

Heavy-tailed Estimation is Easier than Adversarial Contamination

Yeshwanth Cherapanamjeri

CSAIL, Massachusetts Institute of Technology

YESH@MIT.EDU

Daniel Lee

CSAIL, Massachusetts Institute of Technology

LEE.D@MIT.EDU

Editors: Nika Haghtalab and Ankur Moitra

Abstract

A large body of work in the statistics and computer science communities dating back to Huber (Huber, 1960) has led to the development of statistically and computationally efficient estimators robust to the presence of outliers in data. In the course of these developments, two particular outlier models have received significant attention: the adversarial and heavy-tailed contamination models. While the former models outliers as the result of a potentially malicious adversary inspecting and manipulating the data, the latter instead relaxes the assumptions on the distribution generating the data allowing outliers to naturally occur as part of the data generating process. In the first setting, the goal is to develop estimators robust to the largest fraction of outliers while in the second, one seeks estimators to combat the loss of *statistical* efficiency caused by outliers, where the dependence on the failure probability is paramount.

Surprisingly, despite these distinct motivations, the algorithmic approaches to both these settings have converged, prompting questions on the relationship between the corruption models. In this paper, we investigate and provide a principled explanation for this phenomenon. First, we prove that *any* adversarially robust estimator is also resilient to heavy-tailed outliers for *any* statistical estimation problem with i.i.d data. As a corollary, optimal adversarially robust estimators for mean estimation, linear regression, and covariance estimation are also optimal heavy-tailed estimators. Conversely, for arguably the simplest high-dimensional estimation task of mean estimation, we establish the existence of optimal heavy-tailed estimators whose application to the adversarial setting *requires* any black-box reduction to remove *almost all the outliers* in the data. Taken together, our results imply that heavy-tailed estimation is likely easier than adversarially robust estimation opening the door to novel algorithmic approaches bypassing the computational barriers inherent to the adversarial setting. Additionally, *any* confidence intervals obtained for adversarially robust estimation also hold with high-probability.

The proof of our reduction from heavy-tailed to adversarially robust estimation rests on the isoperimetry properties of the set of adversarially robust datasets. Conversely, we identify novel structural properties for samples drawn from a heavy-tailed distribution. We show that such a sample obeys a logarithmic tail-decay condition scaling with the target failure probability. This allows for a quantile-smoothed heavy-tailed estimator which *requires arbitrarily large* stable subsets of the input data to succeed. In the process of analyzing this estimator, we also strengthen the analysis of algorithms utilized previously in the literature.

Keywords: robust statistics, heavy-tailed estimation, high-dimensional statistics

1. Introduction

Over the past several decades, statistical methods have been, and continue to be, employed in an increasingly diverse range of applications. However, classical statistical estimation procedures such as ordinary least squares (OLS) are extremely susceptible to the presence of *outliers* in the data, an

increasingly prevalent issue in the large-scale datasets ubiquitous in modern machine learning. Furthermore, the sheer scale of the data used in these systems makes manual data curation extremely challenging. The challenge of mitigating the effects of these outliers has been extensively investigated over the past 60 years by the statistics and computer science communities leading to the development of estimators with near-optimal statistical and algorithmic guarantees.

For several fundamental high-dimensional estimation tasks, such as mean estimation, covariance estimation, and linear regression, much recent work has focused on two distinct models for the presence of outliers in data. In the *adversarial model*, one assumes that a computationally unbounded adversary is allowed to inspect the data and arbitrarily corrupt a fraction of the samples. In contrast, the *heavy-tailed model* relaxes the distributional assumptions on the data, removing the stringent requirements on the higher-order moments of the data distribution¹, allowing the natural occurrence of outliers as part of the data generation process. Surprisingly, despite their distinct motivations, the algorithmic approaches for these settings have converged with several estimators from one applicable to the other. The goal of the current paper is to shed light on this phenomenon by addressing the following fundamental question:

What is the fundamental relationship between these corruption models that enables algorithmic transference? Are they algorithmically equivalent? Is one corruption model strictly weaker than the other?

In this paper, we address the above question by establishing a formal relationship between the two corruption models. For a general class of statistical estimation problems, we show through a black-box analysis that *any* estimator resilient to *adversarial outliers* is also resilient to *heavy-tailed* ones². Conversely, *no* such guarantees hold for the other direction. Even for the arguably simplest robust estimation problem of *high-dimensional mean estimation*, there exist estimators resilient to heavy-tailed outliers that break down when faced with adversarial corruption. In fact, we establish the much stronger statement that *any* black-box reduction, even one which may potentially alter the points presented to the heavy-tailed estimator, must produce a pointset with a negligible fraction of adversarial outliers, effectively filtering out most of the outliers in the dataset.

Adversarial contamination. The adversarial contamination model is a family of models dating back to the work of Huber (1964), and the statistical and computational limits of estimation under this model are well understood. The idea is to model outliers via an adversary who is able to corrupt the sample. In particular, in this paper we focus on the *strong contamination model*. In this model, a computationally unbounded adversary is allowed to inspect the entire sample, and arbitrarily change any ε fraction of the data.

One simple consequence of this definition is that, immediately, most classical estimators (such as empirical mean, empirical covariance, or ordinary least squares) fail to provide even *finite* bounds. For example, even by changing only one point in the sample, the adversary can arbitrarily skew the empirical mean. The difficulty in this model is therefore even getting guarantees with a *constant* probability of success.

-
1. Typically, these only require the finiteness of a small number of lower-order moments as opposed to much stronger ones such as Gaussianity or sub-Gaussianity which restrict *all* the moments of the distribution.
 2. In this paper, we restrict our analysis to *deterministic* estimators. This avoids pathological scenarios such as those where the estimator itself may fail with constant probability (say, 0.05). However, our analysis is also applicable to settings where the failure probability of the *estimator* is chosen to be sufficiently small (say, $\exp(-n)$).

Heavy-tailed Model. In the heavy-tailed setting, outliers are not modeled through an explicit adversary but rather a natural consequence of relaxing the assumptions on the data distribution. Here, one studies the loss in *statistical efficiency* incurred by relaxing, say a Gaussian or sub-Gaussian assumption on the data, to a broader family of distributions which only restrict a small number of lower-order moments (2 or 4 for the applications in the present paper). By studying the decay of the recovery guarantees with respect to the failure probability, this model allows for outliers to occur as samples from the tail events of the underlying distribution. Surprisingly, in many settings, it is possible to obtain statistical rates which match those obtainable under a sub-Gaussian assumption.

Relationship between the two models. Despite the a priori very different motivations and challenges of these two models, there has been a convergence in the algorithmic approaches to solving them. Recent work has constructed algorithms that achieve optimal statistical rates *while* being optimally robust to outliers. In fact, in several cases, algorithms developed for one setting are directly applicable to the other with minimal modifications [Depersin and Lecué \(2019\)](#); [Lugosi and Mendelson \(2020\)](#); [Mendelson and Zhivotovskiy \(2020\)](#); [Hopkins et al. \(2020b\)](#); [Mourtada et al. \(2021\)](#); [Diakonikolas et al. \(2021\)](#); [Lugosi and Mendelson \(2021\)](#); [Cherapanamjeri et al. \(2022\)](#); [Abdalla and Zhivotovskiy \(2024\)](#). The works of [Diakonikolas et al. \(2021\)](#) and [Hopkins et al. \(2020b\)](#) explain this phenomenon by identifying *sufficient* structural properties that enable this transference of *specific* algorithmic approaches for restricted setting of *mean estimation*.

In contrast, the goal of this paper is to understand the fundamental relationship between these two contamination models that allows algorithms developed for one setting to transfer to the other for a broad range of estimation tasks. We aim to understand the limits of *black-box* reductions (see [Section 1.1.2](#)) between these two models for arbitrary estimation tasks with i.i.d inputs.

1.1. Our Contributions

We now move on to the presentation of our results which are a pair of theorems concerning the black-box reducibility of one model to another.

1.1.1. THE ADVERSARIAL CONTAMINATION MODEL IS AT LEAST AS STRONG.

Our first main result shows that the adversarial contamination model supersedes the heavy-tailed model. We show that *any* adversarially robust algorithm which uses no internal randomness is simultaneously resilient to adversarial contamination while also succeeding with high-probability. In order to state our result in full generality, we formally define the adversarial model we’re considering as well as a definition of adversarially robust algorithms for generic estimation problems.

Definition 1 (Strong Adversarial Contamination Model) First, n data points $\widetilde{\mathbf{X}} := \{X_i\}_{i=1}^n$ are drawn i.i.d from a distribution \mathcal{D} . An adversary then inspects $\widetilde{\mathbf{X}}$ and constructs \mathbf{X}_ε by arbitrarily corrupting any ε fraction of $\widetilde{\mathbf{X}}$. We refer to \mathbf{X}_ε as an ε -corrupted sample of $\widetilde{\mathbf{X}}$.

Definition 2 (Generic Adversarially Robust Estimator) Let \mathcal{D} be a distribution. $G_{\mathcal{D}}(\alpha) \subseteq \mathcal{Z}$ the set of acceptable solutions, for corruption factor ε . We say that \mathcal{A} is an adversarially robust estimator, if

$$\mathbb{P}_{\mathbf{X}, \mathcal{A}} \{ \mathcal{A}(\mathbf{X}_\varepsilon; \varepsilon) \in G_{\mathcal{D}}(\varepsilon) \} \geq 9/10$$

The definition of $G_{\mathcal{D}}(\varepsilon)$ corresponds to the optimal confidence intervals for ε -corruption setting and its definition varies by the particular instantiations of the estimation problem. Note that in the above definition, there's a minimization over all potential adversaries which is left implicit. Our main result here is that asking for constant probability of success *immediately implies* high-probability guarantees (therefore, satisfying the conditions of the heavy-tailed model).

Theorem 3 (Informal - see Theorem 15) *For any constant $\varepsilon \in (0, 0.1)$, there exists an absolute constant $c > 0$ such that any adversarially-robust estimation algorithm with no internal randomness satisfying Definition 2, when given corruption parameter ε and clean sample \mathbf{X} satisfies:*

$$\mathbb{P} \{ \mathcal{A}(\mathbf{X}, \varepsilon) \in G_{\mathcal{D}}(\varepsilon) \} \geq 1 - \exp(-cn).$$

As stated previously, we observe that this restriction on having no internal randomness is necessary for a simple reason: one can have an adversarially robust algorithm that just flips a biased coin such that 5 percent of the time it always fails. Clearly, this algorithm will not satisfy exponentially small tail bounds. Unless otherwise stated, we will assume in this paper that algorithms are deterministic and do not rely on any internal sources of randomness.

Note that in the language of black-box reductions, the *trivial* reduction which passes the input to the algorithm directly suffices. Furthermore, our result places no assumptions on how the adversarially robust algorithm is implemented as opposed to prior work which analyze *specific* algorithms Diakonikolas et al. (2021); Hopkins et al. (2020b) for the restricted setting of mean estimation.

The proof of Theorem 3 rests on the following observation: if an algorithm produces an accurate estimate on \mathbf{X} , it must also produce an accurate estimate on data sets \mathbf{X}' whose Hamming distance from \mathbf{X} is at most $0.1n$. With this observation, Theorem 3 follows from classical results on isoperimetry and concentration to show that all but an exponentially small fraction (in terms of the product measure over \mathcal{D}) of possible datasets must be within $0.1n$ of the set of good samples.

To illustrate the power of this theorem, we apply it to the problems of mean estimation, covariance estimation, and linear regression, in each case showing that an optimal algorithm in the adversarial contamination model implies an algorithm that satisfies optimal high probability guarantees in the heavy-tailed model. Our results are presented in the extreme regime of constant corruption fraction ε (corresponding to $\delta \sim e^{-\Theta(n)}$).

Mean estimation. Our first application is the canonical estimation problem of high-dimensional mean estimation. Here, we are given access to n i.i.d samples from a distribution with mean μ and covariance Σ with the goal of optimally estimating μ in Euclidean norm. Perhaps surprisingly, only recently, a long line of work in the computer science literature Diakonikolas et al. (2016); Lai et al. (2016); Diakonikolas et al. (2017); Depersin and Lecu   (2019); Hopkins et al. (2020b); Diakonikolas et al. (2021) has culminated in the development of statistically and computationally efficient estimators satisfying the following optimal guarantee for mean estimation.

Definition 4 (Optimal Adversarially Robust Mean Estimator) *We say \mathcal{A} is an optimal adversarially robust mean estimator if given any $\varepsilon \in [0, 1/10]$, and any ε -corrupted sample of $\tilde{\mathbf{X}}$, drawn i.i.d from \mathcal{D} produces an estimate $\hat{\mu}$ satisfying:*

$$\mathbb{P}_{\mathbf{X}} \left\{ \|\hat{\mu} - \mu\| \leq C \left(\sqrt{\frac{\text{Tr}(\Sigma)}{n}} + \sqrt{\|\Sigma\|\varepsilon} \right) \right\} \geq \frac{9}{10}$$

for some absolute constant $C > 0$.

Note that, in contrast with the heavy-tailed model to follow, we only ask for a *constant probability guarantee*. The problem isn't getting a high probability bound, it's to get any bound at all.

In contrast, in the heavy-tailed model, some finite bounds are possible on the performance of the empirical mean in this setting. It can be shown (Catoni (2012)) that the following bounds are tight for the empirical mean, denoted by $\hat{\mu}_{\text{EM}}$:

$$\mathbb{E} [\|\hat{\mu}_{\text{EM}} - \mu\|^2] \leq \frac{\text{Tr}(\Sigma)}{n} \text{ and } \mathbb{P} \left\{ \|\hat{\mu}_{\text{EM}} - \mu\| \leq \sqrt{\frac{\text{Tr}(\Sigma)}{n\delta}} \right\} \geq 1 - \delta. \quad (1)$$

While the first in-expectation result is the optimal achievable, the second is substantially worse than the guarantees obtained in more restricted settings, such as when the data is obtained from a *Gaussian* distribution. In a surprising development, the seminal work of Lugosi and Mendelson (2019) shows that such rates are in fact statistically achievable for the bounded covariance setting as outlined in the following definition:

Definition 5 We say \mathcal{A} is an optimal heavy-tailed mean estimator if given any $\delta \in [e^{-cn}, 1/2]$, and sample $\mathbf{X} = \{X_i\}_{i=1}^n$ drawn i.i.d from \mathcal{D} returns an estimate, $\hat{\mu}$, satisfying:

$$\mathbb{P}_{\mathbf{X}, \mathcal{A}} \left\{ \|\hat{\mu} - \mu\| \leq C \left(\sqrt{\frac{\text{Tr}(\Sigma)}{n}} + \sqrt{\frac{\|\Sigma\| \log(1/\delta)}{n}} \right) \right\} \geq 1 - \delta$$

for some absolute constants $C, c > 0$.

By comparing the above rates with those of the empirical mean, we see that the dependence on δ has improved exponentially and furthermore, is decoupled from the dimension term, $\text{Tr}(\Sigma)$. In fact, the rate in Definition 5 is known to be unimprovable even if the data were indeed *Gaussian*. That such a rate is achievable in one dimension was independently discovered by several groups Nemirovsky and Yudin (1983); Jerrum et al. (1986); Alon et al. (1999). Surprisingly, the study of optimal achievable guarantees for heavy-tailed estimation was only initiated much more recently (Catoni (2012); Devroye et al. (2016)). This culminated in the construction of an (inefficient) estimator achieving the above optimal guarantee in the multivariate setting (Lugosi and Mendelson (2019)). Soon after, Hopkins (2020) devised the first computationally efficient estimator with the same guarantees.

We observe that the optimal rates in the *adversarial* model and the *heavy tailed* model coincide under the mapping $\varepsilon \leftrightarrow \frac{\log(1/\delta)}{n}$. In the regime of constant ε (corresponding to $\delta \sim e^{-\Theta(n)}$), Theorem 3 implies the surprising corollary that any optimal adversarially robust mean estimation algorithm automatically gives an optimal estimator in the heavy-tailed model.

Corollary 6 In the regime of constant ε and δ inverse exponential in n , any optimal algorithm for mean estimation under adversarial contamination is also optimal in the heavy-tailed setting.

This follows by instantiating Definition 2 and Theorem 3 with $\mathcal{Z} = \mathbb{R}^d$ and

$$G_{\mathcal{D}}(\varepsilon) = \left\{ \hat{\mu} \in \mathbb{R}^d : \|\hat{\mu} - \mu\| \leq C \left(\sqrt{\frac{\text{Tr}(\Sigma)}{n}} + \sqrt{\|\Sigma\|\varepsilon} \right) \right\}$$

For the remaining applications, we'll leave implicit that we consider ε to be a constant.

Covariance estimation. We now consider an application of [Theorem 3](#) to covariance estimation, another fundamental statistical estimation task. Here, we are given n i.i.d samples from a high-dimensional distribution with the goal of recovering the covariance matrix of the distribution in spectral norm. We instantiate this task with an $L_4 - L_2$ hypercontractivity assumption, standard for high-dimensional heavy-tailed covariance estimation [Mendelson and Zhivotovskiy \(2020\)](#); [Abdalla and Zhivotovskiy \(2024\)](#). We formally state this assumption below:

Definition 7 *We say that a zero mean random vector X with covariance Σ satisfies an $L_4 - L_2$ hypercontractivity assumption if:*

$$\forall \|v\| = 1 : (\mathbb{E}[\langle X, v \rangle^4])^{1/4} \leq C \sqrt{v^\top \Sigma v}$$

for an absolute constant $C > 0$.

Note that no further assumptions are made on the distribution of X beyond its 4th moments allowing for the existence of heavy-tailed outliers. For this setting, there exist *optimal* estimators [Abdalla and Zhivotovskiy \(2024\)](#) which when given n i.i.d samples, with ε fraction arbitrarily corrupted by an adversary, produce estimate, $\hat{\Sigma}$ satisfying with probability at least $1 - \delta$:

$$\|\hat{\Sigma} - \Sigma\| \leq C \left(\sqrt{\frac{\text{Tr}(\Sigma) + \|\Sigma\| \log(1/\delta)}{n}} + \sqrt{\|\Sigma\| \varepsilon} \right)$$

which is known to be optimal even for the heavy-tailed ($\varepsilon = 0$) and adversarial settings (δ set to a constant). Applying [Theorem 3](#) to the covariance estimation setting, we establish the surprising fact that any optimal algorithm for covariance estimation with adversarial contamination is optimal for heavy-tailed contaminations as well.

Corollary 8 *Any optimal algorithm for covariance estimation under [Definition 7](#) with adversarial contamination is also optimal in the heavy-tailed contamination model.*

Linear regression. Our final application is to linear regression. We present a simplified setting for illustrative purposes. However, our results are also applicable more generally. Here, we consider the setting where we are given i.i.d samples generated from a distribution \mathcal{D} with $(X, Y) \sim \mathcal{D}$ generated as follows:

$$Y = \langle X, w^* \rangle + \eta$$

where X has zero-mean, covariance I , and satisfies [Definition 7](#) and η is zero-mean, independent of X , and has variance bounded by 1. Our goal now is to recover the w^* ³. Here, again, there exist estimates [Lugosi and Mendelson \(2020\)](#), \hat{w} , satisfying with probability at least $1 - \delta$:

$$\|\hat{w} - w^*\| \leq C \left(\sqrt{\frac{d + \log(1/\delta)}{n}} + \sqrt{\varepsilon} \right)$$

which are again known to be optimal for both the adversarial (δ set to a constant) and heavy-tailed settings ($\varepsilon = 0$). Hence, [Theorem 3](#) again yields the following corollary:

Corollary 9 *Any optimal estimator for linear regression with adversarial contamination is also optimal in the heavy-tailed contamination model.*

3. There exist alternative objectives for linear regression such as prediction error. We restrict to the simpler objective of parameter recovery for the sake of exposition. However, our results are also applicable to these alternative objectives.

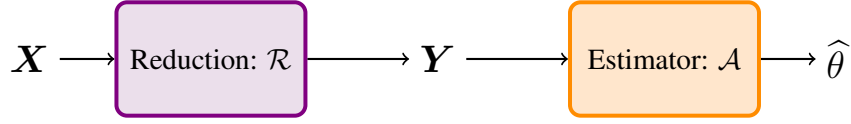


Figure 1: Illustration of the class of reductions [Theorem 12](#) applies to. The input dataset \mathbf{X} is processed by the reduction, \mathcal{R} to produce another dataset \mathbf{Y} that is ultimately used as input to the estimator \mathcal{A} which produces the final output $\hat{\theta}$.

1.1.2. THE ADVERSARIAL CONTAMINATION MODEL IS STRICTLY STRONGER.

[Theorem 3](#) prompts the natural question of whether the two models are equivalent? That is, are algorithms for the heavy-tailed setting automatically adversarially robust? We show in a strong sense that the answer to this question is **no**. The following result shows that, even for the simplest high-dimensional estimation problem of *mean estimation*, any reduction from adversarial to heavy-tailed estimation must either remove *most* outliers in the data or must ensure that their *empirical mean* is close to the true mean and furthermore, that their *variance* are small. That is, the reduction itself has to do a significant amount of heavy lifting.

Before we proceed, we define the class of black-box reductions our results apply to below:

Definition 10 For two statistical estimation problems, P and P' , we say that \mathcal{R} is a black box reduction from P' to P if for any estimator \mathcal{A} for P , $\mathcal{A}(\mathcal{R}(\cdot))$ is an estimator for P' .

An illustration of this definition is provided in [Fig. 1](#). Here, \mathbf{X} denotes a sample from the estimation problem P' . \mathbf{X} is then used by the reduction \mathcal{R} to produce another sample \mathbf{Y} which is input to the estimator \mathcal{A} for P to produce the final output $\hat{\theta}$. In our context, P and P' typically denote corresponding heavy-tailed and adversarial contamination estimation problems.

We pause for a few remarks on [Definition 10](#). Firstly, note that it captures the setting of [Theorem 3](#). Indeed, [Theorem 3](#) establishes that the *trivial* reduction which simply outputs its input, \mathbf{X} , suffices as a black-box reduction from heavy-tailed to adversarially robust estimation. On the other hand, [Definition 10](#) places no restrictions on *how* the intermediate output \mathbf{Y} is produced from \mathcal{R} . Nevertheless, our results establish strong structural properties on \mathbf{Y} that must be satisfied for *any* black-box reduction from adversarially-robust to heavy-tailed mean estimation. Lastly, observe that the requirements of a black-box reduction are agnostic to the specific estimation algorithms used to complete the reduction. Hence, they allow for a formal investigation into the relationship between different *corruption models*.

The exposition of our result additionally requires the concept of *stability*, a core component of recent adversarially robust estimation algorithms [Diakonikolas and Kane \(2019\)](#); [Diakonikolas et al. \(2020\)](#); [Hopkins et al. \(2020a\)](#):

Definition 11 A finite set $S = \{y_i\}_{i=1}^n \subset \mathbb{R}^d$, is (γ, ν) -stable with respect to $\mu \in \mathbb{R}^d$ for $\gamma > 0, \nu \in (0, 1/10)$ if there exists $S' \subset S$ with $|S'| \geq (1 - \nu)n$ such that:

$$\left\| \frac{1}{|S'|} \sum_{i \in S'} y_i - \mu \right\| \leq \gamma \text{ and } \left\| \frac{1}{|S'|} \sum_{i \in S'} (y_i - \mu)(y_i - \mu)^\top \right\| \leq \gamma^2.$$

The existence of such stable sets plays a key role in designing adversarially robust estimators. They may be shown to exist with high-probability for a chosen subset of the *inlier* data points. Hence, in the adversarial contamination setting, $\nu > \varepsilon$. In [Diakonikolas et al. \(2020\)](#), it is shown that any *stability-based* mean-estimation algorithm is simultaneously robust to adversarial and heavy-tailed corruptions. In fact, virtually all known adversarially robust estimators rely on such stability properties of the inlier data points and typically operate by iteratively pruning the dataset until one is left with a large stable subset of the data whose mean is a good estimate of the population mean. Conversely, the presence of (almost optimal) heavy-tailed estimators that do *not* rely on stability [Catoni and Giulini \(2017\)](#) is suggestive that heavy-tailed outliers are qualitatively weaker than adversarial outliers.

The main technical result of this section provides a formal justification for the separation between adversarial and heavy-tailed corruptions through the lens of black-box reductions. Concretely, we show that *any* black-box reduction from adversarial to heavy-tailed estimation must produce pointsets with arbitrarily large stable sets.

Theorem 12 (Informal – see [Theorem 17](#)) *Any black-box reduction ([Definition 10](#)) from adversarially robust mean estimation (with $\varepsilon = 0.1$) to heavy-tailed mean estimation (with $\log(1/\delta) = \Omega(n)$), must produce a pointset $\mathbf{Z} = \{z_i\}_{i=1}^m$ with $\mathcal{I} \subset [m]$ satisfying $|\mathcal{I}| \geq 0.99m$ and:*

$$\left\| \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} z_i - \mu \right\| \leq C \left(\sqrt{\frac{\text{Tr}(\Sigma)}{n}} + \sqrt{\|\Sigma\|} \right) \text{ and } \left\| \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} (z_i - \mu)(z_i - \mu)^\top \right\| \leq C \left(\frac{\text{Tr}(\Sigma)}{n} + \|\Sigma\| \right).$$

That is, this stable set must account for all but a small-constant fraction of the outliers.

We briefly remark on [Theorem 12](#) to simplify the interpretation of its implications. As previously stated, adversarially corrupted pointsets are only guaranteed to contain stable sets (with high probability) of size at most $0.9n^4$. Hence, the pointsets that are produced by *any* black-box adversarial-to-heavy-tailed reduction are qualitatively different from adversarially corrupted datasets. Furthermore, several algorithmic approaches for adversarially robust estimation rely on the extraction of a large stable subset of the input point set. Hence, the above result establishes that any such reduction must already produce a pointset, essentially all of which is a stable subset. Consequently, in the setting where the reduction passes *subsets* of the input to the heavy-tailed estimation algorithm, it must either *filter* out adversarial points or alternatively, ensure that all but a *negligible fraction* of them are part of a stable set. Hence, the reduction itself operates similarly to existing stability-based algorithms for adversarially robust estimation.

To establish [Theorem 12](#), we construct the novel heavy-tailed estimator that provides the guarantees of [Theorem 3](#). The design of our estimator leverages a novel tail-decay condition that is guaranteed to hold when the data is drawn from a distribution but does *not* hold when an adversary engineers a large fraction of the outlier set to deviate from the mean. The estimator then estimates a quantile *smoothed* notion of spread over all one-dimensional projections of the data and adds a random perturbation of this magnitude to the output of a standard heavy-tailed estimator. However, care must be taken in the choice of the underlying estimator. To avoid cancellations in magnitude to establish [Theorem 12](#) on the points provided to the algorithm as *input*, we use a standard stability-based framework for robust estimation [Steinhardt et al. \(2018\)](#). Unfortunately, these estimators are

4. For $\varepsilon = 0.1$, it is easy to construct adversarial contaminations such that all subsets larger than $0.9n$ are *not* stable.

not known to be optimal heavy-tailed estimators with prior analyses incurring an additional multiplicative logarithmic factor [Diakonikolas et al. \(2020\)](#). We provide a novel analysis of this estimator by providing a fine-grained analysis of a Gaussian rounding scheme previously used to analyze a median-of-means approach [Depersin and Lecué \(2019\)](#) and leveraging recent developments in the analysis of the trimmed median [Lugosi and Mendelson \(2021\)](#).

Notation. In this paper, \mathbb{R} and \mathbb{N} denote the set of real and natural numbers respectively. For $n \in \mathbb{N}$, $[n]$ denotes the set of integers from 1 through n . For $d, m, n \in \mathbb{N}$, \mathbb{R}^d and $\mathbb{R}^{m \times n}$ denote the set of d -dimensional vectors and $m \times n$ dimensional matrices respectively. For $v \in \mathbb{R}^d$ and $M \in \mathbb{R}^{d \times d}$, $\|v\|$ and $\|M\|$ are the standard Euclidean and spectral norms respectively while $\text{Tr}(M)$ denotes the trace of M . For measure \mathcal{D} and $n \in \mathbb{N}$, $\mathcal{D}^{\otimes n}$ denotes the n -time produce measure of \mathcal{D} . For a set $T \subset \mathbb{R}^d$, $\text{mean}(T)$ denotes the mean of the uniform distribution over the set and when $d = 1$, $\text{Med}(T)$ denotes its median. For $x \in \mathbb{R}$, $\text{sgn}(x)$ denotes the signum function of x . For distributions $\mathcal{D}_1, \mathcal{D}_2$, $\text{TV}(\mathcal{D}_1, \mathcal{D}_2)$ denotes the TV distance between them. For $v \in \mathbb{R}^d$, v_i denotes the i^{th} entry of v . For $\mathcal{S} \subset \mathbb{R}^d$, $\text{Unif}(\mathcal{S})$ denotes the uniform distribution over \mathcal{S} . Finally, for an event E , $\mathbf{1}\{E\}$ denotes the indicator of that event.

Organization. In the remainder of the paper, we establish [Theorem 15](#), the formalization of [Theorem 3](#). Subsequently, in [Section 3](#), we design and analyze the sub-Gaussian estimator used in the proof of [Theorem 12](#) whose generalized counterpart is [Theorem 17](#).

2. From heavy-tailed to adversarially robust X estimation

Here, we prove that an algorithm, robust to a constant factor of adversarial corruptions, with no additional assumptions required and no additional computation, automatically achieves sub-Gaussian guarantees. The argument reduces to a statement about isoperimetry in the Hamming metric. Recall that in the adversarial corruption setting, we require that our algorithm be robust to arbitrary εn sized corruptions of the sample with constant probability. That is, there is some set with constant mass such that, for any point within εn Hamming distance of this set, the algorithm outputs a “good” mean estimate. The question becomes, what is the measure of this εn Hamming ball when the data is drawn from a heavy-tailed distribution? Classical results on the connection between isoperimetry and concentration show that this is all but an exponentially small fraction.

For an abstract space, \mathcal{X} , consider the set of possible samples obtained from n iid draws from a distribution over \mathcal{X} , \mathcal{X}^n . For instance, in our setting $\mathcal{X} = \mathbb{R}^d$. For two possible sample sets, $x, y \in \mathcal{X}^n$ and $S \subset \mathcal{X}^n$, define the Hamming distance as follows:

$$\text{dist}_{\text{ham}}(x, y) := \sum_i \mathbf{1}[x_i - y_i \neq 0] \text{ and } \text{dist}_{\text{ham}}(x, S) = \min_{y \in S} \text{dist}_{\text{ham}}(x, y).$$

We recall the following fact about expansion with respect to dist_{ham} .

Lemma 13 ([Boucheron et al. \(2013\) Corollary 7.4](#)) *Let $S \subseteq \mathcal{X}^n$ and ν a measure on \mathcal{X} s.t. $\nu^{\otimes n}(S) = p$. We have:*

$$\forall t \geq 0 : \mathbb{P} \left\{ \text{dist}_{\text{ham}}(X, S) \geq t + \sqrt{\frac{n}{2} \log(1/p)} \right\} \leq e^{-2t^2/n}.$$

As a simple corollary, we obtain the following lower bound on the blowup of a set.

Corollary 14 *Let $S \subseteq \mathcal{X}^n$ and ν be a distribution over \mathcal{X} be such that:*

$$\nu^{\otimes n}(S) \geq 0.9.$$

Then,

$$\forall \varepsilon \geq 0 : \nu^{\otimes n}(S_\varepsilon) \geq 1 - 2e^{-\varepsilon^2 n} \text{ where } S_\varepsilon := \{x \in \mathcal{X}^n : \text{dist}_{\text{ham}}(x, S) \leq \varepsilon n\}.$$

Proof Set $t = \varepsilon n - \sqrt{\frac{n}{2} \log(1/p)}$ in [Lemma 13](#).

$$\nu^{\otimes n}(S_\varepsilon) = \mathbb{P}\{\text{dist}_{\text{ham}}(X, S) \geq \varepsilon n\} \leq \exp\left\{-\frac{2\left(\varepsilon n - \sqrt{\frac{n}{2} \log(1/p)}\right)^2}{n}\right\}$$

Let $\alpha > 0$ be a parameter to be determined subsequently. We now have two cases:

Case 1: $n \geq \frac{1}{\varepsilon^2 2(1-\alpha)^2} \log(1/p)$. This implies that $\sqrt{\frac{n}{2} \log(1/p)} \leq \varepsilon n(1-\alpha)$. Plugging in, we get

$$\mathbb{P}\{\text{dist}_{\text{ham}}(X, S) \geq \varepsilon n\} \leq e^{-2\varepsilon^2 \alpha^2 n}$$

Case 2: $n < \frac{1}{\varepsilon^2 2(1-\alpha)^2} \log(1/p)$. Then, by setting $\alpha = 1/\sqrt{2}$

$$2e^{-2\varepsilon^2 \alpha^2 n} \geq 2 \exp\left\{-\log(1/p) \frac{\alpha^2}{(1-\alpha)^2}\right\} = 2p^{\frac{\alpha^2}{(1-\alpha)^2}} \geq 1 \geq 1 - \nu^{\otimes n}(S_\varepsilon),$$

concluding the claim. ■

[Corollary 14](#) is precisely what we need to formalize our argument that the robustness of our algorithm ensures all but a vanishingly small fraction of the samples will be good.

Theorem 15 *Let \mathcal{A} be an adversarially robust algorithm with no internal randomness satisfying [Definition 2](#). Then, given samples $\mathbf{X} = \{x_i\}_{i=1}^n$ from a distribution, \mathcal{D} and corruption parameter $\varepsilon \in (0, 1/10)$, there exists a constant $c(\varepsilon) > 0$ s.t.*

$$\hat{\theta} \in G_{\mathcal{D}}(\varepsilon)$$

with probability $1 - 2e^{-c(\varepsilon)n}$, where $\hat{\theta} = \mathcal{A}(\mathbf{X}; \varepsilon)$. More generally, for $\varepsilon' \in [0, \varepsilon)$, \mathcal{A} outputs $\hat{\theta}$ satisfying the above guarantee with probability $1 - 2e^{-n(\varepsilon-\varepsilon')^2}$ when the sample it receives is ε' -corrupted.

Proof Let $A(\eta) \subseteq (\mathbb{R}^d)^n$ denote the set of samples on which \mathcal{A} satisfies η robustness. Let $S = \mathcal{A}(\varepsilon)$. We first observe that for any $\eta, \eta', \alpha > 0$ where $\alpha = \eta - \eta'$:

$$A(\eta)_\alpha \subseteq A(\eta').$$

That is, the α blowup of the set on which \mathcal{A} satisfies η robustness is contained in the set on which \mathcal{A} satisfies η' robustness. This follows from the triangle inequality for dist_{ham} .

Now, by assumption, $\mathcal{D}^{\otimes n}(S) \geq 0.9$. Let S_α denote its α blowup, instantiating $\alpha = \varepsilon - \varepsilon'$. By [Corollary 14](#), we know that $\mathcal{D}^{\otimes n}(S_\alpha) \geq 1 - 2e^{-\alpha^2 n}$. From our above observation, S_α is a subset

of the points on which \mathcal{A} is ε' robust; i.e. the samples on which, even with an adversary who can corrupt ε' of the samples, \mathcal{A} outputs $\hat{\theta}$ satisfying:

$$\theta \in G_{\mathcal{D}}(\varepsilon)$$

This observation along with the fact that $\mathcal{D}^{\otimes n}(S_\alpha) \geq 1 - 2e^{-\alpha^2 n}$ conclude the proof. \blacksquare

Remark *The reason we state our results in the constant error regime is that the dependence on ε in [Theorem 15](#) is quadratic. This can be seen to be necessary. That is, there exist distributions and adversarially robust algorithms for which the probability of success scales as $1 - \exp\{O(\varepsilon^2 n)\}$. We include a sketch of this proof in [Appendix D](#).*

However, a standard median-of-means preprocessing step, a staple in the robust statistics literature, reduces from the general case to the constant ε setting, overcoming this limitation.

3. A Hard sub-Gaussian Algorithm

In this section, we overview and present guarantees on a robust estimation algorithm that achieves sub-Gaussian rates while, at the same time, being vulnerable to *adversarial* perturbations. The formal details of the algorithm are presented in [Algorithm 2](#). [Algorithm 2](#) parameterizes two algorithms, one for each choice of s , both of which are guaranteed to have sub-Gaussian performance. The algorithm exploits a logarithmic tail-decay condition whose strength scales with the targeted failure probability. This condition is used to extract a notion of scale that remains bounded when the data is obtained from a heavy-tailed distribution but not when the data is adversarially corrupted.

The algorithm operates by first finding a point, x , such that there exists a large subset of points (a $1 - \varepsilon$ fraction) of points where $\varepsilon \approx \log(1/\delta)/n$ having bounded second-moment with respect to x . That is it solves the following program:

$$\arg \min_{x \in \mathbb{R}^d} \min_{w \in \mathcal{W}_\varepsilon} \left\| \sum_{i=1}^n w_i (x_i - x)(x_i - x)^\top \right\|$$

where $\forall \rho \geq 0 : \mathcal{W}_\rho := \left\{ \{w_i\}_{i=1}^n : 0 \leq w_i \leq \frac{1}{(1-\rho)n} \text{ and } \sum_{i=1}^n w_i = 1 \right\}$.

For the next step, define the following notion of variation, denoted by spread, for a one-dimensional point set, $\mathbf{Y} = \{y_i\}_{i=1}^n$ with respect to a centering y and $y_{(i)}$ denoting the ordered elements of \mathbf{Y} :

$$\forall i \in [n] : \tilde{w}_i = \exp\left(-C \cdot \frac{n\varepsilon}{\min(i, n-i)}\right) \text{ and } w_i := \frac{\tilde{w}_i}{\sum_{i=1}^n \tilde{w}_i}$$

$$\text{spread}(\mathbf{Y}, y) := \sqrt{\sum_{i \in [n]} w_i (y_{(i)} - y)^2}.$$

The algorithm then searches over directions for one with large spread. The final output of the algorithm is a random perturbation along the direction found in the previous step whose scale depends on the value of the variation along that direction and the failure probability. As we will see, this notion of spread forces the algorithm to be successful only when there exists a stable set of size

$(1 - c\varepsilon)n$ for any $c < 1$. The tail decay property of samples drawn from heavy-tailed distributions bounds the performance of the estimator in this setting but this may not necessarily hold for the adversarial contamination setting.

Algorithm 1 Comparison

```

1: Input: Vector  $v$ 
2: if  $v = 0$  then
3:   Return:  $v$ 
4: else
5:   Return:  $\text{sgn}(v_{i^*})v$  where  $i^* = \min\{i : v_i \neq 0\}$ 
6: end if
    
```

Algorithm 2 Sub-Gaussian Mean Estimator

```

1: Input: Point set  $\mathbf{X} = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ , Failure Probability  $\delta$ , Version  $s \in \{+1, -1\}$ 
2:  $\varepsilon \leftarrow \log(4/\delta)/n$ 
3:  $\tilde{\mu} \leftarrow \arg \min_{x \in \mathbb{R}^d} \min_{w \in \mathcal{W}_\varepsilon} \left\| \sum_{i=1}^n w_i(x_i - x)(x_i - x)^\top \right\|$ 
4:  $v \leftarrow \arg \max_{\|u\|=1} \text{spread}(\{\langle x_i, u \rangle\}_{i=1}^n, \langle \tilde{\mu}, u \rangle), \sigma_v \leftarrow \text{spread}(\{\langle x_i, v \rangle\}_{i=1}^n, \langle \tilde{\mu}, v \rangle)$ 
5:  $v \leftarrow \text{Comparison}(v)$ 
6:  $\hat{\mu} = \tilde{\mu} + s\sqrt{\varepsilon}\sigma_v v$ 
7: Return:  $\hat{\mu}$ 
    
```

The following pair of theorems now establish that the [Algorithm 2](#) achieves sub-Gaussian rates ([Theorem 16](#)) while remaining vulnerable to outliers ([Theorem 17](#)). In the interest of space, the proofs of [Theorem 16](#) and [Theorem 17](#) are deferred to [Appendix B](#) and [Appendix C](#) respectively.

Theorem 16 *There exists an absolute constants $C, c > 0$ such that the following hold. Let X_1, \dots, X_n be drawn iid from a distribution with mean μ and covariance Σ . Then, for any $\delta \in [e^{-cn}, 1/4]$, the outputs of [Algorithm 2](#), μ_+ and μ_- , on input X_1, \dots, X_n , δ , and $s = \{+1, -1\}$ respectively satisfy:*

$$\max \{\|\hat{\mu}_+ - \mu\|, \|\hat{\mu}_- - \mu\|\} \leq C \left(\sqrt{\frac{\text{Tr}(\Sigma)}{n}} + \sqrt{\frac{\|\Sigma\| \log(1/\delta)}{n}} \right)$$

with probability at least $1 - \delta$.

Theorem 17 *There exists an absolute constant $c' > 0$ and for any constant $c \in [0, 1]$, there exists an absolute constant $C > 0$ such that the following holds. Let $\mathbf{Z} = \{z_i\}_{i=1}^n \subset \mathbb{R}^d$ be a pointset, $\varepsilon \in [0, c']$ and $\eta \geq 0$. Suppose there exists $z \in \mathbb{R}^d$ such that the outputs of [Algorithm 2](#), $\{\hat{\mu}_+, \hat{\mu}_-\}$, when run with inputs \mathbf{Z} , $\delta = \exp(-\varepsilon n)$, and $s = \{+1, -1\}$ respectively satisfy:*

$$\max \{\|\hat{\mu}_+ - z\|, \|\hat{\mu}_- - z\|\} \leq \eta.$$

Then, there exists a subset $\mathcal{I} \subset [n]$ and $|\mathcal{I}| \geq 1 - c\eta n$ satisfying:

$$\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} (z_i - \mu)(z_i - \mu)^\top \preceq C \frac{\eta^2}{\varepsilon} \quad \text{and} \quad \|\mu - z\| \leq C\eta \quad \text{where} \quad \mu := \frac{1}{|\mathcal{I}|} \sum_{i=1}^n z_i.$$

Acknowledgments

The authors would like to thank Sam Hopkins for proposing the project as part of a course on Algorithmic Statistics.

References

- Pedro Abdalla and Nikita Zhivotovskiy. Covariance estimation: Optimal dimension-free guarantees for adversarial corruption and heavy tails. *Journal of the European Mathematical Society*, 2024.
- Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. volume 58, pages 137–147. 1999. doi: 10.1006/jcss.1997.1545. URL <https://doi.org/10.1006/jcss.1997.1545>. Twenty-eighth Annual ACM Symposium on the Theory of Computing (Philadelphia, PA, 1996).
- Robert B. Ash. *Information Theory*. Dover Publications, 1990. ISBN 9780486665214.
- Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. 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>. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.
- Olivier Bousquet. *Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms*. PhD thesis, Ecole Polytechnique, 2002.
- O. Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.*, 48(4):1148–1185, 2012.
- O. Catoni and I. Giulini. Dimension-free PAC-Bayesian bounds for the estimation of the mean of a random vector. *NIPS 2017 Workshop; (Almost) 50 Shades of Bayesian Learning: PAC-Bayesian Trends and Insights*, 2017.
- Yeshwanth Cherapanamjeri, Nilesh Tripuraneni, Peter L. Bartlett, and Michael I. Jordan. Optimal mean estimation without a variance. In Po-Ling Loh and Maxim Raginsky, editors, *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 356–357. PMLR, 2022. URL <https://proceedings.mlr.press/v178/cherapanamjeri22a.html>.
- Jules Depersin and Guillaume Lecué. Robust subgaussian estimation of a mean vector in nearly linear time, 2019.
- Luc Devroye, Matthieu Lerasle, Gábor Lugosi, and Roberto I. Oliveira. Sub-Gaussian mean estimators. *Ann. Statist.*, 44(6):2695–2725, 2016. ISSN 0090-5364. doi: 10.1214/16-AOS1440. URL <https://doi.org/10.1214/16-AOS1440>.
- Ilias Diakonikolas and Daniel M Kane. Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*, 2019.

- Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In [Dinur \(2016\)](#), pages 655–664. ISBN 978-1-5090-3933-3. doi: 10.1109/FOCS.2016.85. URL <https://doi.org/10.1109/FOCS.2016.85>.
- Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 999–1008. PMLR, 2017. URL <http://proceedings.mlr.press/v70/diakonikolas17a.html>.
- Ilias Diakonikolas, Daniel M. Kane, and Ankit Pensia. Outlier robust mean estimation with subgaussian rates via stability. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/13ec9935e17e00bed6ec8f06230e33a9-Abstract.html>.
- Ilias Diakonikolas, Daniel M. Kane, and Ankit Pensia. Outlier Robust Mean Estimation with Subgaussian Rates via Stability. (arXiv:2007.15618), March 2021.
- Irit Dinur, editor. *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, 2016. IEEE Computer Society. ISBN 978-1-5090-3933-3. URL <https://ieeexplore.ieee.org/xpl/conhome/7781469/proceeding>.
- Samuel B. Hopkins. Mean estimation with sub-Gaussian rates in polynomial time. *Ann. Statist.*, 48(2):1193–1213, 2020. ISSN 0090-5364. doi: 10.1214/19-AOS1843. URL <https://doi.org/10.1214/19-AOS1843>.
- Samuel B. Hopkins, Jerry Li, and Fred Zhang. Robust and heavy-tailed mean estimation made simple, via regret minimization. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020a. URL <https://proceedings.neurips.cc/paper/2020/hash/8a1276c25f5efe85f0fc4020fbf5b4f8-Abstract.html>.
- Samuel B. Hopkins, Jerry Li, and Fred Zhang. Robust and Heavy-Tailed Mean Estimation Made Simple, via Regret Minimization. *arXiv.org*, July 2020b.
- Peter J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35:73–101, 1964. ISSN 0003-4851. doi: 10.1214/aoms/1177703732. URL <https://doi.org/10.1214/aoms/1177703732>.
- Mark R. Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci.*, 43(2-3):169–188, 1986. ISSN 0304-3975. doi: 10.1016/0304-3975(86)90174-X. URL [https://doi.org/10.1016/0304-3975\(86\)90174-X](https://doi.org/10.1016/0304-3975(86)90174-X).

- Kevin A. Lai, Anup B. Rao, and Santosh S. Vempala. Agnostic estimation of mean and covariance. In [Dinur \(2016\)](#), pages 665–674. ISBN 978-1-5090-3933-3. doi: 10.1109/FOCS.2016.76. URL <https://doi.org/10.1109/FOCS.2016.76>.
- Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*. Classics in Mathematics. Springer-Verlag, Berlin, 2011. ISBN 978-3-642-20211-7. Isoperimetry and processes, Reprint of the 1991 edition.
- Gábor Lugosi and Shahar Mendelson. Sub-Gaussian estimators of the mean of a random vector. *Ann. Statist.*, 47(2):783–794, 2019. ISSN 0090-5364. doi: 10.1214/17-AOS1639. URL <https://doi.org/10.1214/17-AOS1639>.
- Gábor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *J. Eur. Math. Soc. (JEMS)*, 22(3):925–965, 2020. ISSN 1435-9855. doi: 10.4171/jems/937. URL <https://doi.org/10.4171/jems/937>.
- Gábor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: the optimality of trimmed mean. *Ann. Statist.*, 49(1):393–410, 2021. ISSN 0090-5364. doi: 10.1214/20-AOS1961. URL <https://doi.org/10.1214/20-AOS1961>.
- P. Massart. The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality. *The Annals of Probability*, 18(3):1269–1283, July 1990. ISSN 0091-1798, 2168-894X. doi: 10.1214/aop/1176990746.
- Colin McDiarmid. On the method of bounded differences. In *Surveys in combinatorics, 1989 (Norwich, 1989)*, volume 141 of *London Math. Soc. Lecture Note Ser.*, pages 148–188. Cambridge Univ. Press, Cambridge, 1989.
- Shahar Mendelson and Nikita Zhivotovskiy. Robust covariance estimation under $L_4 - L_2$ norm equivalence. *Ann. Statist.*, 48(3):1648–1664, 2020. ISSN 0090-5364. doi: 10.1214/19-AOS1862. URL <https://doi.org/10.1214/19-AOS1862>.
- Jaouad Mourtada, Tomas Vaškevičius, and Nikita Zhivotovskiy. Distribution-free robust linear regression. *Math. Stat. Learn.*, 4(3-4):253–292, 2021. ISSN 2520-2316. doi: 10.4171/msl/27. URL <https://doi.org/10.4171/msl/27>.
- Arkadi S. Nemirovsky and David B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons, Inc., New York, 1983. ISBN 0-471-10345-4. Translated from the Russian and with a preface by E. R. Dawson.
- Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. In Anna R. Karlin, editor, *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA*, volume 94 of *LIPIcs*, pages 45:1–45:21. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018. ISBN 978-3-95977-060-6. doi: 10.4230/LIPIcs.ITCS.2018.45. URL <https://doi.org/10.4230/LIPIcs.ITCS.2018.45>.
- Michel Talagrand. Sharper bounds for Gaussian and empirical processes. *Ann. Probab.*, 22(1):28–76, 1994. ISSN 0091-1798. URL [http://links.jstor.org/sici?sici=0091-1798\(199401\)22:1<28:SBFGAE>2.0.CO;2-W&origin=MSN](http://links.jstor.org/sici?sici=0091-1798(199401)22:1<28:SBFGAE>2.0.CO;2-W&origin=MSN).

Michel Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126(3):505–563, 1996. ISSN 0020-9910. doi: 10.1007/s002220050108. URL <https://doi.org/10.1007/s002220050108>.

Appendix A. Auxiliary Technical Results

A.1. Probability and Empirical Process Theory

We present Bousquet's inequality on the suprema of empirical processes [Bousquet \(2002\)](#) which builds on prior results by Talagrand [Talagrand \(1994, 1996\)](#).

Theorem 18 ([Bousquet \(2002\)](#); [Boucheron et al. \(2013\)](#)) *Let X_1, \dots, X_n be independent identically distributed random vectors indexed by an index set \mathcal{T} . Assume that $\mathbb{E}[X_{i,s}] = 0$, and $X_{i,s} \leq 1$ 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 $u \geq -1$. Then for all $\lambda \geq 0$,*

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

Also, for all $t \geq 0$,

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

We also recall McDiarmid's classical bounded differences inequality.

Theorem 19 ([McDiarmid \(1989\)](#); [Boucheron et al. \(2013\)](#)) *Let $n \in \mathbb{N}$, \mathcal{X} denote some domain and assume $f : \mathcal{X}^n \rightarrow \mathbb{R}$ satisfies:*

$$\forall i \in [n] : \sup_{\substack{x_1, \dots, x_n \\ x'_i \in \mathcal{X}}} |f(x_1, \dots, x_n) - f(x_1, \dots, x'_i, \dots, x_n)| \leq 1.$$

Then

$$\mathbb{P}\{f(X) - \mathbb{E}f(X) \geq t\} \leq e^{-2t^2/n}.$$

We also require the Ledoux-Talagrand contraction inequality [Ledoux and Talagrand \(2011\)](#) (as stated in [Boucheron et al. \(2013\)](#)).

Theorem 20 ([Ledoux and Talagrand \(2011\)](#); [Boucheron et al. \(2013\)](#)) *Let x_1, \dots, 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 $i = 1, \dots, n$, let $\phi_i : \mathbb{R} \rightarrow \mathbb{R}$ be a 1-Lipschitz function such that $\phi_i(0) = 0$. Let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher random variables, and let $\Psi : [0, \infty) \rightarrow \mathbb{R}$ be a non-decreasing convex function. Then,*

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

and

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

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

Corollary 21 *Assume the setting of [Theorem 20](#). Then,*

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

Theorem 22 (Massart (1990)) Let X_1, \dots, X_n be real-valued, iid random variables with CDF F . Let F_n denote the empirical distribution function:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}[X_i \leq x]$$

For any $\epsilon > 0$:

$$\mathbb{P} \left\{ \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| > \epsilon \right\} \leq 2e^{-2n\epsilon^2}$$

Theorem 23 (Ash (1990) Lemma 4.7.2) Let $S \sim B(n, p)$ be the sum of n independent p -biased coins. Then:

$$\mathbb{P} \{S \geq \lambda n\} \geq \frac{1}{\sqrt{8\lambda n(1-\lambda)}} e^{-nD(\lambda||p)}$$

where $D(\lambda||p) = \lambda \log \frac{\lambda}{p} + (1-\lambda) \log \frac{1-\lambda}{1-p}$

Corollary 24 Let $S \sim B(n, p)$ be the sum of n independent p -biased coins. Then:

$$\mathbb{P} \left\{ \frac{S}{n} \geq p + \epsilon \right\} \geq \frac{1}{\sqrt{2n}} e^{-\frac{n\epsilon^2}{p(1-p)}}$$

Proof We first observe that $\lambda(1-\lambda) \geq \frac{1}{4}$. Therefore,

$$\frac{1}{\sqrt{8\lambda n(1-\lambda)}} \geq \frac{1}{\sqrt{2n}}$$

Next, we show that $D(\lambda||p) \leq \epsilon^2$ where $\lambda = p + \epsilon$

$$\begin{aligned} \lambda \log \frac{\lambda}{p} + (1-\lambda) \log \frac{1-\lambda}{1-p} &\leq \lambda \left(\frac{\lambda}{p} - 1 \right) + (1-\lambda) \left(\frac{1-\lambda}{1-p} - 1 \right) \\ &= \frac{\lambda(\lambda-p)}{p} + \frac{(1-\lambda)(p-\lambda)}{1-p} \\ &= \frac{\lambda(\lambda-p)(1-p) + (1-\lambda)(p-\lambda)p}{p(1-p)} \\ &= \frac{\epsilon\lambda(1-p) - \epsilon p(1-\lambda)}{p(1-p)} \\ &= \frac{\epsilon^2}{p(1-p)} \end{aligned}$$

■

A.2. Tools from Robust Statistics

Lemma 25 For any $\rho \in [0, 1/2]$ and for any $w, w' \in \mathcal{W}_\rho$, we have:

$$\text{TV}(w, w') \leq 2\rho.$$

Proof Observe that:

$$\text{TV}(w, w') = \sum_{i=1}^n \max(w_i, w'_i) - 1 \leq \frac{1}{1-\rho} - 1 = \frac{\rho}{1-\rho} \leq 2\rho$$

and the lemma follows. ■

Lemma 26 Let $Y = \{y_1, \dots, y_n\} \subset \mathbb{R}$ and $\rho \in [0, 1/4]$. Now, let $w, w' \in \mathcal{W}_\rho$ be such that:

$$\sum_{i=1}^n w_i(y_i - \mu)^2 \leq \sigma^2, \quad \sum_{i=1}^n w_i(y_i - \mu')^2 \leq \sigma'^2 \text{ where } \mu = \sum_{i=1}^n w_i y_i, \quad \mu' = \sum_{i=1}^n w'_i y_i.$$

Then, we have:

$$|\mu - \mu'| \leq 2\sqrt{\rho}(\sigma + \sigma').$$

Proof Note that we may assume $w \neq w'$ as this case is trivial. Let \tilde{w} be the distribution be such that $\tilde{w}_i \propto \min(w_i, w'_i)$. Furthermore, let \hat{w} and \hat{w}' distributions such that $\hat{w}_i \propto w_i - \min(w_i, w'_i)$ and $\hat{w}'_i \propto w'_i - \min(w_i, w'_i)$. Note that \tilde{w} is well defined as $\text{TV}(w, w') \leq 1/2$ from [Lemma 25](#) and \hat{w}, \hat{w}' are well-defined from the fact that $\text{TV}(w, w') > 0$. Letting $\nu = \text{TV}(w, w')$:

$$w = (1 - \nu)\tilde{w} + \nu\hat{w}, \quad w' = (1 - \nu)\tilde{w} + \nu\hat{w}'.$$

Letting $\tilde{\mu}, \hat{\mu}, \hat{\mu}'$ be the means of \tilde{w}, \hat{w} , and \hat{w}' respectively, we have:

$$\begin{aligned} |\mu - \tilde{\mu}| &= \nu|\hat{\mu} - \tilde{\mu}| \\ |\mu' - \tilde{\mu}| &= \nu|\hat{\mu}' - \tilde{\mu}|. \end{aligned}$$

As a consequence, by the law of total variation:

$$\begin{aligned} \nu(\hat{\mu} - \mu)^2 + (1 - \nu)(\tilde{\mu} - \mu)^2 &= \nu(1 - \nu)(\tilde{\mu} - \hat{\mu})^2 \leq \sigma^2 \\ \nu(\hat{\mu}' - \mu')^2 + (1 - \nu)(\tilde{\mu} - \mu')^2 &= \nu(1 - \nu)(\tilde{\mu} - \hat{\mu}')^2 \leq \sigma'^2. \end{aligned}$$

Therefore, we get:

$$\begin{aligned} |\tilde{\mu} - \hat{\mu}| &\leq \frac{\sqrt{2}\sigma}{\sqrt{\nu}} \implies |\mu - \tilde{\mu}| \leq \sqrt{2\nu}\sigma \\ |\tilde{\mu} - \hat{\mu}'| &\leq \frac{\sqrt{2}\sigma'}{\sqrt{\nu}} \implies |\mu' - \tilde{\mu}| \leq \sqrt{2\nu}\sigma'. \end{aligned}$$

By the triangle inequality, we get:

$$|\mu - \mu'| \leq 2\sqrt{\rho}(\sigma + \sigma')$$

concluding the proof of the lemma. ■

Appendix B. Proof that Algorithm 2 achieves sub-Gaussian Rates - Theorem 16

In this section, we will show that Algorithm 2 achieves sub-Gaussian rates on heavy-tailed distribution. The main technical difficulty in this section is that the notion of spread used in Algorithm 2 remains suitably bounded. For this step, we prove a tail-decay condition on samples drawn from heavy-tailed distributions and show that this suffices to establish the required bounds. In addition, even *without* the additional spread detection step, it is not known whether the stability-based intermediate estimate ($\tilde{\mu}$ in Algorithm 2) achieves sub-Gaussian rates. Unfortunately, prior analyses of the estimator are sub-optimal by a logarithmic factor Hopkins et al. (2020b); Diakonikolas et al. (2021). We provide an improved statistical analysis of these estimators by proving tighter bounds on the truncated variances of the one-dimensional point sets obtained by projecting the dataset onto one-dimensional subspaces, and secondly, by improving the analysis of a Gaussian rounding scheme of Depersin and Lecué (2019). Our analysis of the performance of the intermediate estimate $\tilde{\mu}$ borrows heavily from the analysis of Lugosi and Mendelson (2021) and may be viewed as a generalization of their results to the setting of stability-based estimators, where the set of points to be truncated are chosen uniformly for every direction. The main theorem of the section is the following which shows that Algorithm 2 achieves sub-Gaussian rates.

Theorem 16 *There exists an absolute constants $C, c > 0$ such that the following hold. Let X_1, \dots, X_n be drawn iid from a distribution with mean μ and covariance Σ . Then, for any $\delta \in [e^{-cn}, 1/4]$, the outputs of Algorithm 2, μ_+ and μ_- , on input X_1, \dots, X_n , δ , and $s = \{+1, -1\}$ respectively satisfy:*

$$\max \{ \|\hat{\mu}_+ - \mu\|, \|\hat{\mu}_- - \mu\| \} \leq C \left(\sqrt{\frac{\text{Tr}(\Sigma)}{n}} + \sqrt{\frac{\|\Sigma\| \log(1/\delta)}{n}} \right)$$

with probability at least $1 - \delta$.

Throughout this section, we assume that X_1, \dots, X_n are drawn iid from a distribution with mean 0 and covariance Σ . Furthermore, noting that Algorithm 2 is shift and scale-equivariant, it suffices to consider the case where $\mu = 0$ and $\|\Sigma\| = 1$. We start by defining a truncation parameter and function below following the analysis of Lugosi and Mendelson (2021):

$$\tau := C_1 \max \left\{ \frac{1}{\varepsilon} \sqrt{\frac{\text{Tr}(\Sigma)}{n}}, \sqrt{\frac{\|\Sigma\|}{\varepsilon}} \right\} \text{ and } \forall \gamma \geq 0 : \phi_\gamma(x) = \begin{cases} x & \text{if } |x| \leq \gamma \\ \text{sgn}(x)\gamma & \text{otherwise} \end{cases}.$$

The next lemma shows that the second moment of the truncated one-dimensional pointsets obtained by projecting the dataset onto any dimension is small. This lemma is used to analyze both the performance of the stability-based intermediate estimate, $\tilde{\mu}$, and the scale of the variation, σ_v , utilized in the algorithm. Intuitively, the lemma is used to bound the variance of the points whose projections are *small* while the logarithmic tail bound established subsequently bounds the variance of the points with *large* projections. Its proof relies on ideas from Lugosi and Mendelson (2021).

Lemma 27 *We have:*

$$\forall \|v\| = 1 : \sum_{i=1}^n \phi_\tau(\langle X_i, v \rangle)^2 \leq C \cdot \left(\frac{\text{Tr}(\Sigma)}{\varepsilon} + n\|\Sigma\| \right).$$

with probability at least $1 - \exp(-\varepsilon n)$.

Proof We may assume without loss of generality that $\|\Sigma\| = 1$. First, observe that:

$$\forall \|v\| = 1 : \mathbb{E}_{X \sim \mathcal{D}}[\phi_\tau(\langle X_i, v \rangle)^2] \leq \mathbb{E}_{X \sim \mathcal{D}}[\langle X_i, v \rangle^2] \leq 1.$$

And, as a consequence, we get where X'_i and γ_i are independent copies of X_i and independent Rademacher random variables respectively:

$$\begin{aligned} \mathbb{E} \left[\max_{\|v\|=1} \sum_{i=1}^n \phi_\tau(\langle X_i, v \rangle)^2 \right] &\leq \mathbb{E} \left[\max_{\|v\|=1} \sum_{i=1}^n \phi_\tau(\langle X_i, v \rangle)^2 - \mathbb{E} [\phi_\tau(\langle X'_i, v \rangle)^2] \right] + \max_{\|v\|=1} n \mathbb{E} [\phi_\tau(\langle X, v \rangle)^2] \\ &\leq n + \mathbb{E} \left[\max_{\|v\|=1} \sum_{i=1}^n \phi_\tau(\langle X_i, v \rangle)^2 - \mathbb{E} [\phi_\tau(\langle X'_i, v \rangle)^2] \right] \\ &\leq n + \mathbb{E}_{X_i, X'_i} \left[\max_{\|v\|=1} \sum_{i=1}^n \phi_\tau(\langle X_i, v \rangle)^2 - \phi_\tau(\langle X'_i, v \rangle)^2 \right] \\ &\leq n + \mathbb{E}_{X_i, X'_i, \gamma_i} \left[\max_{\|v\|=1} \sum_{i=1}^n \gamma_i (\phi_\tau(\langle X_i, v \rangle)^2 - \phi_\tau(\langle X'_i, v \rangle)^2) \right] \\ &\leq n + 2 \mathbb{E}_{X_i, \gamma_i} \left[\max_{\|v\|=1} \sum_{i=1}^n \gamma_i \phi_\tau(\langle X_i, v \rangle)^2 \right] \\ &\leq n + 8 \tau \mathbb{E}_{X_i, \gamma_i} \left[\max_{\|v\|=1} \sum_{i=1}^n \gamma_i \langle X_i, v \rangle \right] \leq n + 8 \tau \mathbb{E}_{X_i, \gamma_i} \left[\left\| \sum_{i=1}^n \gamma_i X_i \right\| \right] \\ &\leq n + 8 \tau \sqrt{\mathbb{E}_{X_i, \gamma_i} \left[\left\| \sum_{i=1}^n \gamma_i X_i \right\|^2 \right]} = n + 8 \tau \sqrt{n \operatorname{Tr}(\Sigma)}, \end{aligned} \quad (2)$$

the sixth inequality following from Ledoux-Talagrand contraction ([Corollary 21](#)) and noting that ϕ_τ is 2τ -Lipschitz. To establish high-probability concentration, define:

$$Z = \max_{\|v\|=1} \sum_{i=1}^n W_{i,v} \text{ where } W_{i,v} := \phi_\tau(\langle X_i, v \rangle)^2 - \mathbb{E}_{X \sim \mathcal{D}}[\phi_\tau(\langle X, v \rangle)^2].$$

We get as a consequence of [Eq. \(2\)](#) that:

$$\mathbb{E}[Z] \leq 8 \tau \sqrt{n \operatorname{Tr}(\Sigma)}.$$

Furthermore, we get:

$$\forall \|v\| = 1 : \mathbb{E}[W_{i,v}^2] \leq \mathbb{E}[\phi_\tau(\langle X_i, v \rangle)^4] \leq \tau^2 \mathbb{E}[\phi_\tau(\langle X_i, v \rangle)^2] \leq \tau^2.$$

Finally, note that almost surely:

$$W_{i,v} \leq \tau^2.$$

We now obtain via [Theorem 18](#) by renormalizing $Y_{i,v} := W_{i,v}/\tau^2$:

$$\mathbb{P}\{Z \geq r\} \leq \exp \left(- \frac{r^2}{32(\tau^3 \sqrt{n \operatorname{Tr} \Sigma} + n\tau^2 + r\tau^2)} \right).$$

Now, by the following setting for r :

$$r := C \max \left\{ \frac{\text{Tr } \Sigma}{\varepsilon}, n \right\},$$

and the above inequality, we get that:

$$\mathbb{P} \{ Z \geq r \} \leq \exp(-\varepsilon n).$$

By observing that:

$$\max_v \sum_{i=1}^n \phi_\tau(\langle X_i, v \rangle)^2 \leq Z + n,$$

the lemma follows. ■

The next lemma establishes the logarithmic tail bound alluded to previously. We show that the tail of the empirical distribution over samples scales as $\varepsilon / \log(t)$ with probability at least $1 - e^{-n\varepsilon}$. This is then exploited to bound the scale of the spread estimate, σ_v , in [Algorithm 2](#).

Lemma 28 *We have:*

$$\forall \|v\| = 1, j \geq 1 : \sum_{i=1}^n \mathbf{1} \{ |\langle X_i, v \rangle| \geq \tau_j \} \leq \frac{2(n\varepsilon + j)}{3j} \text{ where } \tau_j = \frac{C_1}{64} \cdot e^{4j} \cdot \max \left\{ \frac{1}{\varepsilon} \sqrt{\frac{\text{Tr}(\Sigma)}{n}}, \frac{1}{\sqrt{\varepsilon}} \right\}$$

with probability at least $1 - e^{-\varepsilon n}$.

Proof First define:

$$\psi_j(x) := \begin{cases} 1 & \text{if } |x| \geq \tau_j \\ 0 & \text{if } |x| \leq \frac{\tau_j}{2} \\ \frac{2|x|}{\tau_j} - 1 & \text{otherwise} \end{cases}.$$

As in the proof of [Lemma 27](#), define random variable:

$$Z := \max_{v=1} \sum_{i=1}^n W_{i,v} \text{ where } W_{i,v} := \psi_j(\langle X_i, v \rangle) - \mathbb{E}_{X \sim \mathcal{D}}[\psi_j(\langle X, v \rangle)].$$

We get where X'_i and γ_i are independent copies of X_i and independent Rademacher random variables respectively:

$$\begin{aligned}
 \mathbb{E}[Z] &= \mathbb{E} \left[\max_{\|v\|=1} \sum_{i=1}^n \psi_j(\langle X_i, v \rangle) - \mathbb{E}_{X \sim \mathcal{D}} [\psi_j(\langle X, v \rangle)] \right] \\
 &\leq \mathbb{E}_{X_i, X'_i} \left[\max_{\|v\|=1} \sum_{i=1}^n \psi_j(\langle X_i, v \rangle) - \psi_j(\langle X'_i, v \rangle) \right] \\
 &\leq \mathbb{E}_{X_i, X'_i, \gamma_i} \left[\max_{\|v\|=1} \sum_{i=1}^n \gamma_i (\psi_j(\langle X_i, v \rangle) - \psi_j(\langle X'_i, v \rangle)) \right] \\
 &\leq 2 \mathbb{E}_{X_i, \gamma_i} \left[\max_{\|v\|=1} \sum_{i=1}^n \gamma_i \psi_j(\langle X_i, v \rangle) \right] \leq \frac{8}{\tau_j} \mathbb{E}_{X_i, \gamma_i} \left[\max_{\|v\|=1} \sum_{i=1}^n \gamma_i \langle X_i, v \rangle \right] \\
 &\leq \frac{8}{\tau_j} \mathbb{E}_{X_i, \gamma_i} \left[\left\| \sum_{i=1}^n \gamma_i X_i \right\| \right] \leq \frac{8}{\tau_j} \sqrt{\mathbb{E}_{X_i, \gamma_i} \left[\left\| \sum_{i=1}^n \gamma_i X_i \right\|^2 \right]} = \frac{8}{\tau_j} \sqrt{n \operatorname{Tr}(\Sigma)},
 \end{aligned}$$

the fourth inequality following from the Ledoux-Talagrand contraction inequality ([Corollary 21](#)) and noting that ψ_j is $2/\tau_j$ -Lipschitz. Furthermore, observe that from Chebyshev's inequality:

$$\forall \|v\| = 1 : \mathbb{E}[\psi_j(\langle X_i, v \rangle)^2] \leq \mathbb{E} \left[\mathbf{1} \left\{ |\langle X_i, v \rangle| \geq \frac{\tau_j}{2} \right\} \right] \leq \frac{4}{\tau_j^2}.$$

Hence, we get:

$$\mathbb{E}[Z] \leq \frac{8}{\tau_j} \sqrt{n \operatorname{Tr}(\Sigma)} \text{ and } \forall \|v\| = 1 : \mathbb{E}[\psi_j(\langle X_i, v \rangle)^2] \leq \frac{4}{\tau_j^2}.$$

Defining

$$\nu := 32 \max \left\{ \frac{\sqrt{n \operatorname{Tr} \Sigma}}{\tau_j}, \frac{n}{\tau_j^2} \right\} \text{ and } r_j := \frac{n\varepsilon + j}{2j},$$

we get from [Theorem 18](#)

$$\mathbb{P} \{ Z \geq r_j \} \leq \exp \left\{ -\nu h \left(\frac{r_j}{\nu} \right) \right\} = \exp \left\{ - \left((\nu + r_j) \log \left(1 + \frac{r_j}{\nu} \right) - r_j \right) \right\}.$$

Now, we get:

$$\begin{aligned}
 (\nu + r_j) \log \left(1 + \frac{r_j}{\nu} \right) - r_j &\geq r_j \log \left(\frac{r_j}{\nu} \right) - r_j \geq r_j \log \left(\frac{Ce^{4j}}{32n\varepsilon} \cdot r_j \right) - r_j \geq r_j \log \left(\frac{Ce^{4j-1}}{32n\varepsilon} \cdot r_j \right) \\
 &= r_j(2j) + r_j \log \left(\frac{Ce^{2j-1}}{32n\varepsilon} \cdot r_j \right) \geq n\varepsilon + j + r_j \log \left(\frac{Ce^{2j-1}}{64j} \right) \geq n\varepsilon + j.
 \end{aligned}$$

Therefore, we get:

$$\forall j \geq 1 : \mathbb{P} \{ Z_j \geq r_j \} \leq \exp \{ -n\varepsilon + j \}.$$

A union bound now yields:

$$\mathbb{P} \{ \exists j \geq 1 : Z_j \geq r_j \} \leq e^{-n\varepsilon}.$$

Observing that:

$$\forall \|v\| = 1 : \mathbb{E}_{X \sim \mathcal{D}} [\psi_j(\langle X, v \rangle)] \leq \frac{\varepsilon}{Ce^{8j}}$$

the conclusion of the lemma follows. \blacksquare

The next lemma shows that the value of the spread estimate σ_v in [Algorithm 2](#) is small. It utilizes the results of [Lemma 27](#) to bound the contribution of points with *small* projection along v while [Lemma 28](#) bounds the contribution of the large points as the guarantees of [Lemma 28](#) are sub-optimal when applied to the smaller projections.

Lemma 29 *We have:*

$$\forall \|v\| = 1 : \text{spread}(\{\langle X_i, v \rangle\}_{i=1}^n, 0) \leq C \left(\sqrt{\frac{\text{Tr}(\Sigma)}{\varepsilon n}} + \sqrt{\|\Sigma\|} \right)$$

with probability at least $1 - 2e^{-n\varepsilon}$.

Proof As before, it suffices to restrict to the setting $\|\Sigma\| = 1$. First, condition on the events in the conclusions of [Lemmas 27](#) and [28](#). Now, fix a particular v with $\|v\| = 1$ and define \mathbf{Y} to be the ordered set of elements $\{\langle X_i, v \rangle\}_{i=1}^n$. Then, we have:

$$\forall \frac{n}{4} \leq i \leq \frac{3n}{4} : \tilde{w}_i \geq c \implies \sum_{i=1}^n \tilde{w}_i \geq cn \implies \forall j \in [n] : w_j \leq \frac{1}{cn}.$$

We will now bound the spread separately for the subsets $\mathcal{G} = \mathbf{Y} \cap [-\tau, \tau]$ and $\mathcal{B} = \mathbf{Y} \setminus \mathcal{G}$. For the first set, we have by [Lemma 27](#):

$$\sum_{i \in \mathcal{G}} w_i y_i^2 \leq \frac{1}{cn} \sum_{i \in \mathcal{G}} y_i^2 \leq C \cdot \left(\frac{\text{Tr}(\Sigma)}{\varepsilon} + 1 \right).$$

For the second, we further decompose and bound the sum as follows with $\mathcal{B}_j = \mathbf{Y} \cap (-\tau_{j+1}, -\tau_j] \cup [\tau_j, \tau_{j+1})$ and noting that $\mathcal{B} \cup_{j=1}^{4n\varepsilon} \mathcal{B}_j$ as $\mathcal{B}_j = \emptyset$ for $j < 4n\varepsilon$ by [Lemma 28](#):

$$\begin{aligned} \sum_{y \in \mathcal{B}} w_y y^2 &\leq \sum_{j=1}^{4n\varepsilon} \sum_{y \in \mathcal{B}_j} w_y y^2 \leq \frac{1}{cn} \sum_{j=1}^{4n\varepsilon} \sum_{y \in \mathcal{B}_j} \tilde{w}_y \tau_{j+1}^2 = \frac{C}{n} \sum_{j=1}^{4n\varepsilon} \sum_{y \in \mathcal{B}_j} \exp\left(-C \cdot \frac{3nj\varepsilon}{2(n\varepsilon + j)}\right) \tau_{j+1}^2 \\ &\leq \frac{C}{n} \sum_{j=1}^{4n\varepsilon} \frac{2(n\varepsilon + j)}{3j} \cdot \exp(-Cj) \tau_j^2 \leq \frac{C}{n} \sum_{j=1}^{4n\varepsilon} \frac{n\varepsilon + j}{j} \cdot \exp(-Cj) \cdot e^{8j} \left(\frac{1}{\varepsilon^2} \cdot \frac{\text{Tr}(\Sigma)}{n} + \frac{1}{\varepsilon} \right) \\ &\leq \frac{C}{n} \sum_{j=1}^{4n\varepsilon} (n\varepsilon + j) \cdot \exp(-j) \cdot \left(\frac{1}{\varepsilon^2} \cdot \frac{\text{Tr}(\Sigma)}{n} + \frac{1}{\varepsilon} \right) \leq \frac{C}{n} \left(\sum_{j=1}^{4n\varepsilon} (n\varepsilon + j) e^{-j} \right) \cdot \left(\frac{1}{\varepsilon^2} \cdot \frac{\text{Tr}(\Sigma)}{n} + \frac{1}{\varepsilon} \right) \\ &\leq \frac{C}{n} \cdot n\varepsilon \cdot \left(\frac{1}{\varepsilon^2} \cdot \frac{\text{Tr}(\Sigma)}{n} + \frac{1}{\varepsilon} \right) \leq C \left(\frac{\text{Tr}(\Sigma)}{\varepsilon n} + 1 \right) \end{aligned}$$

which concludes the proof of the lemma. \blacksquare

The next lemma is a refinement of a Gaussian rounding scheme by [Depersin and Lecué \(2019\)](#). These improvements along with [Lemma 27](#) allow us to establish that the stability-based estimators achieve sub-Gaussian recovery guarantees. On the other hand, in prior work, the guarantees incurred an additional multiplicative logarithmic factor. Before we proceed, we define the convex program analyzed in the proof:

$$\min_{w \in \mathcal{W}_\rho} \left\| \sum_{i=1}^n w_i z_i z_i^\top \right\|. \quad (\text{ROB-SDP})$$

For a dataset \mathbf{Z} and ρ , let $\text{ROB-SDP}(\mathbf{Z}, \rho)$ refer to the convex program above. The next lemma establishes a vectorization of the above program.

Lemma 30 *Let $\mathbf{Z} = \{z_1, \dots, z_n\} \subset \mathbb{R}^d$ and $\rho \in [0, 1/2]$. Then:*

$$\min_{w \in \mathcal{W}_\rho} \left\| \sum_{i=1}^n w_i z_i z_i^\top \right\| \leq 1024 \max_{\|v\|=1} \min_{w \in \mathcal{W}_{\rho/4}} \sum_{i=1}^n w_i \langle v, z_i \rangle^2.$$

Proof Note that [ROB-SDP](#) may be recast as the following min-max program.

$$\min_{w \in \mathcal{W}_\rho} \max_{\substack{M \succeq 0 \\ \text{Tr}(M)=1}} \left\langle M, \sum_{i=1}^n w_i z_i z_i^\top \right\rangle = \max_{\substack{M \succeq 0 \\ \text{Tr}(M)=1}} \min_{w \in \mathcal{W}_\rho} \left\langle M, \sum_{i=1}^n w_i z_i z_i^\top \right\rangle$$

where the exchange of the min and max follows from von Neumann's minimax theorem. Let M^* and m^* denote the optimal solution and value to the program on the right. Consider now a Gaussian random vector g drawn with mean 0 and covariance M . Note that for any $i \in [n]$, $\langle z_i, g \rangle$ is a Gaussian with mean 0 and variance $z_i^\top M z_i$. Therefore, we have:

$$\forall i \in [n] : \mathbb{P} \left\{ |\langle z_i, g \rangle| \geq \frac{\sqrt{z_i^\top M z_i}}{4} \right\} \geq \frac{3}{4}$$

from the fact that the pdf of a standard Gaussian random variable is bounded above by $1/\sqrt{2\pi}$. Furthermore, we have from the fact that $\|M\| \leq 1$ that:

$$\mathbb{P} \{ \|g\| \leq 4 \} \geq \frac{9}{10}.$$

Hence, we get by the union bound that:

$$\mathbb{P} \left\{ \|g\| \leq 4 \text{ and } |\langle z_i, g \rangle| \geq \frac{\sqrt{z_i^\top M z_i}}{4} \right\} \geq \frac{1}{2}. \quad (3)$$

For the rest of the proof, we divide into two cases based on the set:

$$\mathcal{H} := \left\{ i : z_i^\top M z_i \geq \frac{m^*}{4\rho} \right\}.$$

The two cases we consider are as follows:

Case 1: $|\mathcal{H}| \geq \rho n$. In this case, consider g which satisfies:

$$\sum_{i \in \mathcal{H}} \mathbf{1} \left\{ |\langle z_i, g \rangle| \geq \frac{1}{8} \sqrt{\frac{m^*}{\rho}} \text{ and } \|g\| \leq 4 \right\} \geq \frac{\rho n}{2}.$$

Such a g exists from Eq. (3). On this event, we have for the vector $\tilde{g} = g/\|g\|$:

$$\min_{w \in \mathcal{W}_{\rho/4}} \sum_{i=1}^n w_i \langle z_i, \tilde{g} \rangle^2 \geq \sum_{i \in \mathcal{H}} w_i \langle z_i, \tilde{g} \rangle^2 \geq \frac{\rho}{4} \cdot \frac{1}{16} \cdot \frac{m^*}{64\rho} \geq \frac{m^*}{1024}.$$

This concludes the proof of the lemma in this case.

Case 2: $|\mathcal{H}| < \rho n$. Then, we must have by the minimax formulation:

$$\frac{1}{n - |\mathcal{H}|} \sum_{i \notin \mathcal{H}} z_i^\top M z_i \geq m^*$$

as this is a feasible solution to the minimax optimization problem. Now, consider the event:

$$\frac{1}{n - |\mathcal{H}|} \sum_{i \notin \mathcal{H}} z_i^\top M z_i \cdot \mathbf{1} \left\{ |\langle z_i, g \rangle| \geq \frac{\sqrt{z_i^\top M z_i}}{4} \text{ and } \|g\| \leq 4 \right\} \geq \frac{m^*}{2}.$$

Note that such a g exists by the probabilistic method and Eq. (3). Furthermore, we have for any $\mathcal{G} \subseteq [n] \setminus \mathcal{H}$ with $|\mathcal{G}| \leq \frac{\rho n}{4}$ from the definition of \mathcal{H} :

$$\frac{1}{n - |\mathcal{H}|} \sum_{i \in \mathcal{G}} z_i^\top M z_i \leq \frac{2}{n} \cdot \frac{\rho n}{4} \cdot \frac{m^*}{4\rho} = \frac{m^*}{8}.$$

Hence, we get for any subset $\mathcal{I} \subset [n] \setminus \mathcal{H}$ with $|\mathcal{I}| \geq n - |\mathcal{H}| - \rho n/4$:

$$\frac{1}{n} \sum_{i \in \mathcal{I}} z_i^\top M z_i \mathbf{1} \left\{ |\langle z_i, g \rangle| \geq \frac{\sqrt{z_i^\top M z_i}}{4} \right\} \geq \frac{m^*}{4}.$$

Noticing that for $\tilde{g} = g/\|g\|$:

$$z_i^\top M z_i \mathbf{1} \left\{ |\langle z_i, g \rangle| \geq \frac{\sqrt{z_i^\top M z_i}}{4} \right\} \leq 256 \langle z_i, \tilde{g} \rangle^2 \mathbf{1} \left\{ |\langle z_i, g \rangle| \geq \frac{\sqrt{z_i^\top M z_i}}{4} \right\},$$

we get that for all such \mathcal{I} :

$$\frac{1}{n} \sum_{i \in \mathcal{I}} \langle z_i, \tilde{g} \rangle^2 \mathbf{1} \left\{ |\langle z_i, g \rangle| \geq \frac{\sqrt{z_i^\top M z_i}}{4} \right\} \geq \frac{m^*}{1024}.$$

We get as a consequence

$$\min_{w \in \mathcal{W}_{\rho/4}} \sum_{i=1}^n w_i \langle z_i, \tilde{g} \rangle^2 \geq \min_{\substack{\mathcal{I} \subset [n] \setminus \mathcal{H} \\ |\mathcal{I}| \geq n - |\mathcal{H}| - \rho n/4}} \frac{1}{n} \sum_{i \in \mathcal{I}} \langle z_i, \tilde{g} \rangle^2 \geq \frac{m^*}{1024}$$

which concludes this case and the proof of the lemma. ■

The final technical result required to prove [Theorem 16](#) shows that the mean of the truncated data points when projected onto any direction is close to the true mean of the distribution. The proof is identical to that in [Lugosi and Mendelson \(2021\)](#) and is included for completion.

Lemma 31 *We have:*

$$\forall \|v\| = 1 : \sum_{i=1}^n \phi_\tau(\langle X_i, v \rangle) \leq C \cdot \left(\sqrt{n \operatorname{Tr}(\Sigma)} + n\sqrt{\varepsilon} \right)$$

with probability at least $1 - \exp(-\varepsilon n)$.

Proof We proceed as follows. Consider the random variable:

$$Z := \max_{\|v\|=1} W_{i,v} \text{ where } W_{i,v} := \sum_{i=1}^n \phi_\tau(\langle X_i, v \rangle) - \mathbb{E}_{X \sim \mathcal{D}}[\phi_\tau(\langle X, v \rangle)].$$

We have where X'_i and γ_i are independent copies of X_i and independent Rademacher random variables respectively:

$$\begin{aligned} \mathbb{E}[Z] &= \mathbb{E} \left[\max_{\|v\|=1} \sum_{i=1}^n \phi_\tau(\langle X_i, v \rangle) - \mathbb{E}[\phi_\tau(\langle X'_i, v \rangle)] \right] \\ &\leq \mathbb{E}_{X_i, X'_i} \left[\max_{\|v\|=1} \sum_{i=1}^n \phi_\tau(\langle X_i, v \rangle) - \phi_\tau(\langle X_i, v \rangle) \right] \\ &= \mathbb{E}_{X_i, X'_i, \gamma_i} \left[\max_{\|v\|=1} \sum_{i=1}^n \gamma_i (\phi_\tau(\langle X_i, v \rangle) - \phi_\tau(\langle X_i, v \rangle)) \right] \\ &\leq 2 \mathbb{E}_{X_i, \gamma_i} \left[\max_{\|v\|=1} \sum_{i=1}^n \gamma_i \phi_\tau(\langle X_i, v \rangle) \right] \\ &\leq 4 \mathbb{E}_{X_i, \gamma_i} \left[\max_{\|v\|=1} \sum_{i=1}^n \gamma_i \langle X_i, v \rangle \right] \leq 4 \mathbb{E}_{X_i, \gamma_i} \left[\left\| \sum_{i=1}^n \gamma_i X_i \right\| \right] \\ &\leq 4 \sqrt{\mathbb{E}_{X_i, \gamma_i} \left[\left\| \sum_{i=1}^n \gamma_i X_i \right\|^2 \right]} = 4 \sqrt{n \operatorname{Tr}(\Sigma)}, \end{aligned}$$

the third inequality following from Ledoux-Talagrand contraction ([Corollary 21](#)). Furthermore, note that for all $\|v\| = 1$:

$$\mathbb{E}[W_{i,v}^2] \leq \mathbb{E}[\phi_\tau(\langle X_i, v \rangle)^2] \leq \mathbb{E}[\langle X_i, v \rangle^2] \leq 1$$

and that $W_{i,v} \leq 2\tau$ almost surely. Therefore, we get from [Theorem 18](#) that:

$$\mathbb{P}\{Z \geq r\} \leq \exp \left(-\frac{r^2}{64} \cdot \frac{1}{\tau \sqrt{n \operatorname{Tr}(\Sigma)} + n + r\tau} \right).$$

By picking $r = Cn\varepsilon\tau$, we get that:

$$\mathbb{P}\{Z \geq r\} \leq \exp(-n\varepsilon).$$

By observing that for all $\|v\| = 1$ and $X \sim \mathcal{D}$,

$$\begin{aligned} |\mathbb{E}[\phi_\tau(\langle X, v \rangle)]| &= |\mathbb{E}[\langle X, v \rangle] - \mathbb{E}[(\langle X, v \rangle - \text{sgn}(\langle X, v \rangle)\tau)\mathbf{1}\{|\langle X, v \rangle| \geq \tau}]| \\ &= |\mathbb{E}[(\langle X, v \rangle - \text{sgn}(\langle X, v \rangle)\tau)\mathbf{1}\{|\langle X, v \rangle| \geq \tau}]| \\ &\leq |\mathbb{E}[\langle X, v \rangle\mathbf{1}\{|\langle X, v \rangle| \geq \tau}]| + \tau\mathbb{P}\{|\langle X, v \rangle| \geq \tau\} \\ &\leq \sqrt{\mathbb{E}[\langle X, v \rangle^2]}\sqrt{\mathbb{P}\{|\langle X, v \rangle| \geq \tau\}} + \tau \cdot \frac{1}{\tau^2} \leq \frac{2}{\tau} \leq C\sqrt{\varepsilon} \end{aligned}$$

the lemma follows. ■

We are now ready to prove [Theorem 16](#). We will condition on the conclusions of [Lemmas 27](#) to [29](#) and [31](#). Note that this happens with probability at least $1 - \delta$. Observe that we have from [Lemma 28](#):

$$\forall \|v\| = 1 : \sum_{i=1}^n \mathbf{1}\{|\langle X_i, v \rangle| \geq \tau\} \leq \frac{\varepsilon n}{4}.$$

As a consequence, we get from [Lemma 31](#) for any $\|v\| = 1$:

$$\begin{aligned} \left| \sum_{i=1}^n \langle X_i, v \rangle \mathbf{1}\{|\langle X_i, v \rangle| \leq \tau\} \right| &\leq C \left(\sqrt{n \text{Tr}(\Sigma)} + n\sqrt{\varepsilon} \right) + \tau \sum_{i=1}^n \mathbf{1}\{|\langle X_i, v \rangle| \geq \tau\} \\ &\leq C \cdot \left(\sqrt{n \text{Tr}(\Sigma)} + n\sqrt{\varepsilon} \right). \end{aligned}$$

Furthermore, we have from [Lemma 27](#) that:

$$\sum_{i=1}^n \langle X_i, v \rangle^2 \mathbf{1}\{|\langle X_i, v \rangle| \leq \tau\} \leq C \left(\frac{\text{Tr}(\Sigma)}{\varepsilon} + n \right).$$

As a consequence of the above, [Lemma 30](#) yields for $x = \mathbf{0}$:

$$\min_{w \in \mathcal{W}_\varepsilon} \left\| \sum_{i=1}^n w_i (X_i - x)(X_i - x)^\top \right\| \leq C \left(\frac{\text{Tr}(\Sigma)}{\varepsilon n} + 1 \right).$$

Therefore, $\tilde{\mu}$ satisfies:

$$\min_{w \in \mathcal{W}_\varepsilon} \left\| \sum_{i=1}^n w_i (X_i - \tilde{\mu})(X_i - \tilde{\mu})^\top \right\| \leq C \left(\frac{\text{Tr}(\Sigma)}{\varepsilon n} + 1 \right).$$

Let w^* be the optimal solution to the above program and notice that $\tilde{\mu} = \sum_{i=1}^n w_i^* X_i$. Furthermore, let w' be the uniform distribution over the set $\{i : |\langle X_i, v \rangle| \leq \tau\}$, μ' denote its mean and observe $w' \in \mathcal{W}_\tau$. Then, we have by [Lemma 26](#) that:

$$|\langle \tilde{\mu}, v \rangle - \mu'| \leq C \left(\sqrt{\frac{\text{Tr}(\Sigma)}{n}} + \sqrt{\varepsilon} \right).$$

As a consequence, since the above holds for all $\|v\| = 1$, we have:

$$\|\tilde{\mu} - \mu\| = \|\tilde{\mu}\| \leq C \cdot \left(\sqrt{\frac{\text{Tr}(\Sigma)}{n}} + \sqrt{\varepsilon} \right).$$

Now, consider any u such that $\|u\| = 1$. We have

$$\begin{aligned} \text{spread}(\{\langle X_i, u \rangle\}_{i=1}^n, \langle \tilde{\mu}, u \rangle) &= \sqrt{\sum_{i=1}^n w_i (\langle X_i, u \rangle - \langle \tilde{\mu}, u \rangle)^2} \leq \sqrt{2 \sum_{i=1}^n w_i (\langle X_i, u \rangle^2 + \langle \tilde{\mu}, u \rangle^2)} \\ &\leq \sqrt{2} \cdot (\text{spread}(\{\langle X_i, u \rangle\}_{i=1}^n, \langle \tilde{\mu}, u \rangle) + \|\tilde{\mu}\|) \leq C \cdot \left(\sqrt{\frac{\text{Tr}(\Sigma)}{\varepsilon n}} + 1 \right). \end{aligned}$$

Therefore, we get:

$$\|\hat{\mu}\| \leq \|\tilde{\mu}\| + \sqrt{\varepsilon} \cdot \text{spread}(\{\langle X_i, u \rangle\}_{i=1}^n, \langle \tilde{\mu}, u \rangle) \leq C \left(\sqrt{\frac{\text{Tr}(\Sigma)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right)$$

thus establishing the theorem. ■

Appendix C. When is [Algorithm 2](#) adversarially robust? - [Theorem 17](#)

In this section, we establish the following result which proves that for [Algorithm 2](#) to be adversarially robust, the set of points provided as input is required to have a stable set that includes an arbitrarily large fraction of the input points. Informally, an overwhelming fraction of outliers must either be filtered out, or their inclusion in the set of good points used to compute the mean does not substantially change the value of the estimate.

Theorem 17 *There exists an absolute constant $c' > 0$ and for any constant $c \in [0, 1]$, there exists an absolute constant $C > 0$ such that the following holds. Let $\mathbf{Z} = \{z_i\}_{i=1}^n \subset \mathbb{R}^d$ be a pointset, $\varepsilon \in [0, c']$ and $\eta \geq 0$. Suppose there exists $z \in \mathbb{R}^d$ such that the outputs of [Algorithm 2](#), $\{\hat{\mu}_+, \hat{\mu}_-\}$, when run with inputs \mathbf{Z} , $\delta = \exp(-\varepsilon n)$, and $s = \{+1, -1\}$ respectively satisfy:*

$$\max \{\|\hat{\mu}_+ - z\|, \|\hat{\mu}_- - z\|\} \leq \eta.$$

Then, there exists a subset $\mathcal{I} \subset [n]$ and $|\mathcal{I}| \geq 1 - c\eta n$ satisfying:

$$\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} (z_i - \mu)(z_i - \mu)^\top \preceq C \frac{\eta^2}{\varepsilon} \quad \text{and} \quad \|\mu - z\| \leq C\eta \quad \text{where} \quad \mu := \frac{1}{|\mathcal{I}|} \sum_{i=1}^n z_i.$$

Proof Noting that $\hat{\mu}_+ = \tilde{\mu} + \sqrt{\varepsilon} \sigma_v v$ and $\hat{\mu}_- = \tilde{\mu} - \sqrt{\varepsilon} \sigma_v v$, we must have $\|\tilde{\mu} - z\| \leq \eta$ by convexity. Now, we have by the parallelogram law:

$$2\eta^2 \geq \|\hat{\mu}_+ - z\|^2 + \|\hat{\mu}_- - z\|^2 = 2(\|\tilde{\mu} - z\|^2 + \varepsilon \sigma_v^2) \implies \sigma_v^2 \leq \frac{\eta^2}{\varepsilon}.$$

Notice that in the computation of spread, for any $c_1\varepsilon n/16 \leq i \leq n - c_1\varepsilon n/16$, we have $c \leq \tilde{w}_i \leq 1$ and as a consequence, we have:

$$cn \leq \sum_{i=1}^n \tilde{w}_i \leq n \implies \forall \frac{c_1\varepsilon n}{16} \leq i \leq n - \frac{c_1\varepsilon n}{16} : \frac{c}{n} \leq w_i \leq \frac{1}{cn}.$$

As a consequence, we have for $\|v\| = 1$:

$$\min_{w \in \mathcal{W}_{c_1\varepsilon/16}} \sum_{i=1}^n w_i \langle v, z_i - \tilde{\mu} \rangle^2 \leq C \cdot \frac{\eta^2}{\varepsilon}.$$

From [Lemma 30](#), we also get that:

$$\min_{w \in \mathcal{W}_{c_1\varepsilon/4}} \sum_{i=1}^n w_i (z_i - \tilde{\mu})(z_i - \tilde{\mu})^\top \preceq C \frac{\eta^2}{\varepsilon}.$$

Let w^* be the solution to the above problem. Now, consider the set:

$$\mathcal{I} := \left\{ i : w_i^* \geq \frac{1}{2(1 - c_1\varepsilon/4)n} \right\}.$$

To bound the size of \mathcal{I} , observe that:

$$|\mathcal{I}| \cdot \frac{1}{(1 - c_1\varepsilon/4)n} + (n - |\mathcal{I}|) \cdot \frac{1}{2(1 - c_1\varepsilon/4)n} \geq 1 \implies |\mathcal{I}| \geq \left(1 - \frac{c_1\varepsilon}{2}\right)n.$$

Hence, we get for this set by the definition of \mathcal{I} :

$$\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} w_i (z_i - \mu)(z_i - \mu)^\top \preceq \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} w_i (z_i - \tilde{\mu})(z_i - \tilde{\mu})^\top \preceq 2 \sum_{i=1}^n w_i (z_i - \tilde{\mu})(z_i - \tilde{\mu})^\top \preceq C \frac{\eta^2}{\varepsilon}.$$

Finally, to conclude the proof, note that there exists $w^\dagger \in \mathcal{W}_\varepsilon$ such that:

$$w^\dagger := \arg \min_{w \in \mathcal{W}_\varepsilon} \left\| \sum_{i=1}^n w_i (z_i - \tilde{\mu})(z_i - \tilde{\mu})^\top \right\| \text{ and } \tilde{\mu} = \sum_{i=1}^n w_i^\dagger z_i.$$

Note, also that:

$$\sum_{i=1}^n \left\| w_i^\dagger (z_i - \tilde{\mu})(z_i - \tilde{\mu})^\top \right\| \preceq C \frac{\eta^2}{\varepsilon}.$$

Observe, now, that the uniform distribution over \mathcal{I} , $w_\mathcal{I}$ and the distribution w^\dagger both satisfy $w_\mathcal{I}, w^\dagger \in \mathcal{W}_\varepsilon$. From [Lemma 25](#) and applying [Lemma 26](#) we get:

$$\forall \|v\| = 1 : |\langle v, \mu \rangle - \langle v, \tilde{\mu} \rangle| \leq C\eta.$$

As a consequence, we obtain:

$$\|\mu - \tilde{\mu}\| \leq C\eta$$

concluding the proof of the theorem. ■

Appendix D. The Quadratic Dependence on ε in Theorem 15 is Necessary

Here, we show that the quadratic dependence on ε in Theorem 15 is inherent and not merely an artefact of our analysis. Thus, the bucketing step employed in the median-of-means framework is required to build optimal sub-Gaussian estimators out of adversarially robust ones. In our proof, we build an adversarially robust estimator which satisfies Definition 4 but does not obtain optimal sub-Gaussian rates even in the *Gaussian* setting.

Given sample $\mathbf{X} = (x_1, \dots, x_n)$, assuming for simplicity that n is even, let $\mathbf{X}_1 = (x_1, \dots, x_{n/2})$ and $\mathbf{X}_2 = (x_{n/2+1}, \dots, x_n)$. Let $m := \text{Med}(\mathbf{X}_1)$, and let $\hat{p} := \frac{2}{n} \sum_{i=n/2+1}^n \mathbf{1}[x_i < m]$ be the fraction of samples in \mathbf{X}_2 which lie below m . Consider the event E defined below:

$$\hat{p} \geq \frac{1}{2} + 16\varepsilon.$$

Now, consider the algorithm which is a adversarial mean estimation algorithm, except that it always fails or outputs something arbitrarily bad when E occurs and the distribution does not have substantial mass concentrated on a particular point. Intuitively speaking, this will still be a valid, optimal, robust mean estimation algorithm because the robustness of median means that an adversary usually cannot, only changing εn of the samples, cause E to occur. However, it does not satisfy the requirements of an optimal heavy-tailed mean estimator, as the event E occurs with probability $\exp\{-O(\varepsilon^2 n)\}$ (rather than the $\exp\{-\Omega(\varepsilon n)\}$ we want). The rest of this section is dedicated to formalizing this intuition.

Algorithm 3 Adversarially Robust but Non Heavy Tailed Mean Estimator

- 1: **Input:** Point set $\mathbf{X} = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$, Corruption Factor ε
 - 2: $\mathbf{X}_1 \leftarrow \{x_1, \dots, x_{n/2}\}$, $\mathbf{X}_2 \leftarrow \{x_{n/2+1}, \dots, x_n\}$
 - 3: $m \leftarrow \text{Med}(\mathbf{X}_1)$
 - 4: $\hat{p} \leftarrow \frac{2}{n} \sum_{i=n/2+1}^n \mathbf{1}[x_i \leq m]$
 - 5: If $\hat{p} > \frac{1}{2} + 16\varepsilon$ and $\frac{2}{n} \sum_{i=n/2+1}^n \mathbf{1}[x_i = m] < 4\varepsilon$ then fail. (call this event E)
 - 6: Otherwise, run adversarially robust algorithm \mathcal{A} on \mathbf{X} and ε and output its answer.
-

The proof of the statement rests on the following two claims. In the first claim, we show that the E_ε -blowup of E remains small for any univariate distribution \mathcal{D} . Thus, the probability that an adversary could perturb an ε fraction of points to cause E to occur is less than 0.01.

Claim 32 *Let E_ε denote the ε blowup of E ; i.e $E_\varepsilon := \{\mathbf{X} : \exists \mathbf{X}' \in E \text{ s.t } \text{dist}_{\text{ham}}(\mathbf{X}, \mathbf{X}') \leq \varepsilon n\}$. Then,*

$$\mathbb{P}\{E_\varepsilon\} \leq 0.01.$$

Next, we show that specifically for *Gaussian* inputs, the probability that E occurs can be *lower bounded*. Hence, the probability of E occurring naturally, while less than a constant, is still larger than what is required by an optimal heavy-tailed estimator.

Claim 33 *We have:*

$$\mathbb{P}\{E\} \geq e^{-\theta(\varepsilon^2 n)}.$$

Together, these two claims imply that our algorithm is adversarially robust while not being an optimal heavy-tailed estimator. ■