version of $\boldsymbol{X}$ satisfies:*

$$\|\widehat{\mu}(\boldsymbol{X}) - \mu\|_\Sigma \leqslant C_1 \left(\sqrt{\frac{d \log(1/\delta)}{n}} + \sqrt{d\eta}\right)$$

with probability at least $1 - \delta$ over $\boldsymbol{X}$ when $n \geqslant C_2 d \log(2/\delta)$.

Furthermore, we exhibit lower bounds establishing that the above rate and restrictions on η are essentially tight. Our lower bounds are proved separately for the heavy-tailed and adversarial settings and hence, our upper bound which hold for *both* these settings simultaneously is optimal. Our first is for the heavy-tailed setting where $\Sigma(D)$ denotes the covariance matrix of D.

Theorem 4.0.2. *There exist absolute constants, $C_1, C_2, c > 0$ such that the following holds. Let $n, d \in \mathbb{N}$ and $\delta \in (0,1)$ be such that $n \geqslant C_1 d \log(1/\delta)$ and $\log(1/\delta) \geqslant C_2 \log(2d)$. Then, there exists a family of distributions $\mathcal{D}$ such that for any estimator $\widehat{\mu}$:*

$$\max_{D \in \mathcal{D}} \mathbf{Pr}_{\boldsymbol{X} \sim D^n} \left\{ \|\widehat{\mu}(\boldsymbol{X}) - \mu(D)\|_{\Sigma(D)} \geqslant c\sqrt{\frac{d \log(1/\delta)}{n \log(d)}} \right\} \geqslant \delta.$$

Next, we present our lower bounds for the adversarial corruption model. Furthermore, our lower bounds hold for the weaker *Huber* contamination model where an adversary is only allowed to *add* corrupted points to the dataset as opposed to corrupting existing points. In the first, we show that the error is unbounded if the corruption fraction exceeds $1/(d+1)$ for estimators that are eventually (as $n \to \infty$) even *approximately* consistent.

Theorem 4.0.3. *For any $d > 3$ and $r > 1$, there exists a family of distributions, $\mathcal{D}$, and a distribution $D_0 \in \mathcal{D}$ such that for any $D \in \mathcal{D}$, there exists distribution P satisfying:*

$$D_0 = \frac{d}{d+1} D + \frac{1}{d+1} P.$$

Furthermore, we have for any estimator, $\widehat{\mu}$, and any $n \in \mathbb{N}$:

$$\sup_{D \in \mathcal{D}} \mathbf{Pr}_{\boldsymbol{X} \sim D_0^n} \left\{ \|\widehat{\mu}(\boldsymbol{X}) - \mu(D)\|_{\Sigma(D)} \geqslant r \right\} \geqslant \frac{1}{d+1}.$$

Next, we show that the dependence on the corruption fraction when $\eta < 1/d$ in Theorem 4.0.1 is tight.

Theorem 4.0.4. *For any $d > 3$ and $\eta < 1/(d+1)$, there exists a family of distributions, $\mathcal{D}$, and a distribution $D_0 \in \mathcal{D}$ such that for any $D \in \mathcal{D}$, there exists distribution P satisfying:*

$$D_0 = (1-\eta) D + \eta P.$$

Furthermore, we have for any estimator, $\widehat{\mu}$, and any $n \in \mathbb{N}$:

$$\sup_{D \in \mathcal{D}} \mathbf{Pr}_{\boldsymbol{X} \sim D_0^n} \left\{ \|\widehat{\mu}(\boldsymbol{X}) - \mu(D)\|_{\Sigma(D)} \geqslant \frac{1}{2} \sqrt{\frac{d\eta}{1 - d\eta}} \right\} \geqslant \frac{1}{d+1}.$$

Taken together, our bounds imply a marked departure from the Euclidean setting. The breakdown point and the dependence of the recovery guarantees on the failure probability and corruption factor *all* decay by a factor of d. In the Euclidean setting, the optimal recovery guarantees essentially match what one would achieve when working with *Gaussian* data. However, in the affine equivariant setting, a significant cost is incurred when weaker assumptions are placed on the data distribution.

Our estimator is based on a novel notion of a high-dimensional median, inspired by the well-known Tukey median [61] and the Stahel-Donoho estimator [57, 23] and may be of independent interest. We aim to find a point whose distance to the mean along any direction is small with respect to (a robust notion of) the variance along that direction. However, the main difficulty in analyzing the estimator is establishing that such a point always exists. We define an appropriate proxy for the variance which guarantees the existence of such a median while allowing for optimal recovery guarantees. Interestingly, our analysis, similar to that of the Tukey Median, relies strongly on Helly's Theorem, a central result in convex geometry.

The key challenge in proving our lower bounds is establishing the correct dependency on the failure probability in the heavy-tailed setting. Our lower bound construction uses a family of distributions with different covariances in a standard Bayesian estimation-to-testing framework for proving minimax lower bounds (see, for example, [62]). In the typical heavy-tailed setting, two lower bounds are established separately, one for the failure probability and another for the dimension, and then subsequently combined to obtain the final bound. However, in our case, these two elements are intimately coupled making the application of standard techniques challenging. To overcome this, we perform an explicit analysis of the posterior over the set of candidate distributions once the data points have been generated, but only for a carefully chosen set of observations. We show that when such samples are obtained, the posterior is well-spread and that any proposed estimate performs poorly on at least some distributions in the support of the posterior. Here, the differences in the covariance matrices across the distributions in the family play a critical role and the sensitivity of the Mahalanobis norm to such differences yield our lower bound.

In this final chapter, we discuss the failure of two classical estimators, the Tukey Median [61] and the Stahel-Donoho estimator [57, 23], in Section 4.1 where we will see that they each fail in complementary ways. We then present our high-dimensional median in Section 4.2 in relation to these two classical notions. Our estimator based on this high-dimensional median is described in Section 4.3 and finally, Section 4.4 contains lower bounds which prove the near-optimality of our estimator.

4.1 Failure of Classical Estimators

In this section, we provide some intuition for our estimator. We analyze the performance of two prominent affine equivariant estimators: the Tukey Median and the Stahel-Donoho estimator. We consider a simple setting where both these estimators perform poorly. We then informally describe how our estimator addresses the shortcomings of these two approaches. We defer the rigorous definition and analysis of our estimator to subsequent sections.

For now, recall the Tukey Median [61] and its associated depth function from Chapter 1

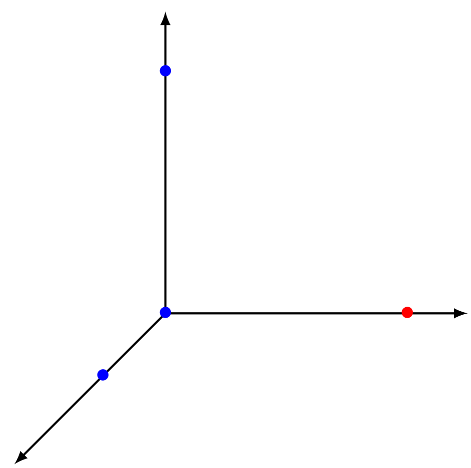
$$
\begin{gathered}
D^1_\tau(y; \boldsymbol{Y}) = \min\left(\lvert\{i : y_i \geqslant y\}\rvert, \lvert\{i : y_i \leqslant y\}\rvert\right) \\
\widehat{\mu}_\tau(\boldsymbol{X}) = \arg\max D^d_\tau(x; \boldsymbol{X}) \text{ where } D^d_\tau(x; \boldsymbol{X}) = \min_{\lVert u\rVert = 1} D^1_\tau\left(\langle u, x\rangle; \{\langle u, x_i\rangle\}_{i=1}^n\right).
\end{gathered}
$$

And the Stahel-Donoho estimator [57, 23] utilizes an alternative notion of *outlyingness*:

$$
\begin{gathered}
D^1_{\mathrm{SD}}(y; \boldsymbol{Y}) = \frac{\lvert y - \mathrm{Med}(\boldsymbol{Y})\rvert}{\mathrm{MAD}(\boldsymbol{Y})} \text{ where } \mathrm{MAD}(\boldsymbol{Y}) = \mathrm{Med}\left(\{\lvert y_i - \mathrm{Med}(\boldsymbol{Y})\rvert\}_{i=1}^n\right) \\
\widehat{\mu}_{\mathrm{SD}}(\boldsymbol{X}) = \arg\min D^d_{\mathrm{SD}}(x; \boldsymbol{X}) \text{ where } D^d_{\mathrm{SD}}(x; \boldsymbol{X}) = \max_{\lVert u\rVert = 1} D^1_{\mathrm{SD}}\left(\langle u, x\rangle; \{\langle u, x_i\rangle\}_{i=1}^n\right)
\end{gathered}
$$

Our hard example will essentially be the simple uniform distribution over the simplex and the origin. However, we will assume one of the standard basis vectors (say e_1) is mildly

Figure 4.1: Illustration of hard distribution. The red dot on e_1 denotes higher probability.



more likely to be observed. Formally, the distribution is defined for parameter ν as follows:

$$\mathbf{Pr}_{X\sim D_\nu}\left\{X = x\right\} = \begin{cases} \frac{1}{d+1} + \nu & \text{if } x = e_1 \\ \frac{1}{d+1} - \frac{\nu}{d} & \text{otherwise} \end{cases}.$$

The example is illustrated in 3 dimensions in Fig. 4.1.

We also assume for the sake of simplicity that the estimators are run directly on the distribution itself as opposed to samples from the distribution where Med and MAD are replaced by their population counterparts. We start with the Tukey median and establish that e_1 is the unique point with largest Tukey depth. Notice that the depth of e_1 is $1/(d+1)+\nu$. Let the support of the distribution be S. For any point not in the convex hull of S, the separating hyperplane theorem ensures that they have Tukey depth 0. Now, consider the case where x belongs to the convex hull and $x \neq e_1$. We consider the two possibilities $x \in T$ where $T = \{e_i\}_{i=2}^d \cup \{\mathbf{0}\}$ and $x \notin T$ separately. First, let $x \in T$ and consider the vector $v = x - \frac{1}{d}\sum_{y\in S\setminus x} y$. We have $\langle x, v\rangle > \langle y, v\rangle$ for all $y \in S \setminus x$. Hence, the depth of x is at most $1/(d+1) - \frac{\nu}{d}$. Secondly, consider the alternative case when $x \notin S$ (but lies in its convex hull). We must have:

$$x = \sum_{y\in S} w_y y \text{ where } w_y \geqslant 0 \text{ and } \sum_{y\in S} w_y = 1.$$

Furthermore, since $x \notin S$, there exists $y \in S$ with $y \neq e_1$ and $0 < w_y < 1$. Consider such a vector y and the vector $v = y - \frac{1}{d}\sum_{z\in S\setminus y} z$. We now get:

$$\forall z \in S \setminus y : \langle v, y\rangle > \langle v, x\rangle > \langle v, z\rangle.$$

Since $y \neq e_1$, the depth of x is also at most $1/(d+1) - \frac{\nu}{d}$. The previous two cases establish that e_1 is the unique point of maximum Tukey depth. Unfortunately, the error of e_1 is rather

Figure 4.2: One-dimensional projection onto e_1.

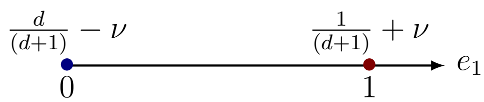


Figure 4.3: One-dimensional projections onto e_i for $i \neq 1$ and $\mathbf{1}$.

large. Consider the one-dimensional projection of the distribution onto e_1:

$$\mathbf{Pr}_{Y \sim D_\nu^1} \{Y = y\} = \begin{cases} \frac{1}{d+1} + \nu & \text{if } y = 1 \\ \frac{d}{d+1} - \nu & \text{if } y = 0 \end{cases}.$$

By considering the error along e_1, we get for $\mu_\nu = \mu(D_\nu)$ for $\nu \leqslant 1/(10d)$:

$$\|e_1 - \mu_\nu\|_{\Sigma(D_\nu)} \geqslant \frac{d/(d+1) - \nu}{\sqrt{(1/(d+1) + \nu)\,(d/(d+1) - \nu)}} = \sqrt{\frac{d/(d+1) - \nu}{1/(d+1) + \nu}} \geqslant \frac{\sqrt{d}}{2}.$$

As we will see later, this error is larger than optimal by a $\sqrt{d}$ factor. The main drawback of the Tukey median is that it remains insensitive to the *variance* along different directions. As illustrated in Fig. 4.2, the true mean (along e_1) lies at $\frac{1}{(d+1)} + \nu$ while the Tukey estimate projects to 1. The variance along e_1 is also at most $E[\langle e_1, X\rangle^2] = \frac{1}{(d+1)} + \nu$. Therefore, incorporating variance information into the estimator can help mitigate some of this degradation. Note, however, that the Tukey median exists for *any* set of data points.

The Stahel-Donoho estimator attempts to incorporate such variance information. However, analyzing the estimator requires non-degeneracy assumptions on the data and even after making these assumptions, they do not provide any *quantifiable* bounds on its performance. For our example, the Stahel-Donoho estimator is not even *defined*. Consider the projection of the distribution onto the standard basis vectors e_i and the all-ones $\mathbf{1}$ direction. From the one-dimensional projections in Fig. 4.3, we have the following straightforward observations:

$$\mathrm{Med}(D_v^1) = \begin{cases} 0 & \text{if } v = e_i \\ 1 & \text{if } v = \mathbf{1} \end{cases} \text{ and } \forall v \in \{e_i\}_{i=1}^d \cup \{\mathbf{1}\} : \mathrm{MAD}(D_v^1) = 0.$$

Consequently, for an estimate x to have finite Stahel-Donoho outlyingness, it must satisfy $\langle x, e_i\rangle = 0$ for all i and $\langle x, \mathbf{1}\rangle = 1$ which is a contradiction.

Our previous discussion shows that the Tukey median and the Stahel-Donoho estimator fail in two complementary ways. The Tukey median is always defined for any set of data points but its failure to incorporate directional variances into its estimation procedure lead to large error. On the other hand, the variance estimates used in the Stahel-Donoho estimator may not allow for a well-defined estimate in certain settings and even when it is defined, existing analyses do not yield quantitative bounds on its performance. As we will see subsequently, our estimator simultaneously addresses the shortcomings of both the Tukey median and the Stahel-Donoho estimator. Our median estimator accounts for directional variances like the Stahel-Donoho median but at the same time, is defined for *any* collection for data points like the Tukey median.

4.2 A High-dimensional Median

In this section, we formally present our high-dimensional median. We demonstrate how it addresses the shortcomings of the Tukey median and the Stahel-Donoho estimator by simultaneously, being well-defined for all point sets and accounting for directional variances. Our estimator is inspired by the Stahel-Donoho estimator but differs in how the robust location and scale parameters are estimated. Recall, the one-dimensional outlyingness function used by the Stahel-Donoho estimator:

$$D^1_{\mathrm{SD}}(y;\boldsymbol{Y}) = \frac{|y - \mathrm{Med}(\boldsymbol{Y})|}{\mathrm{MAD}(\boldsymbol{Y})} \text{ where } \mathrm{MAD}(\boldsymbol{Y}) = \mathrm{Med}\left(\{|y_i - \mathrm{Med}(\boldsymbol{Y})|\}_{i=1}^n\right).$$

The location parameter is robustly estimated by the median and the scale by the median-absolute deviation (MAD) of the one-dimensional point set. The key point of difference between our median and the Stahel-Donoho estimator is a pair of novel location and scale estimation procedures. Defining for a subset $S \subseteq [n]$:

$$\mu_S(\boldsymbol{Y}) \coloneqq \frac{1}{|S|}\sum_{i\in S} y_i \text{ and } \sigma_{1,S}(\boldsymbol{Y}) \coloneqq \frac{1}{|S|}\sum_{i\in S}|y_i - \mu_S(\boldsymbol{Y})|,$$

our location and scale estimates are obtained as follows where $\nu = 1/(3d)$:

1. First, find S satisfying $|S| \geqslant (1-\nu)n$ that minimizes $\sigma_{1,S}(\boldsymbol{Y})$.
2. Second, define location estimate $\widetilde{\mu}(\boldsymbol{Y}) \coloneqq \mu_S(\boldsymbol{Y})$ and scale estimate $\widetilde{\sigma}(\boldsymbol{Y}) \coloneqq \sigma_{1,S}(\boldsymbol{Y})$.

With these one-dimensional location and scale estimates, our median is defined below:

$$D^1_{\mathrm{Ours}}(y;\boldsymbol{Y}) = \frac{|y - \widetilde{\mu}(\boldsymbol{Y})|}{\widetilde{\sigma}(\boldsymbol{Y})}$$

$$\widehat{\mu}_{\mathrm{Ours}}(\boldsymbol{X}) = \arg\min D^d_{\mathrm{Ours}}(x;\boldsymbol{X}) \text{ where } D^d_{\mathrm{Ours}}(x;\boldsymbol{X}) \coloneqq \max_{\|v\|=1} D^1_{\mathrm{Ours}}\left(\langle x, v\rangle; \{\langle x_i, v\rangle\}_{i=1}^n\right)$$

We show that with these definitions, our estimate *always* exists for *any* dataset with finite outlyingness. In fact, we establish the following *strengthening* of this statement:

1. Firstly, we show that the estimate always has *constant* outlyingness. This allows us to prove sharp quantitative bounds in our setting of interest.

2. Secondly, while our definition technically requires choosing S to minimize the directional scale estimate, we show that there exists an estimate with finite depth for *all* choices of S satisfying the size constraints.

This estimator is defined in Algorithm 11 where $\mathrm{Conv}(T)$ denotes the convex hull of T and the proof of its existence is provided in Theorem 4.2.1. The proof relies on Helly's Theorem (Theorem A.1.6), a fundamental result in convex geometry.

Algorithm 11 High-dimensional Median

1: **Input**: Point set $\boldsymbol{X} = \{x_i\}_{i=1}^k \subset \mathbb{R}^d$
2: Let $\nu = 1/(3d)$ and $\mathcal{S} = \{S \subset [k] : |S| \geqslant (1-\nu)k\}$
3: Define for all $v \in \mathbb{S}^{d-1}, S \in \mathcal{S}$:

$$\mu_{v,S} = \mu\left(\{\langle x_i, v\rangle\}_{i\in S}\right) \qquad \sigma_{v,S} = \sigma_1\left(\{\langle x_i, v\rangle\}_{i\in S}\right)$$

4: Define convex compact sets:

$$T_{v,S} = \left\{x \in \mathbb{R}^d : |\langle x, v\rangle - \mu_{v,S}| \leqslant 2\sigma_{v,s}\right\} \cap \mathrm{Conv}(\boldsymbol{Y})$$

5: Let $T = \cap_{v\in\mathbb{S}^{d-1}, S\in\mathcal{S}} T_{v,S}$
6: **Return:** $\mu(T)$

For a point set $\boldsymbol{X} = \{x_i\}_{i=1}^k \subset \mathbb{R}^d$, let $\widehat{\mu}(\boldsymbol{X})$ denote the output of Algorithm 11. The main result of this section establishes the existence and affine-equivariance of $\widehat{\mu}(\cdot)$.

Theorem 4.2.1. *For any $k \in \mathbb{N}$ and $\boldsymbol{X} = \{x_i\}_{i=1}^k \subset \mathbb{R}^d$, $\widehat{\mu}(\boldsymbol{X})$ exists and is well defined. Furthermore, $\widehat{\mu}(\cdot)$ is affine-equivariant.*

Proof. We tackle the two claims of the theorem in turn.

Existence of $\widehat{\mu}$: We first show that T is non-empty, convex, and compact implying the first claim. Note that T is the intersection of compact convex sets and is hence, convex and compact. To establish the non-emptiness of T, an application of Helly's Theorem (Theorem A.1.6) allows us to restrict to finite intersections of the sets $T_{v,S}$. Consider any $d+1$ sized collection $H = \{v_j, S_j\}_{j\in[d+1]}$. We have:

$$R := \cap_{j\in[d+1]} S_j, |R| \geqslant (1-(d+1)\nu)k \geqslant \frac{k}{2}.$$

Defining:

$$\mu_R = \frac{1}{|R|} \cdot \sum_{i \in R} x_i,$$

we will show that μ_R lies in $\cap_{(v,S) \in H} T_{v,S}$. For any $(v, S) \in H$, we have:

$$|\langle \mu_R, v \rangle - \mu_{v,S}| = \left| \frac{1}{|R|} \sum_{i \in R} (\langle x_i, v \rangle - \mu_{v,S}) \right| \leqslant \frac{1}{|R|} \cdot \sum_{i \in R} |\langle x_i, v \rangle - \mu_{v,S}| \leqslant \frac{|S|}{|R|} \sigma_{v,S} \leqslant 2\sigma_{v,S}.$$

An application of Helly's theorem now establishes that T is non-empty proving the claim.

Affine-equivariance of $\widehat{\mu}$: Let $\boldsymbol{X} = \{x_i\}_{i=1}^k \subset \mathbb{R}^d$ and $f(x) = Ax + b$ with $A \in \mathbb{R}^{d \times d}, b \in \mathbb{R}^d$ and A be non-singular. Hence, f is an invertible affine transformation. Furthermore, let $\boldsymbol{X}' = f(\boldsymbol{X}) = \{x_i' = f(x_i)\}_{i=1}^k$ and T be the set obtained in Algorithm 11 on input $\boldsymbol{X}$ and T' be the corresponding set on $\boldsymbol{X}'$. We will show $T' = f(T)$ proving the second claim.

First, let $x \in T$ and we prove $f(x) \in T'$. Observe for any $v \in \mathbb{S}^{d-1}, i \in [k]$:

$$\langle v, x_i' \rangle = \langle v, f(x_i) \rangle = \langle v, Ax_i \rangle + \langle v, b \rangle = \langle A^\top v, x_i \rangle + \langle v, b \rangle = \|A^\top v\| \left\langle \frac{A^\top v}{\|A^\top v\|}, x_i \right\rangle + \langle v, b \rangle.$$

We have by defining $v' = \frac{A^\top v}{\|A^\top v\|}$ for any $S \subset [k]$ with $|S| \geqslant (1 - \nu)k$:

$$\begin{aligned} \mu\left(\{\langle v, x_i' \rangle\}_{i \in S}\right) &= \|A^\top v\| \cdot \mu\left(\{\langle v', x_i \rangle\}_{i \in S}\right) + \langle v, b \rangle \\ \sigma_1\left(\{\langle v, x_i' \rangle\}_{i \in S}\right) &= \|A^\top v\| \cdot \sigma_1\left(\{\langle v', x_i \rangle\}_{i \in S}\right). \end{aligned}$$

As a consequence, we get that for $f(x) = Ax + b$:

$$\begin{aligned} |\langle v, Ax + b \rangle - \mu\left(\{\langle v, x_i' \rangle\}_{i \in S}\right)| &= \|A^\top v\| \cdot |\langle v', x \rangle - \mu\left(\{\langle v', x_i \rangle\}_{i \in S}\right)| \\ &\leqslant 2 \cdot \|A^\top v\| \cdot \sigma_1\left(\{\langle v', x_i \rangle\}_{i \in S}\right) = 2\sigma_1\left(\{\langle v, x_i' \rangle\}_{i \in S}\right) \end{aligned}$$

where the inequality follows from $x \in T$. Since, the above inequality holds for all $x \in T, v \in \mathbb{S}^{d-1}, S \subset [k]$ with $|S| \geqslant (1 - \nu)k$, we get that $f(T) \subseteq T'$. By repeating the above argument for $f^{-1}(z) = A^{-1}z - A^{-1}b$, we get that $f^{-1}(T') \subseteq T$ which implies $T' \subseteq f(T)$ concluding the proof of the theorem. □

4.3 Our Estimator

Here, we prove Theorem 4.0.1 using the high-dimensional median described in the previous section. Our estimator achieving the guarantees of Theorem 4.0.1 is defined in Algorithm 12. Note that since our high-dimensional median is affine-equivariant (Theorem 4.2.1), so is Algorithm 12. Hence, it suffices to establish Theorem 4.0.1 in the setting $\mu = 0$ and $\Sigma = I$.

Algorithm 12 Affine-equivariant Estimator

1: **Input**: Point set $\boldsymbol{X} = \{X_i\}_{i=1}^n \subset \mathbb{R}^d$, Confidence Parameter δ
2: $k \leftarrow \max(6\eta dn, Cd\log(1/\delta))$
3: Partition $\boldsymbol{X}$ into k equally sized buckets $\{\mathcal{B}_i\}_{i\in[k]}$
4: Compute $\widehat{\mu}_i = \mu(\mathcal{B}_i)$
5: $\widehat{\mu} =$ High-dimensional Median$(\{\widehat{\mu}_i\}_{i\in[k]})$
6: **Return:** $\widehat{\mu}$

We first prove the following technical lemma which we will use to establish the required concentration properties on the bucketed means, $\widehat{\mu}_i$. Before we proceed, we define the thresholding operator for a threshold $\tau \geqslant 0$ as follows:

$$\psi_\tau(x) = \begin{cases} x & \text{if } |x| \leqslant \tau \\ \mathrm{sgn}(x)\tau & \text{otherwise} \end{cases}.$$

Lemma 4.3.1. *There exists an absolute constant $C > 0$ such that the following holds. Let $Y_1, \dots, Y_k$ be k iid random vectors drawn from a distribution D with mean μ and variance $\sigma^2 I$ and $\delta \in (0, 1)$. Then, we have for $\tau = 24\sigma d$:*

$$\max_{\|v\|=1} \frac{1}{k} \sum_{i=1}^k |\psi_\tau(\langle v, Y_i\rangle)| \leqslant 2\sigma$$

with probability at least $1 - \delta$ when $k \geqslant Cd\log(2/\delta)$.

Proof. We have with Y being an independent copy from D:

$$Z = \max_{v\in\mathbb{S}^{d-1}} \frac{1}{k} \sum_{i=1}^k |\psi_\tau(\langle Y_i, v\rangle)| - \mathbb{E}\left[|\psi_\tau(\langle Y, v\rangle)|\right].$$

We first bound $\mathbb{E}[Z]$ where Y_i' are independent draws from D and γ_i are independent Rademacher random variables. The third inequality follows from the Ledoux-Talagrand contraction inequality (Corollary A.1.9) and the observation that $|\psi_\tau(\cdot)|$ is 1-Lipschitz:

$$\begin{aligned}
\mathbb{E}[Z] &\leqslant \mathbb{E}\left[\max_{v\in\mathbb{S}^{d-1}}\left|\frac{1}{k}\sum_{i=1}^k |\psi_\tau(\langle Y_i, v\rangle)| - \mathbb{E}\left[|\psi_\tau(\langle Y, v\rangle)|\right]\right|\right] \\
&= \mathbb{E}\left[\max_{v\in\mathbb{S}^{d-1}}\left|\frac{1}{k}\sum_{i=1}^k |\psi_\tau(\langle Y_i, v\rangle)| - |\psi_\tau(\langle Y_i', v\rangle)|\right|\right] \\
&= \mathbb{E}\left[\max_{v\in\mathbb{S}^{d-1}}\left|\frac{1}{k}\sum_{i=1}^k \gamma_i\left(|\psi_\tau(\langle Y_i, v\rangle)| - |\psi_\tau(\langle Y_i', v\rangle)|\right)\right|\right]
\end{aligned}$$

$$\begin{aligned}
&\leqslant 2\,\mathbb{E}\left[\max_{v\in\mathbb{S}^{d-1}}\left|\frac{1}{k}\sum_{i=1}^{k}\gamma_i\left|\psi_\tau\left(\langle Y_i,v\rangle\right)\right|\right|\right] \leqslant 4\,\mathbb{E}\left[\max_{v\in\mathbb{S}^{d-1}}\left|\frac{1}{k}\sum_{i=1}^{k}\gamma_i\langle Y_i,v\rangle\right|\right] \\
&= \frac{4}{k}\,\mathbb{E}\left[\left\|\sum_{i=1}^{k}\gamma_i Y_i\right\|\right] \leqslant \frac{4}{k}\sqrt{\mathbb{E}\left[\left\|\sum_{i=1}^{k}\gamma_i Y_i\right\|^2\right]} = 4\sqrt{\frac{d}{k}}\sigma.
\end{aligned}$$

Additionally, noting that $\psi_\tau(x) \leqslant \tau$ for all $x \in \mathbb{R}$:

$$Y_{i,v} \coloneqq \frac{1}{\tau}\left(|\psi_\tau(\langle Y_i, v\rangle)| - \mathbb{E}[|\psi_\tau(\langle Y, v\rangle)|]\right) \leqslant 1.$$

Furthermore, we have for all $v \in \mathbb{S}^{d-1}$:

$$\sum_{i=1}^{k}\mathbb{E}[Y_{i,v}^2] \leqslant \frac{1}{\tau^2}\sum_{i=1}^{k}\mathbb{E}\left[\psi_\tau\left(\langle Y_i,v\rangle\right)^2\right] \leqslant \frac{1}{\tau^2}\sum_{i=1}^{k}\mathbb{E}\left[\langle Y_i,v\rangle^2\right] = \frac{k\sigma^2}{\tau^2}.$$

Hence, we get by an application of Bousquet's inequality (Theorem A.1.7):

$$\mathbf{Pr}\left\{Z \geqslant \mathbb{E}[Z] + t\right\} \leqslant \exp\left(-\left(\frac{k}{\tau}\right)^2 \cdot \frac{t^2}{2(v + kt/(3\tau))}\right) \text{ where } v = \frac{8\sigma\sqrt{kd}}{\tau} + \frac{k\sigma^2}{\tau^2}.$$

Setting $t = \frac{\sigma}{2}$ and from our setting of τ and k, we get:

$$Z \leqslant \mathbb{E}[Z] + \frac{\sigma}{2} \leqslant \sigma$$

with probability at least $1 - \delta$. The lemma now follows as:

$$\forall \|v\| = 1 : \mathbb{E}\left[|\psi_\tau(\langle Y, v\rangle)|\right] \leqslant \mathbb{E}\left[|\langle Y, v\rangle|\right] \leqslant \sqrt{\mathbb{E}\left[\langle Y, v\rangle^2\right]} = \sigma.$$

$\square$

We now proceed to the proof of Theorem 4.0.1. For the sake of analysis let $\widetilde{\mu}_i$ denote the *uncorrupted* versions of the bucketed means $\widehat{\mu}_i$. For these, we have:

$$\mathbb{E}[\widetilde{\mu}_i] = 0 \qquad \text{and} \qquad \mathbb{E}[\widetilde{\mu}_i\widetilde{\mu}_i^\top] = \frac{k}{n}I.$$

Hence, we get by Lemma 4.3.1 and the setting of k in Algorithm 12:

$$\forall \|v\| = 1 : \frac{1}{k}\sum_{i=1}^{k}|\psi_\tau(\langle v, \widetilde{\mu}_i\rangle)| \leqslant 2\widetilde{\sigma} \text{ where } \widetilde{\sigma} = \sqrt{\frac{k}{n}} \text{ and } \tau = 24\widetilde{\sigma}d$$

with probability at least $1-\delta$. We condition on this event in the remainder of the proof. Note, furthermore, that there are at most ηn many corrupted points in $\boldsymbol{X}$. Therefore, we have for $|\{i : \widetilde{\mu}_i = \widehat{\mu}_i\}| \geqslant (1 - 1/(6d))k$, again from the setting of k in Algorithm 12, that:

$$\forall \|v\| = 1 : \frac{1}{k}\sum_{i=1}^{k} \mathbf{1}\left\{|\langle v, \widetilde{\mu}_i \rangle| \geqslant \tau\right\} \leqslant \frac{1}{\tau} \cdot \frac{1}{k}\sum_{i=1}^{k} |\psi_\tau(\langle v, \widetilde{\mu}_i \rangle)| \leqslant \frac{1}{12d}.$$

Therefore, we get from the previous two observations that:

$$\forall \|v\| = 1 : |\mathcal{G}_v| \geqslant \left(1 - \frac{1}{4d}\right) k \text{ where } \mathcal{G}_v = \{i : \widetilde{\mu}_i = \widehat{\mu}_i \text{ and } \psi_\tau(\langle v, \widetilde{\mu}_i \rangle) = \langle v, \widetilde{\mu}_i \rangle\}.$$

Now, let $v \in \mathbb{S}^{d-1}$. We have for $\mathcal{G}_v$ from Algorithm 11 and Theorem 4.2.1:

$$|\langle v, \widehat{\mu} \rangle - \mu(\{\langle v, \widehat{\mu}_i \rangle\}_{i \in \mathcal{G}_v})| \leqslant 2\sigma_1(\{\langle v, \widehat{\mu}_i \rangle\}_{i \in \mathcal{G}_v}).$$

For the mean term, we get:

$$|\mu(\{\langle v, \widehat{\mu}_i \rangle\}_{i \in \mathcal{G}_v})| \leqslant \mu(\{|\langle v, \widehat{\mu}_i \rangle|\}_{i \in \mathcal{G}_v}) \leqslant \frac{1}{(1 - 1/(4d))} \mu(\{|\psi_\tau(\langle v, \widetilde{\mu}_i \rangle)|\}_{i \in [k]}) \leqslant 3\widetilde{\sigma}.$$

For the deviation term, we get:

$$\sigma_1(\{\langle v, \widehat{\mu}_i \rangle\}_{i \in \mathcal{G}_v}) = \mu(\{|\langle v, \widehat{\mu}_i \rangle - \mu(\{\langle v, \widehat{\mu}_i \rangle\}_{i \in \mathcal{G}_v})|\}_{i \in \mathcal{G}_v}) \leqslant \mu(\{|\langle v, \widehat{\mu}_i \rangle|\}_{i \in \mathcal{G}_v}) + 3\widetilde{\sigma} \leqslant 6\widetilde{\sigma}.$$

The above two bounds imply:

$$\forall v \in \mathbb{S}^{d-1} : |\langle v, \widehat{\mu} \rangle| \leqslant 15\widetilde{\sigma} = 15\sqrt{\frac{k}{n}}$$

establishing the theorem. □

4.4 Lower Bounds

Here, we present the proofs of Theorems 4.0.2 to 4.0.4 which show that the guarantees of Theorem 4.0.1 are nearly tight. For the heavy-tailed setting with *no* adversarial corruption (i.e $\eta = 0$), Theorem 4.0.2 shows that the recovery error of our estimator is optimal up to a $\sqrt{\log(d)}$ factor and for the adversarial corruption model, Theorem 4.0.3 establishes that *no* affinely-equivariant estimator can achieve breakdown point greater than $1/(d+1)$ while Theorem 4.0.4 shows that $\sqrt{d\eta}$ is the best achievable recovery error for *any* affinely-equivariant estimator.

Heavy-tailed Lower Bound - Proof of Theorem 4.0.2

To define our class of distributions, let:

$$\varepsilon = \frac{1}{4}\sqrt{\frac{d\log(1/(d\delta))}{n\log(d)}}.$$

Our hard class will contain d distributions with support over the standard basis vectors and the origin, i.e., $\{e_i\}_{i=1}^d \cup \{\mathbf{0}\}$. Each distribution puts a smaller mass at one of the standard basis vectors. More formally, we have $\mathcal{D} = \{D_i\}_{i=1}^d$ with:

$$\mathbf{Pr}_{X\sim D_i}\left\{X = e_j\right\} = \begin{cases} \frac{\varepsilon^2}{d} & \text{if } i \neq j \\ \frac{\varepsilon^2}{d^2} & \text{if } i = j \end{cases},$$

and

$$\mathbf{Pr}_{X\sim D_i}\left\{X = \mathbf{0}\right\} = 1 - \frac{d-1}{d}\varepsilon^2 - \frac{\varepsilon^2}{d^2}.$$

By a straightforward calculation, we have:

$$\Sigma(D_i) \preccurlyeq M^i \text{ where } M^i_{jk} = \begin{cases} 0 & \text{if } j \neq k \\ \frac{\varepsilon^2}{d} & \text{if } j = k \text{ and } j \neq i \\ \frac{\varepsilon^2}{d^2} & \text{if } j = k = i \end{cases}.$$

Now consider the following procedure of generating the data $\boldsymbol{X}$:

1. Sample a random integer I from the index set $\{1, 2, \ldots, d\}$.
2. Given $I = i$, draw n i.i.d samples $\boldsymbol{X} = \{X_1, X_2, \ldots, X_n\}$ from D_i.

Next, it suffices to show that for any estimator $\widehat{\mu}(\cdot)$, we have

$$\mathbf{Pr}\left\{\|\widehat{\mu}(\boldsymbol{X}) - \mu(D_I)\|_{\Sigma(D_I)} \geqslant \frac{1}{4}\varepsilon\right\} \geqslant \delta.$$

For each distribution D_i, consider a set of instances

$$S_i = \left\{\boldsymbol{X} = (X_1, \ldots, X_n) : m_i(\boldsymbol{X}) \geqslant \frac{4\varepsilon^2 n}{d} \text{ and } \sum_{j=1}^{d} \mathbf{1}\left\{m_j(\boldsymbol{X}) < \frac{4\varepsilon^2 n}{d}\right\} \geqslant \frac{d}{2}\right\}$$

$$\text{where } m_j(\boldsymbol{X}) := \sum_{k=1}^{n} \mathbf{1}\left\{X_k = e_j\right\}.$$

Next, consider $S := \cup S_i$. Now for any $\boldsymbol{X} \in S$, let

$$\mathcal{J} := \left\{j : m_j(\boldsymbol{X}) < \frac{4\varepsilon^2 n}{d}\right\} \text{ and } z := \widehat{\mu}(\boldsymbol{X}).$$

Suppose $z_j \leqslant \frac{\varepsilon^2}{2d}$ for any $j \in \mathcal{J}$, by considering the cumulative error on $\mathcal{J}$ we have

$$\|\widehat{\mu}(\boldsymbol{X}) - \mu(D_k)\|^2_{\Sigma(D_k)} \geqslant \left(\frac{d}{2} - 1\right) \frac{\left(\varepsilon^2/2d\right)^2}{\varepsilon^2/d} \geqslant \frac{1}{16}\varepsilon^2$$

for any D_k.

On the other hand, suppose there exists $j \in \mathcal{J}$ such that $z_j > \frac{\varepsilon^2}{2d}$. If $I = j$ is the sampled index, then by considering the error on e_j we have

$$\|\widehat{\mu}(\boldsymbol{X}) - \mu(D_j)\|_{\Sigma(D_j)} \geqslant \frac{|\varepsilon^2/2d - \varepsilon^2/d^2|}{\sqrt{\varepsilon^2/d^2}} \geqslant \frac{1}{4}\varepsilon.$$

By the definition of $\boldsymbol{X}$, let i be the index such that $m_i(\boldsymbol{X}) \geqslant \frac{4\varepsilon^2 n}{d}$, then we have the posterior probability of $I = j$ is at least that of $I = i$. So

$$\mathbf{Pr}\left\{\|\widehat{\mu}(\boldsymbol{X}) - \mu(D_I)\|_{\Sigma(D_I)} \geqslant \frac{1}{4}\varepsilon \middle| \boldsymbol{X}\right\} \geqslant \mathbf{Pr}\left\{I = j|\boldsymbol{X}\right\} \geqslant \mathbf{Pr}\left\{I = i|\boldsymbol{X}\right\}.$$

Putting pieces together, we have

$$\begin{aligned}
\mathbf{Pr}\left\{\|\widehat{\mu}(\boldsymbol{X}) - \mu(D_I)\|_{\Sigma(D_I)} \geqslant \frac{1}{4}\varepsilon\right\} &\geqslant \sum_{\boldsymbol{X} \in S_i} \mathbf{Pr}\left\{\|\widehat{\mu}(\boldsymbol{X}) - \mu(D_I)\|_{\Sigma(D_I)} \geqslant \frac{1}{4}\varepsilon \middle| \boldsymbol{X}\right\} \mathbf{Pr}\left\{\boldsymbol{X}\right\} \\
&\geqslant \sum_{\boldsymbol{X} \in S_i} \mathbf{Pr}\left\{I = i|\boldsymbol{X}\right\} \mathbf{Pr}\left\{\boldsymbol{X}\right\} = \sum_{\boldsymbol{X} \in S_i} \mathbf{Pr}\left\{I = i, \boldsymbol{X}\right\} \\
&= \sum_{\boldsymbol{X} \in S_i} \mathbf{Pr}\left\{\boldsymbol{X}|I = i\right\} \mathbf{Pr}\left\{I = i\right\}.
\end{aligned}$$

It remains to prove that $\mathbf{Pr}\left\{S_i|I = i\right\} \geqslant d\delta$. For the simplicity of notations, denote $\mathbf{Pr}_i$ as the conditional distribution of $\boldsymbol{X}$ under $I = i$. Define events:

$$A = \left\{\boldsymbol{X} = (X_1, \ldots, X_n) : m_i(\boldsymbol{X}) \geqslant \frac{4\varepsilon^2 n}{d}\right\}$$

$$B = \left\{\boldsymbol{X} = (X_1, \ldots, X_n) : \sum_{j=1, j\neq i}^{d} \mathbf{1}\left\{m_j(\boldsymbol{X}) < \frac{4\varepsilon^2 n}{d}\right\} \geqslant \frac{d}{2}\right\}$$

$$C = \left\{\boldsymbol{X} = (X_1, \ldots, X_n) : m_0(\boldsymbol{X}) \geqslant (1 - 2\varepsilon^2)n\right\} \text{ where } m_0(\boldsymbol{X}) := \sum_{k=1}^{n} \mathbf{1}\left\{X_k = 0\right\}.$$

Note that $\mathbf{Pr}_i\left\{S_i\right\} = \mathbf{Pr}_i\left\{A \cap B\right\}$ and $C \subseteq B$. So, we have

$$\mathbf{Pr}_i(A \cap B) = \mathbf{Pr}_i(B|A)\mathbf{Pr}_i(A) \geqslant \mathbf{Pr}_i(B)\mathbf{Pr}_i(A) \geqslant \mathbf{Pr}_i(C)\mathbf{Pr}_i(A).$$

We first use a Binomial tail lower bound to bound $\mathbf{Pr}_i(A)$ (see e.g., [3]):

$$\mathbf{Pr}_i\left\{B(n,p) \geqslant k\right\} \geqslant \frac{1}{\sqrt{8n\frac{k}{n}\left(1-\frac{k}{n}\right)}} \exp\left(-nD\left(\frac{k}{n} \,\middle\|\, p\right)\right),$$

where $B(n,p)$ denotes a Binomial random variable and $D(a \parallel p) = a\log\frac{a}{p} + (1-a)\log\frac{1-a}{1-p}$ denotes the KL divergence.

Plugging in $k = 4dnp$ and $p = \varepsilon^2/d^2$ we obtain that

$$\mathbf{Pr}_i(A) \geqslant \frac{1}{\sqrt{32ndp}} \exp\left(-4ndp\log\left(4d\right)\right) \geqslant 2d\delta.$$

Finally, note that $\mathbf{Pr}(C) \geqslant 1/2$ since $n - m_0(\boldsymbol{X})$ is positive with $\mathbb{E}[n - m_0(\boldsymbol{X})] \leqslant \varepsilon^2 n$. Therefore, we have $\mathbf{Pr}_i(S_i) \geqslant d\delta$ concluding the proof. □

Adversarial Contamination - Proofs of Theorems 4.0.3 and 4.0.4

We start with Theorem 4.0.3 which establishes an upper bound on the breakdown point.

Proof of Theorem 4.0.3 . Let $r > 0$ and $S = \{e_i\}_{i=1}^d \cup \{\mathbf{0}\}$. First define the family $\widetilde{\mathcal{D}} = \{\widetilde{D}_i\}_{i=0}^{d+1}$ with $\widetilde{D}_0$ and $\widetilde{D}_{d+1}$ denoting the uniform distributions over S and $S \setminus \{\mathbf{0}\}$ respectively and for $i \in [d]$, $\widetilde{D}_i$ is defined as follows:

$$\mathbf{Pr}_{X\sim\widetilde{D}_i}\left\{X = x\right\} = \begin{cases} 0 & \text{if } x = e_i \\ \frac{1}{d} & \text{if } x \in S \setminus \{e_i\} \end{cases}.$$

Now, our hard family of distributions $\mathcal{D} = \{D_i\}_{i=0}^{d+1}$ is defined in the following way:

1. First, generate $\widetilde{X} \sim \widetilde{D}_i$
2. Independently, generate $Z \sim \mathrm{Unif}(\{\pm 1\}^d)$
3. Observe $X = \widetilde{X} + \frac{Z}{(2dr)^3}$.

Note, that $\Sigma(D_i)$ is non-singular for each i and D_0 may be written as a mixture of D_i and the distribution with all its mass on e_i for every i. Now, suppose $\widehat{\mu}$ is an estimator that satisfies for some $n \in \mathbb{N}$:

$$\forall D \in \mathcal{D} : \mathbf{Pr}_{\boldsymbol{X}\sim D_0^n}\left\{\|\widehat{\mu}(\boldsymbol{X}) - \mu(D)\|_{\Sigma(D)} \geqslant r\right\} < \frac{1}{d+1}.$$

Then, by the union bound, there must exist a sample $\boldsymbol{X}$ in the support of D_0^n such that:

$$\forall D \in \mathcal{D} \setminus \{D_0\} : \|\widehat{\mu}(\boldsymbol{X}) - \mu(D)\|_{\Sigma(D)} \leqslant r.$$

Then, letting $\widehat{\mu} = \widehat{\mu}(\boldsymbol{X})$, we must have for any $i \in [d]$:

$$r \geqslant \|\widehat{\mu} - \mu(D_i)\|_{\Sigma(D_i)} \geqslant (dr)^3 |\widehat{\mu}_i|.$$

This implies:

$$|\widehat{\mu}_i| \leqslant \frac{1}{d^3 r^2} \implies \sum_{i=1}^{d} |\widehat{\mu}_i| \leqslant \frac{1}{(dr)^2}.$$

However, note that we have for the direction $\mathbf{1}/\sqrt{d}$ and the distribution D_{d+1}:

$$r \geqslant \|\widehat{\mu} - \mu(D_{d+1})\|_{\Sigma(D_{d+1})} \geqslant (dr)^3 \cdot \left(\frac{1}{\sqrt{d}} - \frac{1}{d^{2.5}} \right) \geqslant (dr)^2$$

which is a contradiction thus establishing the theorem. □

We now move on to Theorem 4.0.4.

Proof of Theorem 4.0.4. As before, we will construct a hard family of distributions. For support set $S = \{e_i\}_{i=1}^d \cup \{\mathbf{1}/d\}$, define the set of distributions $\widetilde{\mathcal{D}} = \{\widetilde{D}_i\}_{i=0}^d$ defined as follows:

$$\forall i \in [d] : \mathbf{Pr}_{X \sim \widetilde{D}_i}(X = x) = \begin{cases} 0 & \text{if } x = e_i \\ \frac{d}{d-1}\eta & \text{if } x = e_j \text{ for } j \neq i \\ 1 - d\eta & \text{if } x = \frac{\mathbf{1}}{d} \end{cases} \text{ and}$$

$$\mathbf{Pr}_{X \sim \widetilde{D}_0}(X = x) = \begin{cases} \eta & \text{if } x = e_j \text{ for any } j \in [d] \\ 1 - d\eta & \text{if } x = \frac{\mathbf{1}}{d} \end{cases}.$$

Let $\sigma = ((\eta(1 - d\eta))/8d)^4$ and the hard family of distributions is defined as follows:

1. First, generate $\widetilde{X} \sim \widetilde{D}_i$
2. Independently, generate $Z \sim \mathrm{Unif}(\{\pm 1\}^d)$
3. Observe $X = \widetilde{X} + \sigma Z$.

Note that $\Sigma(D_i)$ is non-singular for each i and D_0 may be written as a mixture of D_i and the distribution with all its mass on e_i for every i. Now, suppose $\widehat{\mu}$ is an estimator that satisfies for some $n \in \mathbb{N}$:

$$\forall D \in \mathcal{D} : \mathbf{Pr}_{\boldsymbol{X} \sim D_0^n} \left\{ \|\widehat{\mu}(\boldsymbol{X}) - \mu(D)\|_{\Sigma(D)} \geqslant \frac{1}{2}\sqrt{\frac{d\eta}{1 - d\eta}} \right\} < \frac{1}{d+1}.$$

Then, by the union bound, there must exist a sample $\boldsymbol{X}$ in the support of D_0^n such that:

$$\forall D \in \mathcal{D} : \|\widehat{\mu}(\boldsymbol{X}) - \mu(D)\|_{\Sigma(D)} \leqslant \frac{1}{2}\sqrt{\frac{d\eta}{1 - d\eta}}.$$

Letting $\widehat{\mu} = \widehat{\mu}(\boldsymbol{X})$, we have for the direction $\mathbf{1}/\sqrt{d}$:

$$\frac{1}{2}\sqrt{\frac{d\eta}{1-d\eta}} \geqslant \|\widehat{\mu} - \mu(D_0)\|_{\Sigma(D_0)} \geqslant \frac{1}{\sqrt{d}\sigma}\left|\sum_{i=1}^{d}\widehat{\mu}_i - 1\right|.$$

This implies for our setting of σ that:

$$\sum_{i=1}^{d}\widehat{\mu}_i \geqslant 1 - \frac{\eta^2}{4}.$$

Therefore, there exists $i \in [d]$ with:

$$\widehat{\mu}_i \geqslant \frac{4-\eta^2}{4d}.$$

For this i, we have:

$$\|\widehat{\mu} - \mu(D_i)\|_{\Sigma(D_i)} \geqslant \frac{d}{\sqrt{d\eta(1-d\eta)}+\sigma} \cdot \left|\frac{4-\eta^2}{4d} - (1-d\eta)\frac{1}{d}\right| > \frac{1}{2}\sqrt{\frac{d\eta}{1-d\eta}}$$

which is a contradiction concluding the proof of the theorem. □

Bibliography

[1] F. Alizadeh. "Interior point methods in semidefinite programming with applications to combinatorial optimization". In: *SIAM Journal on Optimization* 5.1 (1995), pp. 13–51 (page 20).

[2] Noga Alon, Yossi Matias, and Mario Szegedy. "The space complexity of approximating the frequency moments". In: vol. 58. 1, part 2. Twenty-eighth Annual ACM Symposium on the Theory of Computing (Philadelphia, PA, 1996). 1999, pp. 137–147. DOI: 10.1006/jcss.1997.1545. URL: https://doi.org/10.1006/jcss.1997.1545 (pages 3, 5, 9).

[3] Robert B Ash. *Information theory*. Courier Corporation, 2012 (page 58).

[4] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. A nonasymptotic theory of independence, With a foreword by Michel Ledoux. Oxford University Press, Oxford, 2013, pp. x+481. ISBN: 978-0-19-953525-5. DOI: 10.1093/acprof:oso/9780199535255.001.0001. URL: https://doi.org/10.1093/acprof:oso/9780199535255.001.0001 (pages 70–72).

[5] Olivier Bousquet. "Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms". PhD thesis. Ecole Polytechnique, 2002 (page 72).

[6] Gavin Brown, Marco Gaboardi, Adam D. Smith, Jonathan R. Ullman, and Lydia Zakynthinou. "Covariance-Aware Private Mean Estimation Without Private Covariance Estimation". In: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*. Ed. by Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan. 2021, pp. 7950–7964. URL: https://proceedings.neurips.cc/paper/2021/hash/42778ef0b5805a96f9511e20b5611fce-Abstract.html (page 6).

[7] Gavin Brown, Samuel B. Hopkins, and Adam D. Smith. "Fast, Sample-Efficient, Affine-Invariant Private Mean and Covariance Estimation for Subgaussian Distributions". In: *CoRR* abs/2301.12250 (2023). DOI: 10.48550/arXiv.2301.12250. arXiv: 2301.12250. URL: https://doi.org/10.48550/arXiv.2301.12250 (page 6).

[8] Zihao Chen and Yeshwanth Cherapanamjeri. "Statistical Barriers to Affine-equivariant Estimation". In preparation. 2023 (page 7).

[9] Yu Cheng, Ilias Diakonikolas, and Rong Ge. "High-Dimensional Robust Mean Estimation in Nearly-Linear Time". In: *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*. Ed. by Timothy M. Chan. SIAM, 2019, pp. 2755–2771. DOI: 10.1137/1.9781611975482.171. URL: https://doi.org/10.1137/1.9781611975482.171 (page 4).

[10] Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L. Bartlett. "Fast Mean Estimation with Sub-Gaussian Rates". In: *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*. Ed. by Alina Beygelzimer and Daniel Hsu. Vol. 99. Proceedings of Machine Learning Research. PMLR, 2019, pp. 786–806. URL: http://proceedings.mlr.press/v99/cherapanamjeri19b.html (page 6).

[11] Yeshwanth Cherapanamjeri, Samuel B. Hopkins, Tarun Kathuria, Prasad Raghavendra, and Nilesh Tripuraneni. "Algorithms for heavy-tailed statistics: regression, covariance estimation, and beyond". In: *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Chicago, IL, USA, June 22-26, 2020*. Ed. by Konstantin Makarychev, Yury Makarychev, Madhur Tulsiani, Gautam Kamath, and Julia Chuzhoy. ACM, 2020, pp. 601–609. ISBN: 978-1-4503-6979-4. DOI: 10.1145/3357713.3384329. URL: https://doi.org/10.1145/3357713.3384329 (pages 6, 8).

[12] Yeshwanth Cherapanamjeri, Nilesh Tripuraneni, Peter L. Bartlett, and Michael I. Jordan. "Optimal Mean Estimation without a Variance". In: *Conference on Learning Theory, 2-5 July 2022, London, UK*. Ed. by Po-Ling Loh and Maxim Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, 2022, pp. 356–357. URL: https://proceedings.mlr.press/v178/cherapanamjeri22a.html (page 6).

[13] B. S. Cirel′ son, I. A. Ibragimov, and V. N. Sudakov. "Norms of Gaussian sample functions". In: *Proceedings of the Third Japan-USSR Symposium on Probability Theory (Tashkent, 1975)*. Lecture Notes in Math., Vol. 550. Springer, Berlin, 1976, pp. 20–41 (page 71).

[14] P. L. Davies. "Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices". In: *Ann. Statist.* 15.3 (1987), pp. 1269–1292. ISSN: 0090-5364. DOI: 10.1214/aos/1176350505. URL: https://doi.org/10.1214/aos/1176350505 (pages 3, 4, 11).

[15] Jules Depersin and Guillaume Lecué. "On the robustness to adversarial corruption and to heavy-tailed data of the Stahel–Donoho median of means". In: *Information and Inference: A Journal of the IMA* 12.2 (2022), pp. 814–850. DOI: 10.1093/imaiai/iaac026 (page 6).

[16] Jules Depersin and Guillaume Lecué. "Robust sub-Gaussian estimation of a mean vector in nearly linear time". In: *Ann. Statist.* 50.1 (2022), pp. 511–536. ISSN: 0090-5364. DOI: 10.1214/21-aos2118. URL: https://doi.org/10.1214/21-aos2118 (pages 6, 8).

[17] Luc Devroye, Matthieu Lerasle, Gábor Lugosi, and Roberto I. Oliveira. "Sub-Gaussian mean estimators". In: *Ann. Statist.* 44.6 (2016), pp. 2695–2725. ISSN: 0090-5364. DOI: 10.1214/16-AOS1440. URL: https://doi.org/10.1214/16-AOS1440 (pages 32, 44).

[18] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. "Robust estimators in high dimensions without the computational intractability". In: *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*. Ed. by Irit Dinur. IEEE Computer Society, 2016, pp. 655–664. ISBN: 978-1-5090-3933-3. DOI: 10.1109/FOCS.2016.85. URL: https://doi.org/10.1109/FOCS.2016.85 (page 4).

[19] Ilias Diakonikolas and Daniel M Kane. "Recent advances in algorithmic high-dimensional robust statistics". In: *arXiv preprint arXiv:1911.05911* (2019) (page 5).

[20] Ilias Diakonikolas, Daniel M. Kane, and Ankit Pensia. "Outlier Robust Mean Estimation with Subgaussian Rates via Stability". In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Ed. by Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin. 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/13ec9935e17e00bed6ec8f06230e33a9-Abstract.html (page 6).

[21] Yihe Dong, Samuel B. Hopkins, and Jerry Li. "Quantum Entropy Scoring for Fast Robust Mean Estimation and Improved Outlier Detection". In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Ed. by Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett. 2019, pp. 6065–6075. URL: https://proceedings.neurips.cc/paper/2019/hash/a4d92e2cd541fca87e4620aba658316d-Abstract.html (page 4).

[22] David Donoho and Peter J. Huber. "The notion of breakdown point". In: *A Festschrift for Erich L. Lehmann*. Wadsworth Statist./Probab. Ser. Wadsworth, Belmont, CA, 1983, pp. 157–184 (pages 3, 10).

[23] David L Donoho. *Breakdown properties of multivariate location estimators*. Tech. rep. Technical report, Harvard University, Boston., 1982 (pages 3, 4, 11, 45, 47, 48).

[24] David L. Donoho and Miriam Gasko. "Breakdown properties of location estimates based on halfspace depth and projected outlyingness". In: *Ann. Statist.* 20.4 (1992), pp. 1803–1827. ISSN: 0090-5364. DOI: 10.1214/aos/1176348890. URL: https://doi.org/10.1214/aos/1176348890 (pages 3, 4, 11).

[25] David L. Donoho and Richard C. Liu. "The "automatic" robustness of minimum distance functionals". In: *Ann. Statist.* 16.2 (1988), pp. 552–586. ISSN: 0090-5364. DOI: 10.1214/aos/1176350820. URL: https://doi.org/10.1214/aos/1176350820 (pages 3, 4, 11).

[26] John C. Duchi, Saminul Haque, and Rohith Kuditipudi. "A Fast Algorithm for Adaptive Private Mean Estimation". In: *CoRR* abs/2301.07078 (2023). DOI: 10.48550/arXiv.2301.07078. arXiv: 2301.07078. URL: https://doi.org/10.48550/arXiv.2301.07078 (page 6).

[27] P. M. Gruber and J. M. Wills, eds. *Handbook of convex geometry. Vol. A, B.* North-Holland Publishing Co., Amsterdam, 1993, Vol. A: lxvi+735 pp., Vol. B: pp. i–lxvi and 737–1438. ISBN: 0-444-89598-1 (page 72).

[28] Eduard Helly. "Über Mengen konvexer Körper mit gemeinschaftlichen Punkte." In: *Jahresbericht der Deutschen Mathematiker-Vereinigung* 32 (1923), pp. 175–176. URL: http://eudml.org/doc/145659 (page 72).

[29] Wassily Hoeffding. "Probability inequalities for sums of bounded random variables". In: *J. Amer. Statist. Assoc.* 58 (1963), pp. 13–30. ISSN: 0162-1459. URL: http://links.jstor.org/sici?sici=0162-1459(196303)58:301%3C13:PIFSOB%3E2.0.CO;2-D&origin=MSN (page 70).

[30] Samuel B. Hopkins. "Mean estimation with sub-Gaussian rates in polynomial time". In: *Ann. Statist.* 48.2 (2020), pp. 1193–1213. ISSN: 0090-5364. DOI: 10.1214/19-AOS1843. URL: https://doi.org/10.1214/19-AOS1843 (pages 5, 6, 8, 12, 19, 21, 26, 27).

[31] Samuel B. Hopkins, Gautam Kamath, and Mahbod Majid. "Efficient mean estimation with pure differential privacy via a sum-of-squares exponential mechanism". In: *STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing, Rome, Italy, June 20 - 24, 2022.* Ed. by Stefano Leonardi and Anupam Gupta. ACM, 2022, pp. 1406–1417. DOI: 10.1145/3519935.3519947. URL: https://doi.org/10.1145/3519935.3519947 (page 6).

[32] Samuel B. Hopkins, Gautam Kamath, Mahbod Majid, and Shyam Narayanan. "Robustness Implies Privacy in Statistical Estimation". In: *Proceedings of the 55th Annual ACM Symposium on Theory of Computing, STOC 2023, Orlando, FL, USA, June 20-23, 2023.* Ed. by Barna Saha and Rocco A. Servedio. ACM, 2023, pp. 497–506. DOI: 10.1145/3564246.3585115. URL: https://doi.org/10.1145/3564246.3585115 (page 6).

[33] Samuel B. Hopkins, Jerry Li, and Fred Zhang. "Robust and Heavy-Tailed Mean Estimation Made Simple, via Regret Minimization". In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Ed. by Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin. 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/8a1276c25f5efe85f0fc4020fbf5b4f8-Abstract.html (page 6).

[34] Peter J. Huber. "Robust covariances". In: *Statistical decision theory and related topics, II (Proc. Sympos., Purdue Univ., Lafayette, Ind., 1976)*. Academic Press, New York, 1977, pp. 165–191 (pages 3, 10).

[35] Peter J. Huber. "Robust estimation of a location parameter". In: *Ann. Math. Statist.* 35 (1964), pp. 73–101. ISSN: 0003-4851. DOI: 10.1214/aoms/1177703732. URL: https://doi.org/10.1214/aoms/1177703732 (pages 2, 3).

[36] Mark R. Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. "Random generation of combinatorial structures from a uniform distribution". In: *Theoret. Comput. Sci.* 43.2-3 (1986), pp. 169–188. ISSN: 0304-3975. DOI: 10.1016/0304-3975(86)90174-X. URL: https://doi.org/10.1016/0304-3975(86)90174-X (pages 3, 5, 9).

[37] Pravesh Kothari, Pasin Manurangsi, and Ameya Velingker. "Private Robust Estimation by Stabilizing Convex Relaxations". In: *Conference on Learning Theory, 2-5 July 2022, London, UK*. Ed. by Po-Ling Loh and Maxim Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, 2022, pp. 723–777. URL: https://proceedings.mlr.press/v178/kothari22a.html (page 6).

[38] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*. Classics in Mathematics. Isoperimetry and processes, Reprint of the 1991 edition. Springer-Verlag, Berlin, 2011, pp. xii+480. ISBN: 978-3-642-20211-7 (page 72).

[39] Zhixian Lei, Kyle Luh, Prayaag Venkat, and Fred Zhang. "A Fast Spectral Algorithm for Mean Estimation with Sub-Gaussian Rates". In: *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*. Ed. by Jacob D. Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, 2020, pp. 2598–2612. URL: http://proceedings.mlr.press/v125/lei20a.html (page 6).

[40] Regina Y. Liu. "Data depth and multivariate rank tests". In: *L_1-statistical analysis and related methods (Neuchâtel, 1992)*. North-Holland, Amsterdam, 1992, pp. 279–294 (pages 3, 4, 11).

[41] Regina Y. Liu. "On a notion of data depth based on random simplices". In: *Ann. Statist.* 18.1 (1990), pp. 405–414. ISSN: 0090-5364. DOI: 10.1214/aos/1176347507. URL: https://doi.org/10.1214/aos/1176347507 (pages 4, 11).

[42] Regina Y. Liu, Jesse M. Parelius, and Kesar Singh. "Multivariate analysis by data depth: descriptive statistics, graphics and inference". In: *Ann. Statist.* 27.3 (1999). With discussion and a rejoinder by Liu and Singh, pp. 783–858. ISSN: 0090-5364. DOI: 10.1214/aos/1018031260. URL: https://doi.org/10.1214/aos/1018031260 (pages 3, 4, 11).

[43] Regina Y. Liu and Kesar Singh. "Ordering directional data: concepts of data depth on circles and spheres". In: *Ann. Statist.* 20.3 (1992), pp. 1468–1484. ISSN: 0090-5364. DOI: 10.1214/aos/1176348779. URL: https://doi.org/10.1214/aos/1176348779 (pages 3, 4, 11).

[44] Xiyang Liu, Weihao Kong, Sham M. Kakade, and Sewoong Oh. "Robust and differentially private mean estimation". In: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual.* Ed. by Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan. 2021, pp. 3887–3901. URL: https://proceedings.neurips.cc/paper/2021/hash/1fc5309ccc651bf6b5d22470f67561ea-Abstract.html (page 6).

[45] Hendrik P. Lopuhaä and Peter J. Rousseeuw. "Breakdown points of affine equivariant estimators of multivariate location and covariance matrices". In: *Ann. Statist.* 19.1 (1991), pp. 229–248. ISSN: 0090-5364. DOI: 10.1214/aos/1176347978. URL: https://doi.org/10.1214/aos/1176347978 (pages 3, 4, 11).

[46] Gábor Lugosi and Shahar Mendelson. *Multivariate mean estimation with direction-dependent accuracy.* 2020. arXiv: 2010.11921 [math.ST] (page 6).

[47] Gábor Lugosi and Shahar Mendelson. "Robust multivariate mean estimation: the optimality of trimmed mean". In: *Ann. Statist.* 49.1 (2021), pp. 393–410. ISSN: 0090-5364. DOI: 10.1214/20-AOS1961. URL: https://doi.org/10.1214/20-AOS1961 (page 6).

[48] Gábor Lugosi and Shahar Mendelson. "Sub-Gaussian estimators of the mean of a random vector". In: *Ann. Statist.* 47.2 (2019), pp. 783–794. ISSN: 0090-5364. DOI: 10.1214/17-AOS1639. URL: https://doi.org/10.1214/17-AOS1639 (pages 3, 5, 6, 8, 11–13, 16, 21, 26).

[49] Ricardo Antonio Maronna. "Robust M-estimators of multivariate location and scatter". In: *Ann. Statist.* 4.1 (1976), pp. 51–67. ISSN: 0090-5364. URL: http://links.jstor.org/sici?sici=0090-5364(197601)4:1%3C51:ROMLAS%3E2.0.CO;2-I&origin=MSN (pages 3, 4, 10, 11).

[50] Colin McDiarmid. "On the method of bounded differences". In: *Surveys in combinatorics, 1989 (Norwich, 1989).* Vol. 141. London Math. Soc. Lecture Note Ser. Cambridge Univ. Press, Cambridge, 1989, pp. 148–188 (page 70).

[51] Arshak Minasyan and Nikita Zhivotovskiy. "Statistically Optimal Robust Mean and Covariance Estimation for Anisotropic Gaussians". In: *CoRR* abs/2301.09024 (2023). DOI: 10.48550/arXiv.2301.09024. arXiv: 2301.09024. URL: https://doi.org/10.48550/arXiv.2301.09024 (page 6).

[52] Arkadi S. Nemirovsky and David B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience Series in Discrete Mathematics. Translated from the Russian and with a preface by E. R. Dawson. John Wiley & Sons, Inc., New York, 1983, pp. xv+388. ISBN: 0-471-10345-4 (pages 3, 5, 9).

[53] Y. Nesterov. "Semidefinite relaxation and nonconvex quadratic optimization". In: *Optimization Methods and Software* 9.1-3 (1998), pp. 141–160 (page 27).

[54] Hannu Oja. "Descriptive statistics for multivariate distributions". In: *Statist. Probab. Lett.* 1.6 (1983), pp. 327–332. ISSN: 0167-7152. DOI: 10.1016/0167-7152(83)90054-8. URL: https://doi.org/10.1016/0167-7152(83)90054-8 (pages 3, 4, 11).

[55] P. Rousseeuw and V. Yohai. "Robust regression by means of S-estimators". In: *Robust and nonlinear time series analysis (Heidelberg, 1983)*. Vol. 26. Lect. Notes Stat. Springer, New York, 1984, pp. 256–272. DOI: 10.1007/978-1-4615-7821-5_15. URL: https://doi.org/10.1007/978-1-4615-7821-5_15 (pages 3, 11).

[56] Peter Rousseeuw. "Multivariate estimation with high breakdown point". In: *Mathematical statistics and applications, Vol. B (Bad Tatzmannsdorf, 1983)*. Reidel, Dordrecht, 1985, pp. 283–297 (pages 3, 4, 11).

[57] Werner A. Stahel. "Robuste Schätzungen. infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen". de. Doctoral Thesis. Zürich: ETH Zurich, 1981. DOI: 10.3929/ethz-a-000231580 (pages 3, 4, 11, 45, 47, 48).

[58] Michel Talagrand. "New concentration inequalities in product spaces". In: *Invent. Math.* 126.3 (1996), pp. 505–563. ISSN: 0020-9910. DOI: 10.1007/s002220050108. URL: https://doi.org/10.1007/s002220050108 (page 72).

[59] Michel Talagrand. "Sharper bounds for Gaussian and empirical processes". In: *Ann. Probab.* 22.1 (1994), pp. 28–76. ISSN: 0091-1798. URL: http://links.jstor.org/sici?sici=0091-1798(199401)22:1%3C28:SBFGAE%3E2.0.CO;2-W&origin=MSN (page 72).

[60] J. W. Tukey. "A survey of sampling from contaminated distributions". In: *Contributions to Probability and Statistics* (1960), pp. 448–485 (page 3).

[61] John W. Tukey. "Mathematics and the picturing of data". In: *Proceedings of the International Congress of Mathematicians (Vancouver, B.C., 1974), Vol. 2*. Canad. Math. Congress, Montreal, Que., 1975, pp. 523–531 (pages 3, 4, 10, 45, 47, 48).

[62] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, Cambridge, 2019. DOI: 10.1017/9781108627771. URL: https://doi.org/10.1017/9781108627771 (pages 33, 48).

[63] Víctor J. Yohai. “High breakdown-point and high efficiency robust estimates for regression”. In: *Ann. Statist.* 15.2 (1987), pp. 642–656. ISSN: 0090-5364. DOI: 10.1214/aos/1176350366. URL: https://doi.org/10.1214/aos/1176350366 (pages 3, 4, 11).

[64] Yijun Zuo. “Projection-based depth functions and associated medians”. In: *Ann. Statist.* 31.5 (2003), pp. 1460–1490. ISSN: 0090-5364. DOI: 10.1214/aos/1065705115. URL: https://doi.org/10.1214/aos/1065705115 (pages 4, 11).

[65] Yijun Zuo, Hengjian Cui, and Xuming He. “On the Stahel-Donoho estimator and depth-weighted means of multivariate data”. In: *Ann. Statist.* 32.1 (2004), pp. 167–188. ISSN: 0090-5364. DOI: 10.1214/aos/1079120132. URL: https://doi.org/10.1214/aos/1079120132 (pages 4, 11).

Appendix A

Auxiliary Material

A.1 Empirical Processes and Concentration Results

Here, we collect results from empirical process theory, concentration inequalities, and convex analysis that we use in our proofs. The first is Hoeffding's Inequality [29] as stated in [4]:

Theorem A.1.1 ([29, 4]). *Let $X_1, \ldots X_n$ be independent random variables such that X_i takes its values in $[a_i, b_i]$ almost surely for all $i \leqslant n$. Let*

$$S = \sum_{i=1}^{n} (X_i - \mathbb{E}\, X_i).$$

Then, for every $t > 0$,

$$\mathbf{Pr}\{S \geqslant t\} \leqslant \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

We also require McDiarmid's bounded differences inequality [50].

Theorem A.1.2 ([50, 4]). *Let $n \in \mathbb{N}$, $\mathcal{X}$ denote some domain and assume that $f : \mathcal{X}^n \to \mathbb{R}$ satisfies for some constants $c_1, \ldots, c_n$:*

$$\forall i \in [n] : \sup_{\substack{x_1, \ldots, x_n \\ x_i' \in \mathcal{X}}} |f(x_1, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leqslant c_i.$$

Now, denote:

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

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

$$\mathbf{Pr}\{Z - \mathbb{E}\, Z \geqslant t\} \leqslant e^{-t^2/(2\nu)}.$$

Next, we have the concentration of Lipschitz functions of Gaussians [13].

Theorem A.1.3 ([13, 4])**.** *Let $X = (X_1, \dots, X_n)$ be a vector of n independent standard normal variables. Let $f : \mathbb{R}^n \to \mathbb{R}$ denote an L-Lipschitz function. Then, for all $t \geqslant 0$,*

$$\mathbf{Pr}\left\{f(X) - \mathbb{E}\, f(X) \geqslant t\right\} \leqslant e^{-t^2/(2L^2)}.$$

We also need the Gaussian Poincare Inequality from [4].

Theorem A.1.4. *Let $X = (X_1, \dots, X_n)$ be a vector of n i.i.d standard Gaussian variables. Let $f : \mathbb{R}^n \to \mathbb{R}$ be any continuously differentiable function. Then, we have:*

$$\mathrm{Var}(f(X)) \leqslant \mathbb{E}\left[\|\nabla f(X)\|^2\right].$$

We reprove the following simple lemma.

Lemma A.1.5. *Let $X \sim \mathcal{N}(0, I_n)$. Then, we have for all $\delta \in (0,1)$:*

$$\mathbf{Pr}\left\{\sqrt{n-1} - \sqrt{2\log(2/\delta)} \leqslant \|X\| \leqslant \sqrt{n} + \sqrt{2\log(2/\delta)}\right\} \leqslant \delta.$$

Proof. Consider $f(X) = \|X\|$. Note that $f(\cdot)$ is 1-Lipschitz. Hence, we may apply Theorem A.1.3. It remains to bound $\mathbb{E}[f(X)]$. For the upper bound, we have:

$$\mathbb{E}[\|X\|] \leqslant \sqrt{\mathbb{E}\left[\|X\|^2\right]} \leqslant \sqrt{n}.$$

For the lower bound, consider $f_\gamma(X) = g_\gamma(f(X))$ for $0 \leqslant \gamma \leqslant 1$ where:

$$g_\gamma(x) = \begin{cases} \frac{x^2}{2\gamma} & \text{if } |x| \leqslant \gamma \\ |x| - \frac{\gamma}{2} & \text{o.w} \end{cases}.$$

Note that $f_\gamma(\cdot)$ is differentiable everywhere and $\|\nabla f_\gamma(\cdot)\| \leqslant 1$. Hence, we get by the Gaussian Poincare Inequality (Theorem A.1.4):

$$\begin{aligned}\mathrm{Var}(f_\gamma(X)) &\leqslant \mathbb{E}\left[\|\nabla f_\gamma(X)\|^2\right] \\ &= \mathbb{E}\left[\|\nabla f_\gamma(X)\|^2 \mathbf{1}\left\{\|X\| \geqslant \gamma\right\}\right] + \mathbb{E}\left[\|\nabla f_\gamma(X)\|^2 \mathbf{1}\left\{\|X\| < \gamma\right\}\right] \\ &\leqslant 1 + \mathbf{Pr}\left\{\|X\| \leqslant \gamma\right\}.\end{aligned}$$

By taking $\gamma \to 0$, we get:

$$\mathrm{Var}(f(X)) = \lim_{\gamma \to 0} \mathrm{Var}(f_\gamma(X)) \leqslant 1.$$

Hence, we get:

$$\mathbb{E}[f(X)] \geqslant \sqrt{\mathbb{E}[f^2(X)] - \mathrm{Var}(f(X))} \geqslant \sqrt{d-1}.$$

□

We now recall Helly's celebrated theorem [28] on convex intersections as stated in [27, Theorem 1.1, Chapter 2.1].

Theorem A.1.6 ([28, 27])**.** *Let $\mathcal{K}$ be a family of convex sets in $\mathbb{R}^d$, and suppose $\mathcal{K}$ is finite or each member of $\mathcal{K}$ is compact. If every $d+1$ or fewer members of $\mathcal{K}$ have a common point, then there is a point common to all members of $\mathcal{K}$.*

Next, we present Bousquet's inequality on the suprema of empirical processes [5] which builds on prior results by Talagrand [59, 58].

Theorem A.1.7 ([5, 4])**.** *Let $X_1, \ldots, X_n$ be independent identically distributed random vectors indexed by an index set $\mathcal{T}$. Assume that $\mathbb{E}[X_{i,s}] = 0$, and $X_{i,s} \leqslant 1$ for all $s \in \mathcal{T}$. Let $Z = \sup_{s\in\mathcal{T}} \sum_{i=1}^n X_{i,s}$, $\nu = 2\,\mathbb{E}\,Z + \sigma^2$ where $\sigma^2 = \sup_{s\in\mathcal{T}} \sum_{i=1}^n \mathbb{E}\,X_{i,s}^2$ is the wimpy variance. Let $\phi(u) = e^u - u - 1$ and $h(u) = (1+u)\log(1+u) - u$, for $u \geqslant -1$. Then for all $\lambda \geqslant 0$,*

$$\log \mathbb{E}\, e^{\lambda(Z - \mathbb{E}\,Z)} \leqslant \nu\phi(\lambda).$$

Also, for all $t \geqslant 0$,

$$\mathbb{P}\left\{Z \geqslant \mathbb{E}\,Z + t\right\} \leqslant e^{-\nu h(t/\nu)} \leqslant \exp\left(-\frac{t^2}{2(\nu + t/3)}\right).$$

We also require the Ledoux-Talagrand contraction inequality [38] (again as stated in [4]).

Theorem A.1.8 ([38, 4])**.** *Let $x_1, \ldots, x_n$ be vectors whose real-valued components are indexed by $\mathcal{T}$, that is, $x_i = (x_{i,s})_{s\in\mathcal{T}}$. For each $i = 1, \ldots, n$, let $\phi_i : \mathbb{R} \to \mathbb{R}$ be a 1-Lipschitz function such that $\phi_i(0) = 0$. Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent Rademacher random variables, and let $\Psi : [0, \infty) \to \mathbb{R}$ be a non-decreasing convex function. Then,*

$$\mathbb{E}\left[\Psi\left(\sup_{s\in\mathcal{T}} \sum_{i=1}^n \varepsilon_i \phi_i(x_{i,s})\right)\right] \leqslant \mathbb{E}\left[\Psi\left(\sup_{s\in\mathcal{T}} \sum_{i=1}^n \varepsilon_i x_{i,s}\right)\right]$$

and

$$\mathbb{E}\left[\Psi\left(\frac{1}{2}\sup_{s\in\mathcal{T}} \left|\sum_{i=1}^n \varepsilon_i \phi_i(x_{i,s})\right|\right)\right] \leqslant \mathbb{E}\left[\Psi\left(\sup_{s\in\mathcal{T}} \left|\sum_{i=1}^n \varepsilon_i x_{i,s}\right|\right)\right].$$

We will use the following simple corollary of the second conclusion in our proofs.

Corollary A.1.9. *Assume the setting of Theorem A.1.8. Then,*

$$\mathbb{E}\left[\sup_{s\in\mathcal{T}} \left|\sum_{i=1}^n \varepsilon_i \phi_i(x_{i,s})\right|\right] \leqslant 2\,\mathbb{E}\left[\sup_{s\in\mathcal{T}} \left|\sum_{i=1}^n \varepsilon_i x_{i,s}\right|\right].$$

A.2 Auxiliary Results from Chapter 2

Lemma A.2.1. *There exist absolute constants $c, C > 0$ such that the following holds. Let $\boldsymbol{X} = X_1, \ldots, X_n$ be n i.i.d random vectors drawn from a distribution P with mean μ satisfying:*

$$\mathop{\mathbb{E}}_{X \sim P} [\|X - \mu\|] \leqslant \sigma.$$

Then, Algorithm 5 on input $\boldsymbol{X}$, returns an estimate $\widehat{x}$ satisfying:

$$\|\widehat{x} - \mu\| \leqslant 30\sigma$$

with probability at least $1 - e^{-cn}$.

Proof. We have for any $i \in [n]$:

$$\mathbf{Pr}\left\{\|X_i - \mu\| \leqslant 10\sigma\right\} \geqslant \frac{9}{10}.$$

Hence, we get by Hoeffding's (Theorem A.1.1) inequality:

$$\sum_{i=1}^{n} \mathbf{1}\left\{\|X_i - \mu\| \leqslant 10\sigma\right\} \geqslant 0.75n$$

with probability at least $1 - e^{-cn}$. Condition on this event and let $\mathcal{G} = \{X_i : \|X_i - \mu\| \leqslant 10\sigma\}$. We get for any $x \in \mathcal{G}$ by the triangle inequality:

$$\sum_{i=1}^{n} \mathbf{1}\left\{\|X_i - x\| \leqslant 20\sigma\right\} \geqslant 0.75n.$$

Therefore, we get for the solution $\widehat{x}$ returned by Algorithm 5:

$$\min\left\{r > 0 : \sum_{i=1}^{n} \mathbf{1}\left\{\|X_i - \widehat{x}\| \leqslant r\right\} \geqslant 0.6n\right\} \leqslant 20\sigma.$$

Furthermore, for $\widehat{\mathcal{G}} = \{X_i : \|X_i - \widehat{x}\| \leqslant 20\sigma\}$, we have $\mathcal{G} \cap \widehat{\mathcal{G}} \neq \phi$ and hence, for $y \in \mathcal{G} \cap \widehat{\mathcal{G}}$:

$$\|\widehat{x} - \mu\| \leqslant \|\widehat{x} - y\| + \|y - \mu\| \leqslant 30\sigma$$

concluding the proof. □

Lemma A.2.2. *For any $\boldsymbol{Z} \in \mathbb{R}^{k \times d}$ and $x \in \mathbb{R}^d$, the optimal value of $\boldsymbol{MT}(x, r, \boldsymbol{Z})$ is monotonically non-increasing in r.*

Proof. The lemma follows trivially from the fact that a feasible solution X of $\mathbf{MT}(x, r, \boldsymbol{Z})$ is also a feasible solution for $\mathbf{MT}(x, r', \boldsymbol{Z})$ for $r' \leqslant r$. □

A.3 Auxiliary Results from Chapter 3

In this section, we establish a lower bound for robust mean estimation under weak moments. The lower bound will be a consequence of the following theorem:

Theorem A.3.1. *Given $\eta, \alpha \in (0,1)$, there exist two distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ over $\mathbb{R}$ with means μ_1 and μ_2, respectively, satisfying:*

1. $d_{TV}(\mathcal{D}_1, \mathcal{D}_2) \leqslant \frac{\eta}{4}$
2. $|\mu_1 - \mu_2| \geqslant \frac{1}{4} \cdot \eta^{\alpha/(1+\alpha)}$
3. $\mathbb{E}_{X\sim\mathcal{D}_1}[|X-\mu_1|^{1+\alpha}], \mathbb{E}_{X\sim\mathcal{D}_2}[|X-\mu_2|^{1+\alpha}] \leqslant 1.$

Proof. We prove the theorem by explicit construction. Let $\mathcal{D}_1$ be a δ-distribution on 0: $\mathbb{P}_{X\sim\mathcal{D}_1}(X=0) = 1$. We have $\mu_1 = 0$ and the weak moment condition holds trivially for $\mathcal{D}_1$. Now, for $\mathcal{D}_2$, we have:

$$\mathbb{P}_{X\sim\mathcal{D}_2}(X=x) = \begin{cases} 1-\frac{\eta}{4}, & \text{when } x = 0 \\ \frac{\eta}{4}, & \text{when } x = \left(\frac{1}{\eta}\right)^{1/(1+\alpha)} \\ 0, & \text{otherwise.} \end{cases}.$$

From the definitions of $\mathcal{D}_1$ and $\mathcal{D}_2$, we obtain the first conclusion. By direct computation, we have $\mu_1 = 0$ and $\mu_2 = \frac{1}{4} \cdot \eta^{\alpha/(1+\alpha)}$ establishing the second. Finally, we verify the weak moment condition on $\mathcal{D}_2$ using the convexity of the function $f(x) = |x|^{1+\alpha}$:

$$\mathbb{E}_{X\sim\mathcal{D}_2}[|X-\mu_2|^{1+\alpha}] \leqslant 2^\alpha \cdot \mathbb{E}[|X|^{1+\alpha} + |\mu_2|^{1+\alpha}] \leqslant 2^\alpha \left(\frac{1}{4} + \frac{\eta^\alpha}{4^{(1+\alpha)}}\right) \leqslant 1.$$

This concludes the proof of the theorem. □