1

Introduction to Robust Statistics

1.1 Introduction

Consider the following basic statistical task: Given n independent samples from a Gaussian, $\mathcal{N}(\mu, I)$, in $\mathbf{R}^d$ with identity covariance and unknown mean, estimate its mean vector μ to within small error in the ℓ_2-norm. It is not hard to see that the empirical mean has ℓ_2-error at most $O(\sqrt{d/n})$ with high probability. Moreover, this error upper bound is best possible among all n-sample estimators.

The Achilles heel of the empirical estimator is that it crucially relies on the assumption that the observations were generated by an identity covariance Gaussian. The existence of even a *single* outlier can arbitrarily compromise this estimator's performance. However, the Gaussian assumption is only ever approximately valid, as real datasets are typically exposed to some source of contamination. Hence, any estimator that is to be used in practice must be *robust* in the presence of outliers or model misspecification.

Learning in the presence of outliers is an important goal in statistics and has been studied within the robust statistics community since the 1960s. In recent years, the problem of designing robust and computationally efficient estimators for high-dimensional statistical tasks has become a rather pressing challenge in a number of data analysis applications. These include the analysis of biological datasets, where natural outliers are common and can contaminate the downstream statistical analysis, and data poisoning attacks in machine learning, where even a small fraction of fake data (outliers) can substantially degrade the quality of the learned model.

While classical work in robust statistics managed to determine most of the information-theoretic limits of robust estimation, the computational aspects were left wide open in high dimensions. In particular, a number of known robust estimators for basic high-dimensional statistical problems have been

https://doi.org/10.1017/9781108943161.001 Published online by Cambridge University Press

shown to be computationally intractable to compute. In fact, the conventional wisdom within the statistics community was that some of these problems were not solvable in a computationally efficient manner. In the Conclusions chapter of his book *Robust Statistical Procedures*, Peter J. Huber writes:

> The bad news is that with all currently known algorithms the effort of computing those estimates increases exponentially in d. We might say they break down by failing to give a timely answer! . . .
>
> The current trend toward ever-larger computer-collected and computer-managed data bases poses interesting challenges to statistics and data-analysis in general. [. . .] Only simple algorithms (i.e., with a low degree of computational complexity) will survive the onslaught of huge data sets. This runs counter to recent developments in computational robust statistics.
>
> It appears to me that none of the above problems will be amenable to a treatment through theorems and proofs. They will have to be attacked by heuristics and judgement, and by alternative "what if" analyses.

In the subsequent decades, there was a striking tension between robustness and computational efficiency in high dimensions. Specifically, even for the most basic task of high-dimensional mean estimation, all known estimators were either hard to compute or were very sensitive to outliers. This state of affairs changed fairly recently with the development of the first computationally efficient estimators for high-dimensional robust statistics problems. This book is dedicated to describing these developments and the techniques that have built upon them in the intervening years.

Before getting into the core of these recent developments, it is prudent to first describe the state of affairs before these algorithms were discovered. In this chapter, we will cover this basic background by describing the underlying models that we will be considering, analyzing basic robust estimators in one dimension, and discussing some of the difficulties involved with generalizing these estimators to higher dimensions.

1.2 Contamination Model

In order to specify a robust statistics problem, one needs to know three things:

1. What does the clean (uncorrupted) data look like?
2. What statistics of this data is the algorithm trying to estimate?
3. What kinds of contamination is the algorithm expected to deal with?

As we will see, most robust estimation tasks are provably impossible without imposing some sort of niceness assumptions on the clean data (inliers). At a

https://doi.org/10.1017/9781108943161.001 Published online by Cambridge University Press

high level, this is because without assuming anything about the inlier data, one would have no way of determining whether an extreme outlier is a corruption or simply an uncorrupted datapoint that happens to be far from most of the rest of the data. Thus, for most problems, we will need to make some assumptions on the distribution that the uncorrupted data is drawn from. One of the strongest niceness assumptions is that the inliers are distributed according to a Gaussian distribution. More generally, one might also consider what can be accomplished with weaker assumptions, such as log-concavity or simply some bounds on the higher moments of the inlier distribution. In fact, a lot of the progress in algorithmic robust statistics has involved investigating what kinds of assumptions about the inlier data can be relaxed without sacrificing computational efficiency.

In terms of what our algorithm is trying to estimate, we will usually focus on fairly simple statistics like the mean or covariance of the uncorrupted data. However, sometimes we will have more sophisticated goals, such as trying to learn the entire distribution up to small error, or learn some other more complicated underlying parameter.

Finally, the choice of contamination model bears a deeper discussion. We will elaborate on various natural assumptions over the course of the next few sections.

1.2.1 Contamination Model Basics

There are many ways that datasets might be corrupted, and many models to describe such corruptions. If one is optimistic, one might assume that the corruptions are *random*; that is, some datapoints are randomly replaced by samples from a known error distribution or are otherwise corrupted by some known random process. Given such an understanding of the underlying errors, robust estimation tasks typically become much easier, as one can try to find efficient ways to cancel out the effects of these *predictable* errors.

One might also assume that one is merely dealing with *small* measurement errors. That is, perhaps every datapoint is corrupted in some unpredictable way, but no datapoint is corrupted by very much. In this case, one might hope that these small corruptions will not be enough to substantially change the outcome of the estimator being used. In other words, robustness against these kinds of errors amounts to a question about the numerical stability of the estimators.

Unfortunately, these kinds of assumptions are too optimistic in a number of applications. Random corruption models assume that the source of errors is understood sufficiently well that they can essentially be incorporated as just

https://doi.org/10.1017/9781108943161.001 Published online by Cambridge University Press

another parameter in the model. Meanwhile, algorithms robust to small errors may be unable to cope with the presence of a small number of outliers.

The error models that we focus on in this book generally allow for worst-case corruptions that affect only a small constant fraction (usually denoted by ϵ) of our data. This means that, for example, 1% of our data is not coming from the inlier distribution, but instead from some source of errors that we have no control over. These errors might produce very extreme outliers, and they will not necessarily come from any model that we could predict ahead of time. In the worst case, one might even consider an adversary designing these errors in such a way to best thwart our algorithm. The latter kind of error model might be slightly more pessimistic than necessary about how bad our errors are; but if we can design algorithms that work under these pessimistic models, the results will apply very broadly. Such algorithms will work against *any* kind of corruptions, as long as these corruptions do not affect too large a fraction of our inputs.

That said, there are still a few things that we need to specify about these corruption models, having to do with whether they add or remove points and whether they are adaptive or not.

1.2.2 Additive and Subtractive Nonadaptive Corruptions

Among the types of contamination models that we will consider in this book, one important defining feature is what the errors are allowed to do to the clean (inlier) data. At a high level, this will leave us with three basic types of error models:

- **Additive Contamination:** In additive contamination models, the errors consist of new, incorrect, datapoints being inserted into our dataset.
- **Subtractive Contamination:** In subtractive contamination models, the errors consist of clean datapoints being selectively removed from our dataset.
- **General Contamination:** In general contamination models, both kinds of errors can occur. Erroneous datapoints can be inserted and clean ones can be removed. Equivalently, we can think of these corruptions as *replacing* our clean datapoints with outliers.

We formally define the corresponding contamination models below.

Definition 1.1 (Additive, Nonadaptive Contamination (Huber Model)) Given a parameter $0 < \epsilon < 1$ and a distribution D on inliers, we say that one can sample from D with ϵ-additive contamination if one can obtain independent samples from a distribution X of the following form: A sample from X returns

https://doi.org/10.1017/9781108943161.001 Published online by Cambridge University Press

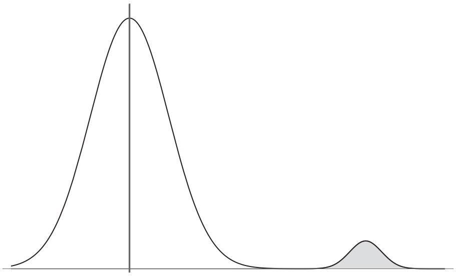

Figure 1.1 Example of a Gaussian with additive contamination. The error distribution corresponds to the gray bump on the right.

a sample from D with probability $(1 - \epsilon)$, and otherwise returns a sample from some (unconstrained and unknown) error distribution E.

See Figure 1.1 for an example of additive contamination.

The parameter ϵ is the proportion of contamination and quantifies the power of the adversary. Among the samples, an unknown $(1-\epsilon)$-fraction are generated from a distribution of interest; we will call these samples *inliers*. The remaining samples are drawn from an arbitrary distribution; we will call these samples *outliers*.

Note that the distribution X being sampled in Definition 1.1 is a mixture of the distribution D over inliers (clean/good samples) and the distribution E over outliers (errors or corruptions). As we will often want to talk about these kinds of mixtures, we introduce the relevant notation.

Notation We will use linear combinations of probability distributions to denote the mixtures defined by the corresponding linear combination of the associated density functions. For example, if X_i are probability distributions and $p_i \geq 0$ are real numbers summing to 1, we will use $p_1X_1 + p_2X_2 + \cdots + p_kX_k$ or $\sum_{i=1}^{k} p_iX_i$ to denote the mixture X, where one can obtain a sample from X by first picking a number $1 \leq i \leq k$ so that a given i is picked with probability p_i, and then returning a sample from the corresponding X_i.

For example, the distribution X sampled from in the additive contamination model can be written as $X = (1 - \epsilon)D + \epsilon E$.

We note that if the distributions X_i are random variables, this is also the standard notation for taking a linear combination of the random variables X_i. In this text, we will typically use this notation to denote mixtures, and will make it clear from the context in the rare cases where we want it to denote a linear combination instead.

https://doi.org/10.1017/9781108943161.001 Published online by Cambridge University Press

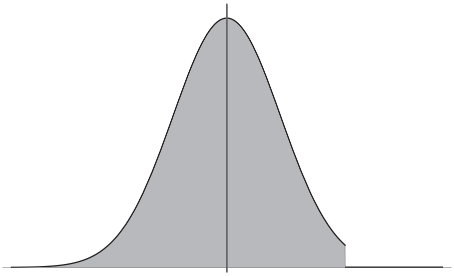

Figure 1.2 Example of a Gaussian with subtractive contamination. In particular, the right tail of the distribution has been removed.

For subtractive contamination, instead of inserting new samples (outliers), there is a probability of at most ϵ that samples are censored from the data that the algorithm observes. One way to define this is as follows.

Definition 1.2 (Subtractive, Nonadaptive Contamination) Given a parameter $0 < \epsilon < 1$ and a distribution D on inliers, we say that one can sample from D with ϵ-subtractive contamination if the following holds: For some event R with probability $1 - \epsilon$, one can obtain independent samples from the distribution of D conditioned on R.

See Figure 1.2 for an example of subtractive contamination.

In other words, with probability ϵ, the event R^c occurs and these samples are removed from the data stream. This allows an adversary to remove an ϵ-fraction of inlier samples. It is tempting to write that the observed distribution is proportional to $D - \epsilon E$, where E is the distribution over samples of D conditioned on R^c (i.e., the distribution over samples that are removed). We can make this rigorous with a slight extension of our above notation.

Notation We define a *pseudo-distribution* to be a real-valued measure. This means that a probability distribution is simply a nonnegative pseudo-distribution normalized to have total mass equal to 1. More generally, for any pseudo-distributions $X_1, \ldots, X_k$ and real numbers $p_1, \ldots, p_k$, we use $p_1X_1 + p_2X_2 + \cdots + p_kX_k$ or $\sum_{i=1}^{k} p_iX_i$ to denote the pseudo-distribution X whose density is given by the corresponding linear combination of densities of the pseudo-distributions X_i. In particular, for any set S, $X(S) = \sum_{i=1}^{k} p_iX_i(S)$.

We will often want to think of pseudo-distributions as some kind of non-normalized or nonpositive probability distributions, and think of these linear combinations as "mixtures" of these "distributions" even when neither term

https://doi.org/10.1017/9781108943161.001 Published online by Cambridge University Press

really applies. For example, one can write the distribution X on the observed samples from subtractive contamination as

$$X = \left(\frac{1}{1-\epsilon}\right) D - \left(\frac{\epsilon}{1-\epsilon}\right) E.$$

While this is not technically a mixture, it is useful to think of X as being obtained from D by first "subtracting" an ϵ-fraction of the distribution E and then renormalizing. Of course, this only makes sense if this ϵE was already contained in the distribution D. To convey this type of information, we introduce one further piece of notation.

Notation Given two pseudo-distributions X and Y, we use $X \leq Y$ to denote that the density of X is pointwise at most the density of Y. Equivalently, for every set S we have $X(S) \leq Y(S)$.

This means, for example, that the distribution E in the subtractive contamination model of Definition 1.2 must satisfy $\epsilon E \leq D$. Equivalently, the final distribution X must satisfy $(1-\epsilon)X \leq D$. Similarly, the additive contamination model of Definition 1.1 is defined by $X \geq (1-\epsilon)D$.

Finally, for the general contamination model, there are a few essentially equivalent reasonable definitions depending on the relative amounts of additive and subtractive contamination allowed. Perhaps the easiest way to deal with things is to allow the adversary to remove an ϵ-fraction of the probability mass of the inlier distribution and replace it with equal mass from some other distribution.

Definition 1.3 (General, Nonadaptive Contamination) Given a parameter $0 < \epsilon < 1$ and an inlier distribution D, we say that one can sample from D with ϵ-general contamination if one can obtain independent samples from a distribution of the form $X = D - \epsilon L + \epsilon E$, for distributions L and E with $\epsilon L \leq D$.

See Figure 1.3 for an example of general contamination.

This leads to a natural question as to which distributions X one can obtain in the general contamination model. It turns out that it is those that are close to D in *total variation distance*.

Definition 1.4 (Total Variation Distance) Given distributions X and Y, the *total variation distance* between them, denoted $d_{\mathrm{TV}}(X, Y)$, is defined to be half the L_1-norm of their difference, namely: $d_{\mathrm{TV}}(X, Y) := \frac{1}{2}\|X - Y\|_1$. If X and Y have probability density functions $p(x)$ and $q(x)$, we have $d_{\mathrm{TV}}(X, Y) = \frac{1}{2}\int |p(x) - q(x)| dx$. We also have the following equivalent definitions:

https://doi.org/10.1017/9781108943161.001 Published online by Cambridge University Press

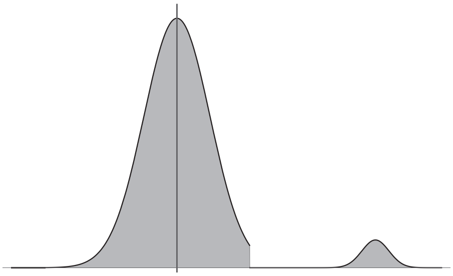

Figure 1.3 Example of a Gaussian with general contamination. The right tail of the distribution has been removed and replaced by the outlying bump on the right.

- The total variation distance is the biggest discrepancy between the probabilities of X and Y on any set, that is, $d_{\mathrm{TV}}(X, Y) = \sup_S(|X(S) - Y(S)|)$.
- If Y is thought of as being a copy of the distribution X with some small probability of error, the total variation distance characterizes how small that error can be. In particular, we can write $d_{\mathrm{TV}}(X, Y) = \inf_{A\sim X, B\sim Y} \mathbf{Pr}[A \neq B]$.

It is not hard to see that the general contamination model is equivalent to saying that one can sample from a distribution X with $d_{\mathrm{TV}}(X, D) \leq \epsilon$. This is particularly informative given the last of the above formulations of total variation distance, as it essentially says that the algorithm is receiving samples from D with probability $1 - \epsilon$ and with probability ϵ is getting some kind of error.

Remark 1.5 In many settings, the subtractive contamination model is much easier to deal with than the additive contamination model. For example, if the goal is to estimate the mean of the inlier distribution D, even a single additive error can corrupt the sample mean by an arbitrary amount. Subtractive errors on the other hand are limited in how much damage they can do, since they are only allowed to remove existing samples. For a single removed sample to have a large effect on the sample mean, it would need to be the case that the initial sample set already had some extreme outliers which could be removed. Because of this, most of this book will focus on the more challenging models of additive or general contamination.

1.2.3 Adaptive Corruptions

There is one aspect in which even the general contamination model is not as strong as it could be. All of the contamination models from the last section are what might be called *nonadaptive*. That is, they replace the distribution D

https://doi.org/10.1017/9781108943161.001 Published online by Cambridge University Press

over inlier samples by a distribution X by introducing some errors. But after doing this, the algorithm is then given honest, independent samples from the distribution X. A more insidious adversary might be able to choose what errors to introduce and which samples to corrupt, based on a knowledge of what the uncorrupted samples are. This idea leads us to our strongest contamination model.

Definition 1.6 (Strong Contamination Model) Given a parameter $0 < \epsilon < 1$ and an inlier distribution D, an algorithm receives samples from D with ϵ-contamination as follows: The algorithm specifies an integer number of samples n, and n samples are drawn independently from the distribution D. An adversary is then allowed to inspect these samples, remove up to $\lceil \epsilon n \rceil$ of them, and replace them with arbitrary points. The modified set of n points are then given to the algorithm.

In analogy with this adaptive version of the general noise model, we can devise an adaptive version of the additive noise model (that inserts $\lceil \epsilon n/(1-\epsilon) \rceil$ new samples into the dataset) and the adaptive subtractive noise model (that selects and removes $\lceil \epsilon n \rceil$ clean samples).

Although there are a few cases where it is useful to know that the errors that an algorithm is observing are i.i.d. samples from some distribution, most of the algorithms developed in this book can be shown to work in the strong contamination model. As this is the most powerful of the corruption models, we will state most of our results in this model.

1.3 Information-Theoretic Limits

Before we get into describing basic algorithms for robust estimation, we provide a succinct outline of the information-theoretic limits of such algorithms. The most basic of these limits is the following: If the samples are ϵ-contaminated (even by a nonadaptive adversary), then one cannot hope to learn the underlying distribution to total variation distance better than (approximately) ϵ. To state this formally, we present the following proposition.

Proposition 1.7 *Let X and Y be distributions with $d_{\mathrm{TV}}(X, Y) \leq 2\epsilon$ for some $0 < \epsilon < 1$. A distribution D is taken to be either X or Y. Then an algorithm, given any number of samples from D with ϵ-general contamination, cannot reliably distinguish between the cases $D = X$ and $D = Y$. Furthermore, the same holds if (i) $d_{\mathrm{TV}}(X, Y) \leq \epsilon/(1-\epsilon)$ and the samples have ϵ-additive contamination or if (ii) $d_{\mathrm{TV}}(X, Y) \leq \epsilon$ and the samples have ϵ-subtractive contamination.*

https://doi.org/10.1017/9781108943161.001 Published online by Cambridge University Press

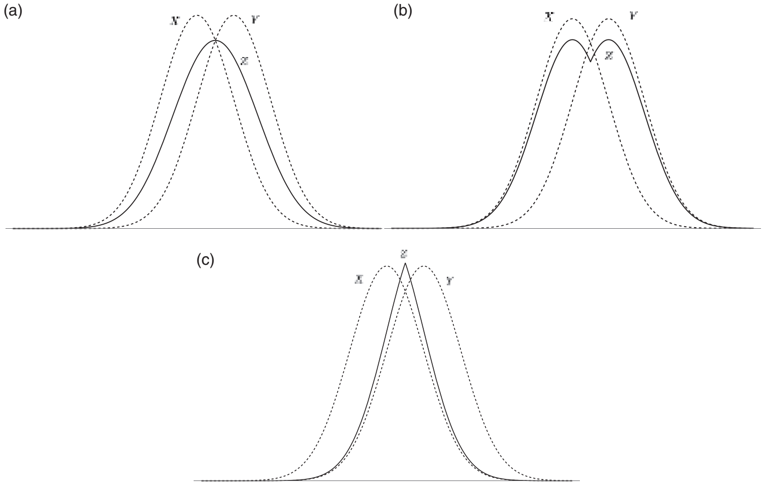

Figure 1.4 Illustration of construction of Z from X and Y: general contamination (a), additive contamination (b), subtractive contamination (c).

Proof In all three cases, the basic idea is that the adversary can find a single distribution Z such that Z is both an ϵ-contaminated version of X and an ϵ-contaminated version of Y. If the algorithm is then presented with independent samples from Z, there is no way to distinguish when these are contaminated samples from $D = X$ or contaminated samples from $D = Y$.

The constructions needed for our three types of contamination will be slightly different. See Figure 1.4 for an example of the construction of Z in each of the three cases.

In the case of general contamination, one can simply take $Z = (X + Y)/2$. In this case, we have

$$d_{\mathrm{TV}}(X, Z) = \frac{1}{2}\|X - Z\|_1 = \frac{1}{2}\|X - (X + Y)/2\|_1 = \frac{1}{2}\|(X - Y)/2\|_1 = \frac{1}{4}\|X - Y\|_1$$
$$= d_{\mathrm{TV}}(X, Y)/2 \leq \epsilon.$$

A similar bound on $d_{\mathrm{TV}}(Y, Z)$ completes the argument.

For additive and subtractive contamination, the argument is slightly more complicated. If X and Y have total variation distance δ, then writing $X - Y$ as a positive part and a negative part, we obtain $X = Y - \delta L + \delta A$, for some distributions L and A with $\delta L \leq Y$. Writing this slightly more symmetrically, we can take $W = (Y - \delta L)/(1 - \delta)$, and we have $X = (1 - \delta)W + \delta A$ and $Y = (1 - \delta)W + \delta L$.

For the case of subtractive contamination, we can take $Z = W$ as above.

https://doi.org/10.1017/9781108943161.001 Published online by Cambridge University Press

Then Z can be obtained from either X or Y by subtracting a δ-fraction of the mass and renormalizing.

For additive contamination, we can take $Z = ((1-\delta)W + \delta A + \delta L)/(1+\delta)$. We note that this can be obtained by adding $\delta/(1+\delta)$ additive contamination to either X and Y (adding L to X or A to Y). As long as $\delta \leq \epsilon/(1-\epsilon)$, we have $\delta/(1+\delta) \leq \epsilon$, which completes our proof. □

Remark 1.8 The distances at which X and Y are indistinguishable given corruptions, presented in Proposition 1.7, are essentially tight. See Exercise 1.3 for more details.

One interesting takeaway from Proposition 1.7 is that if $\epsilon \geq 1/2$, then one cannot reliably distinguish between *any* pair of distributions X and Y in the presence of ϵ additive or general contamination. This is because the total variation distance of any two distributions is at most 1. This means that for essentially every problem that we consider in this book (with the exception of topics covered in Chapter 5), we will need to assume that the proportion of contamination ϵ is less than $1/2$ in order for any guarantees to be possible.

Another implication of Proposition 1.7 is that it puts limits on our ability to robustly estimate basic statistics of the underlying distribution. For example, if one makes no assumptions on the underlying distribution D, it will be impossible (even with an unlimited number of samples) to learn the mean of D to within *any* bounded error. This is simply because one can find pairs of distributions X, Y with $d_{\mathrm{TV}}(X, Y) < \epsilon$ but with $\|\mathbf{E}[X] - \mathbf{E}[Y]\|$ unbounded.

Consequently, in order for meaningful results to be possible, we will need to consider settings where the inlier distribution is restricted to some well-behaved family. Broadly speaking, the best we can hope to achieve is to learn the underlying distribution within error $O(\epsilon)$ in total variation distance. If our distribution family is one where no ϵ-fraction of the probability mass can contribute too much to the mean (which is a measure of the concentration of the distribution), then this may suffice to obtain relatively good estimates of the mean. On the other hand, for families without this kind of concentration, we will be limited in how well we can expect to do.

The information-theoretic limitations for some basic distribution families are summarized below.

Lemma 1.9 *Let $\mathcal{D}$ be the family of one-dimensional Gaussian distributions with standard deviation* 1*. An algorithm with access to ϵ-corrupted samples (additive, subtractive, or general contamination) from an unknown distribution $D \in \mathcal{D}$ cannot reliably estimate $\mathbf{E}[D]$ to additive error $o(\epsilon)$.*

https://doi.org/10.1017/9781108943161.001 Published online by Cambridge University Press

Proof Let δ be a sufficiently small constant multiple of ϵ. It is not hard to see that $d_{\mathrm{TV}}(\mathcal{N}(0,1), \mathcal{N}(\delta,1)) < \epsilon$. Therefore, by Proposition 1.7, no algorithm can reliably distinguish between $G = \mathcal{N}(0,1)$ and $G = \mathcal{N}(\delta,1)$. However, these distributions have means that differ by δ. If an algorithm could estimate the mean to error better than $\delta/2$, it could use this estimate to distinguish between these distributions, yielding a contradiction. □

Using similar logic, we can obtain analogous results for some other natural distribution families.

Lemma 1.10 *Let $\mathcal{D}$ be the family of one-dimensional log-concave distributions with standard deviation* 1. *An algorithm with access to ϵ-corrupted samples from an unknown distribution $D \in \mathcal{D}$ cannot reliably estimate* $\mathbf{E}[D]$ *to additive error $o(\epsilon \log(1/\epsilon))$.*

Lemma 1.11 *Let $\mathcal{D}$ be the family of all one-dimensional distributions with standard deviation at most* 1. *An algorithm with access to ϵ-corrupted samples from an unknown distribution $D \in \mathcal{D}$ cannot reliably estimate* $\mathbf{E}[D]$ *to within additive error $o(\sqrt{\epsilon})$.*

Lemma 1.12 *Let $\mathcal{D}$ be the family of one-dimensional distributions D satisfying* $\mathbf{E}[|D-\mu_D|^k] < 1$, *for some $k \geq 2$, and* $\mu_D = \mathbf{E}[D]$ *(i.e., distributions with kth central moment bounded above by* 1*). An algorithm with access to ϵ-corrupted samples from an unknown distribution $D \in \mathcal{D}$ cannot reliably estimate* $\mathbf{E}[D]$ *to additive error $o(\epsilon^{1-1/k})$.*

1.4 One-Dimensional Robust Estimation

We begin our analysis of computationally efficient robust statistics by solving some of the most fundamental estimation tasks for natural families of one-dimensional distributions. This will allow us to gain a basic understanding of some useful techniques and principles without having to deal with many of the difficulties introduced by high-dimensional versions of these problems. In particular, we focus on robust estimators of the mean and standard deviation. For these problems, we will assume that the distribution D over inlier samples comes from some known family $\mathcal{D}$, and we will give algorithms that robustly estimate the mean and variance of D, given access to ϵ-corrupted samples from D.

1.4.1 Estimators Based on Order Statistics

One of the difficulties of robust mean estimation is that the empirical mean itself is very far from being robust. In particular, a single extreme outlier can

https://doi.org/10.1017/9781108943161.001 Published online by Cambridge University Press

corrupt the mean of a finite sample set by an arbitrarily large error. This is not an issue for the median and other order statistics, thus making them good candidates for designing robust estimators. To set things up, we define the quantiles of a distribution or a set.

Definition 1.13 (Quantiles) Let X be a distribution on $\mathbf{R}$ and $q \in [0,1]$. We define the q-quantile of X to be the infimum over all $t \in \mathbf{R}$ such that $\mathbf{Pr}[X \leq t] \geq q$. If S is a multiset of real numbers, then the q-quantile of S is the q-quantile of the uniform distribution over S.

The basic result about quantiles is that the empirical q-quantile of a distribution is a fairly good empirical estimator of the true q-quantile.

Proposition 1.14 *Let X be a distribution on $\mathbf{R}$ and let $0 < \epsilon, \delta < 1/2$. Let S be a set of n samples from X that are ϵ-corrupted under the strong contamination model. Then with probability at least $1 - \delta$, the q-quantile of S is between the $(q - \epsilon + O(\sqrt{\log(1/\delta)/n}))$-quantile of X and the $(q + \epsilon + O(\sqrt{\log(1/\delta)/n}))$-quantile of X.*

Proof We will show that with probability at least $1 - \delta/2$, the q-quantile of S is at least the $(q - \epsilon + O(\sqrt{\log(1/\delta)/n}))$-quantile of X. The upper bound will follow similarly. By definition, the q-quantile of S is the minimum value that is bigger than at least qn elements of S. In other words, we need to show that if we take t to be the $(q - \epsilon - C(\sqrt{\log(1/\delta)/n}))$-quantile of X for $C > 0$ some sufficiently large constant, then with probability at least $1 - \delta/2$, there are at most qn elements of S less than t.

The proof is quite simple. The set S was generated by first sampling n independent elements from X. Each of these elements independently and with probability at most $(q - \epsilon - C(\sqrt{\log(1/\delta)/n}))$ are less than t. Therefore, by the Chernoff bound, with probability at least $1 - \delta/2$, the number of original samples with value less than t was no more than $(q - \epsilon)n$. Upon corrupting ϵn of these samples, we still have at most qn samples less than t. This completes the proof of the lower bound. The upper bound follows analogously. □

Proposition 1.14 is useful for estimating the mean of distributions for which the mean can be related to an order statistic. Perhaps the most common such case is that of distributions symmetric about their mean, as for such distributions the mean and median will be the same. This result can be applied in particular for the case of Gaussian distributions.

Corollary 1.15 *Let $D = \mathcal{N}(\mu, \sigma^2)$ be a one-dimensional Gaussian distribution. Let S be an ϵ-corrupted set of n samples from D, for some $\epsilon < 1/3$, and*

https://doi.org/10.1017/9781108943161.001 Published online by Cambridge University Press

let m be its median. Then, if δ is at least e^{-an} for some sufficiently small a, with probability $1-\delta$ we have

$$|m-\mu| = O(\epsilon + \sqrt{\log(1/\delta)/n})\sigma.$$

Proof By Proposition 1.14, with probability $1-\delta$ we have that m is between the $(1/2-\epsilon+O(\sqrt{\log(1/\delta)/n}))$-quantile and the $(1/2+\epsilon+O(\sqrt{\log(1/\delta)/n}))$-quantile of D. Since the $(1/2+\eta)$-quantile of D is $\mu+O(\eta\sigma)$ for $\eta < 2/5$, the result follows. □

In order to robustly estimate the standard deviation for certain distribution families, one can express the standard deviation in terms of a difference between order statistics. For example, the Inter-Quartile-Range (IQR) is the difference between the 1/4-quantile and the 3/4-quantile. Specifically, it is not hard to see that for Gaussian distributions, $D = \mathcal{N}(\mu, \sigma^2)$, the IQR of D is equal to $c_{iqr}\,\sigma$, for some universal constant c_{iqr}. Using this fact, we can obtain a robust estimator for the standard deviation of a Gaussian distribution.

Corollary 1.16 *Let $D = \mathcal{N}(\mu, \sigma^2)$ be a one-dimensional Gaussian distribution. Let S be an ϵ-corrupted set of n samples from D, for some $\epsilon < 1/8$, and let r be the IQR of S. Let c_{iqr} be the aforementioned universal constant. Then, if δ is at least e^{-an} for some sufficiently small a, with probability $1-\delta$ we have*

$$\sigma = c_{iqr}\, r\,(1+O(\epsilon+\sqrt{\log(1/\delta)/n})).$$

Proof By Proposition 1.14, with probability at least $1-\delta$, each of the empirical quartiles correspond to the $(1/4 \pm \epsilon + O(\sqrt{\log(1/\delta)/n}))$-quantile and the $(3/4 \pm \epsilon + O(\sqrt{\log(1/\delta)/n}))$-quantile of D. This means that they are each within $O(\epsilon+O(\sqrt{\log(1/\delta)/n}))\sigma$ of the 1/4- and 3/4-quantiles. Thus, r is within $O(\epsilon+O(\sqrt{\log(1/\delta)/n}))\sigma$ of the IQR of D (and also at least a constant multiple of) σ, and the result follows. □

One point worth making about Corollaries 1.15 and 1.16 is that both have error proportional to σ. This means that while the mean can be estimated to an additive error of $O(\epsilon\sigma)$, the standard deviation can only be estimated up to multiplicative error. This is a fairly common phenomenon.

Unfortunately, while the above-described estimators work quite well for Gaussian distributions, they are fairly specific and not generalizable. The median estimator essentially requires that the mean and median be the same; this works for symmetric distributions, but for skewed ones it does not work in general. The IQR, as an estimator of the standard deviation, is even more fragile. While it is not hard to show, using Chebyshev's inequality, that the IQR is never more than a constant factor larger than the standard deviation (and not

https://doi.org/10.1017/9781108943161.001 Published online by Cambridge University Press

much smaller for “nice” distributions), getting a precise relationship between the two essentially only worked here because the family of Gaussians has only one distribution up to affine transformation.

In order to obtain estimators for more complicated families, we will need to do something more similar to computing an actual mean. However, we will need to do this in such a way that an ϵ-fraction of samples being very extreme errors will not significantly affect the estimate. One fairly straightforward way to achieve this is by simply throwing away the few most extreme datapoints on each side and computing a *truncated mean*.

1.4.2 Estimators Based on Truncation

In general, we need an estimator that is not too much affected by an ϵ-fraction of the points being either very large or very small. A natural way to correct this is to take any points in the top or bottom ϵ-fraction and either throw them away or reduce them to something more manageable. There are a few ways to define the relevant truncation operation; the following is perhaps the most efficient version.

Definition 1.17 (Truncation) Given a distribution X on $\mathbf{R}$ and $0 < \epsilon < 1/2$, we define the *ϵ-truncation* of X to be the distribution obtained by taking X conditioned on the values lying between the ϵ-quantile and the $(1-\epsilon)$-quantile.

See Figure 1.5 for an illustration of this definition.

The following proposition shows that, under reasonable assumptions, the mean of the truncated empirical distribution can provide a good robust estimate of the true mean.

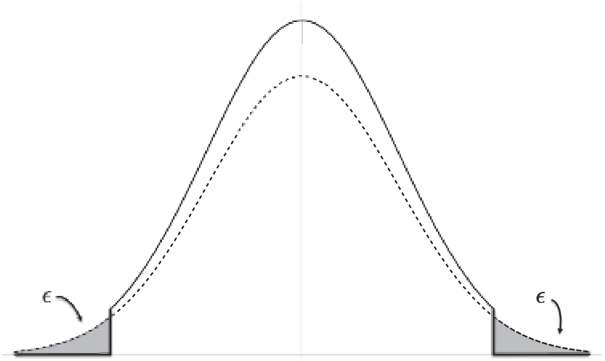

Figure 1.5 A Gaussian (dotted line) and its ϵ-truncation (solid line) for $\epsilon = 0.1$. The ϵ-tails of the distribution on both sides are removed and the remaining distribution rescaled.

https://doi.org/10.1017/9781108943161.001 Published online by Cambridge University Press

Proposition 1.18 *Let $0 < \epsilon < \epsilon' < 1/2$. Let D be a distribution on $\mathbf{R}$ with mean μ such that removing any $2\epsilon'$-fraction of the mass of D changes the mean by at most $\eta > 0$ in absolute value. Let n be an integer at least a sufficiently large constant multiple of $\log(1/\delta)/(\epsilon' - \epsilon)^2$, for some $0 < \delta < 1/2$. Let S_0 be a set of n independent samples from D and let S be obtained from S_0 by adversarially corrupting an ϵ-fraction of its elements. Then, with probability at least $1-\delta$, the mean $\widehat{\mu}$ of the ϵ'-truncated empirical distribution of S satisfies $|\widehat{\mu} - \mu| \leq \eta$.*

Remark 1.19 Some version of the assumption made in Proposition 1.18 – that removing a $2\epsilon'$-fraction of the mass of D does not change the mean by much – is essentially necessary for robust mean estimation to be possible. For example, suppose that D' can be obtained from D by removing a 2ϵ-fraction of its mass and that $|\mathbf{E}[D] - \mathbf{E}[D']| > \eta$. Then, $d_{\mathrm{TV}}(D, D') \leq 2\epsilon$, so by Proposition 1.7 one cannot distinguish between D and D' with any number of samples. Therefore, one cannot hope to estimate the mean to error better than $\eta/2$.

Proof First, we note that for any distribution X and any $m \in \mathbf{R}$, we have

$$\mathbf{E}[X] - m = \int_m^\infty \mathbf{Pr}[X > t]dt - \int_{-\infty}^m \mathbf{Pr}[X < t]dt.$$

If X_ϵ is the ϵ-truncation of X, then we can write $\mathbf{Pr}[X_\epsilon > t]$ as

$$f_\epsilon(\mathbf{Pr}[X > t]) \overset{\text{def}}{=} \begin{cases} 0 & \text{if } \mathbf{Pr}[X > t] < \epsilon, \\ (\mathbf{Pr}[X > t] - \epsilon)/(1 - 2\epsilon) & \text{if } 1 - \epsilon > \mathbf{Pr}[X > t] > \epsilon, \\ 1 & \text{if } \mathbf{Pr}[X > t] > 1 - \epsilon. \end{cases}$$

In particular, letting m be the median of D, we have

$$\mathbf{E}[D] - m = \int_m^\infty \mathbf{Pr}[D > t]dt - \int_{-\infty}^m \mathbf{Pr}[D < t]dt.$$

For the truncated version of S, we can write

$$\mathbf{E}[S_{\epsilon'}] - m = \int_m^\infty f_{\epsilon'}(\mathbf{Pr}_{x\sim_u S}[x > t])dt - \int_{-\infty}^m f_{\epsilon'}(\mathbf{Pr}_{x\sim_u S}[x < t])dt.$$

By definition, the empirical probability $\mathbf{Pr}_{x\sim_u S}[x > t]$ is the fraction of elements of S that are bigger than t. This quantity is within ϵ of $\mathbf{Pr}_{x\sim_u S_0}[x > t]$. For any given value of t, by our choice of n, with probability $1 - \delta/2$, it will hold that

$$\left|\mathbf{Pr}_{x\sim_u S_0}[x > t] - \mathbf{Pr}[D > t]\right| < (\epsilon' - \epsilon).$$

In fact, by the VC inequality (Theorem A.12), with probability $1-\delta$, this holds

https://doi.org/10.1017/9781108943161.001 Published online by Cambridge University Press

simultaneously for all t. If this is the case, we have

$$\mathbf{E}[S_{\epsilon'}] - m = \int_m^\infty f_{\epsilon'}(\mathbf{Pr}[D > t] \pm \epsilon')dt - \int_{-\infty}^m f_{\epsilon'}(\mathbf{Pr}[D < t] \pm \epsilon')dt.$$

This is at most

$$\begin{aligned}\mathbf{E}[S_{\epsilon'}] - m &\leq \int_m^\infty f_{\epsilon'}(\mathbf{Pr}[D > t] + \epsilon')dt - \int_{-\infty}^m f_{\epsilon'}(\mathbf{Pr}[D < t] - \epsilon')dt \\ &\leq \int_m^\infty \mathbf{Pr}[D > t]/(1 - 2\epsilon')dt - \int_{-\infty}^m \max(0, \mathbf{Pr}[D < t] - 2\epsilon')/(1 - 2\epsilon')dt.\end{aligned}$$

Letting $D^+_{2\epsilon'}$ be the distribution obtained by conditioning D on x being larger than the $2\epsilon'$-quantile, then the above can be seen to equal $\mathbf{E}[D^+_{2\epsilon'}] - m$. Since $D^+_{2\epsilon'}$ is obtained from D by removing a $2\epsilon'$-fraction of the mass, we have

$$\begin{aligned}\widehat{\mu} - \mu &= (\mathbf{E}[S_{\epsilon'}] - m) - (\mathbf{E}[D] - m) \\ &\leq (\mathbf{E}[D^+_{2\epsilon'}] - m) - (\mathbf{E}[D] - m) \\ &\leq \eta.\end{aligned}$$

The lower bound follows similarly. □

Proposition 1.18 applies to a much broader family of distributions than just Gaussians. Specifically, it is not hard to see that if $D = \mathcal{N}(\mu, \sigma^2)$ is a Gaussian and ϵ' is $O(\epsilon)$ and at most $1/3$, the error η can be taken to be $O(\epsilon\sqrt{\log(1/\epsilon)}\sigma)$. The exact same guarantee holds if D is any sub-Gaussian distribution with standard deviation σ, that is, a distribution whose tails decay at least as fast as the tails of the Gaussian with the same standard deviation.

On the other hand, if D is a general log-concave distribution with standard deviation σ, η is at most $O(\epsilon \log(1/\epsilon)\sigma)$. More generally, if D has kth central moment at most 1, we have that $\eta = O(\epsilon^{1-1/k})$.

Finally, we note that if one wants to robustly compute the variance of D for these more general families, the simplest technique is to first use a truncated mean to obtain an estimate $\widehat{\mu}$ for the mean of D, and then use another truncated mean to estimate the average value of $(D - \widehat{\mu})^2$.

1.5 Higher-Dimensional Robust Mean Estimation

While the techniques in the previous section do a fairly good job of estimating the mean of a one-dimensional random variable, generalizing these techniques to higher dimensional problems is somewhat tricky. For concreteness, we will work with perhaps the simplest problem in this family. Let $D = \mathcal{N}(\mu, I_d)$ be a

https://doi.org/10.1017/9781108943161.001 Published online by Cambridge University Press

d-dimensional Gaussian with identity covariance matrix and unknown mean μ. Given access to ϵ-corrupted samples from D, the goal is to estimate its mean μ up to a small error in ℓ_2-norm.

We start by discussing the difficulties involved with robustly estimating μ in higher dimensions. First, we would like to understand the information-theoretic limits for this problem. By Proposition 1.7, we know that we cannot hope to distinguish between a pair of Gaussians with total variation distance at most ϵ. For the family of spherical Gaussians, it is not hard to show that $d_{\mathrm{TV}}(\mathcal{N}(\mu, I_d), \mathcal{N}(\mu', I_d)) = \Theta(\min(1, \|\mu - \mu'\|_2))$. Therefore, we cannot hope to learn the mean to ℓ_2-error $o(\epsilon)$ in the presence of ϵ-corruptions.

Switching our attention to algorithms, perhaps the most natural approach is to try to generalize the one-dimensional median-based estimator; alas, it is unclear how to achieve this, as there are various ways to define a notion of "median" in high dimensions. One natural idea is to use the coordinate-wise median: That is, take a number of samples x_i and for each coordinate j take the median of the jth coordinates of the x_i. Since the jth-coordinates are distributed as $\mathcal{N}(\mu_j, 1)$, this gives an $O(\epsilon)$-approximation for each coordinate of μ by Corollary 1.15. Unfortunately, an estimator that guarantees error $O(\epsilon)$ in each coordinate might still have ℓ_2 error as large as $\Omega(\epsilon\sqrt{d})$, which is significantly worse than our desired error.

Interestingly, it turns out that a generalization of this idea does work – leading to a sample-efficient (but computationally inefficient) multivariate robust mean estimator. Note that if v is a unit vector in $\mathbf{R}^d$, then $v \cdot D$ is distributed as $\mathcal{N}(v \cdot \mu, 1)$. Using a one-dimensional robust mean estimator for this Gaussian random variable (such as the empirical median), we can obtain an estimate m_v such that with high probability $|m_v - v \cdot \mu| = O(\epsilon)$. The idea of our high-dimensional robust mean estimator is the following: If we can compute these approximations for *every* unit vector v, this will suffice to estimate μ. In particular, if we can find *any* $\widehat{\mu} \in \mathbf{R}^d$ such that $|m_v - v \cdot \widehat{\mu}| = O(\epsilon)$ for all unit vectors v (note that such vectors $\widehat{\mu}$ exist, since μ satisfies this requirement), then we have

$$\|\mu - \widehat{\mu}\|_2 = \sup_{\|v\|_2=1} |v \cdot (\mu - \widehat{\mu})| \leq \sup_{\|v\|_2=1} (|v \cdot \mu - m_v| + |v \cdot \widehat{\mu} - m_v|) = O(\epsilon). \quad (1.1)$$

In order to be able to actually find such a $\widehat{\mu}$, we will need that the median be a good estimator of the mean in every linear projection. Looking at the proof of Proposition 1.14, it can be seen that this will hold if for our set S of uncorrupted samples and every unit vector v and $t \in \mathbf{R}$, we have

$$\left|\mathbf{Pr}_{x \sim_u S}[v \cdot x > t] - \mathbf{Pr}[v \cdot G > t]\right| < \epsilon.$$

By the VC inequality (Theorem A.12), this holds with high probability, as

https://doi.org/10.1017/9781108943161.001 Published online by Cambridge University Press

long as the number of samples is at least a sufficiently large constant multiple of d/ϵ^2.

This argument shows that multivariate robust mean estimation of a spherical Gaussian with ℓ_2-error of $O(\epsilon)$ – independent of the dimension! – is in fact possible information-theoretically; alas, the implied estimator is highly nontrivial to compute. Taken literally, one would first need to compute m_v for every unit vector v (i.e., for infinitely many directions), and then find some appropriate $\widehat{\mu}$. Via a slight relaxation of the aforementioned argument, the situation is not this bad. If we modify Equation (1.1), we note that it is actually sufficient to have $|v \cdot \widehat{\mu} - m_v| = O(\epsilon)$ for all unit vectors v in some finite cover C of the unit sphere. In particular, we will need to know that

$$\|\mu - \widehat{\mu}\|_2 = O(\sup_{v \in C} |v \cdot (\mu - \widehat{\mu})|).$$

Fortunately, there exist finite covers C of the unit sphere such that for any $x \in \mathbf{R}^d$,

$$\|x\|_2 = O(\sup_{v \in C} |v \cdot x|). \tag{1.2}$$

See Theorem A.10. On the other hand, it is not hard to see that for Equation (1.2) to hold for even a random x, we need to have $|C|$ scale exponentially in d.

In summary, this relaxation *does* give us the following exponential-time algorithm for robust mean estimation: Given such a set C of size $2^{O(d)}$, we first compute m_v for each $v \in C$, and then solve a linear program (of exponential size) to find a $\widehat{\mu}$ satisfying $|v \cdot \widehat{\mu} - m_v| = O(\epsilon)$ for all $v \in C$. This yields an algorithm with runtime $\mathrm{poly}(2^d/\epsilon)$.

This discussion is summarized in the following proposition.

Proposition 1.20 *There exists an algorithm that, on input of an ϵ-corrupted set of samples from $D = \mathcal{N}(\mu, I_d)$ of size $n = \Omega((d + \log(1/\tau))/\epsilon^2)$, runs in* $\mathrm{poly}(n, 2^d)$ *time, and outputs $\widehat{\mu} \in \mathbf{R}^d$ such that with probability at least $1 - \tau$, it holds that $\|\widehat{\mu} - \mu\|_2 = O(\epsilon)$.*

For distributions other than Gaussians, one can provide a similar analysis. As long as there exists a one-dimensional robust mean estimator that can approximate $v \cdot \mu$ to error δ for every unit vector v (and assuming that we can make this hold in all directions simultaneously with a limited number of samples), then one can use this to construct an estimator of μ with ℓ_2 error $O(\delta)$ in exponential time (see Exercise 1.12).

Connection with Tukey Median There is a classical method to robustly estimate the mean of symmetric distributions known as the *Tukey median*, which

https://doi.org/10.1017/9781108943161.001 Published online by Cambridge University Press

can be thought of as a variation on the estimator from Proposition 1.20 by using median-based estimators. In particular, given a distribution D, we define the *Tukey depth* of a point y with respect to D as the minimum over unit vectors v of $\mathbf{Pr}_{x\sim D}[v \cdot x > v \cdot y]$. The Tukey median of a distribution D is then any point with maximum Tukey depth.

If $D = \mathcal{N}(\mu, I_d)$ is a Gaussian distribution, then the Tukey depth of the mean μ will be $1/2$. Similarly, for the sample case, the Tukey depth of μ with respect to the uniform distribution over a sufficiently large number of samples from D will be arbitrarily close to $1/2$ with high probability. If D is replaced by an ϵ-corruption of D (or if an ϵ-fraction of the samples are corrupted), then the Tukey depth of μ_D will still be $1/2 - O(\epsilon)$. Moreover, it is not hard to show that *any* point y with Tukey depth $1/2 - O(\epsilon)$ with respect to the ϵ-corruption of D (or its samples) will satisfy $\|x - y\|_2 = O(\epsilon)$ with high probability.

In summary, for Gaussians and other symmetric distributions, the Tukey median provides another method of robustly estimating the mean to near-optimal error. Unfortunately, computing the Tukey median also leads to computational issues. In particular, it has been shown that computing a Tukey median of an arbitrary point set is NP-Hard.

The kind of results described in this section were the state of the art for several decades. High-dimensional robust mean estimation, even for the simple case of spherical Gaussians, had three kinds of algorithms: Those that were of an entirely heuristic nature (i.e., without provable error guarantees); those that had error guarantees which scaled polynomially in the dimension; and those that had runtimes which scaled exponentially in the dimension. This held until a new class of algorithms arose to circumvent both of these problems; we will discuss these developments in the next chapter.

1.6 Connection with Breakdown Point

The focus of this book is on developing robust estimators to approximate a desired parameter of a distribution given an ϵ-corrupted dataset. Specifically, we want robust estimators that approximate a target parameter as accurately as possible, and in particular with no dependence on the underlying dimensionality of the data. Until recently, such dimension-independent error guarantees could not be achieved in high dimensions with computationally efficient algorithms.

Classical work in robust statistics largely focused on designing robust estimators with large *breakdown point*. The breakdown point of an estimator is a natural notion that quantifies the effect (or influence) of the outliers on its

https://doi.org/10.1017/9781108943161.001 Published online by Cambridge University Press

performance. Here we define a population variant of the breakdown point, specifically for the problem of robust mean estimation. Similar definitions exist for various other parameter estimation tasks.

While estimators typically act on finite sample sets, for simplicity we think about estimators as acting on distributions (by treating samples as the uniform distribution over the samples and considering the infinite sample regime). That is, we view an estimator T as a function mapping a distribution to the desired parameter (mean vector). For a distribution p and an estimator T, we will denote by $T(p)$ the mean vector estimate that T outputs given p.

Given this notation, we start by defining the notion of maximum bias.

Definition 1.21 (Maximum Bias) For a fixed distribution p and a contamination parameter $0 < \epsilon < 1/2$, the maximum ϵ-bias $b_T(p, \epsilon)$ of the estimator T is defined to be the supremum ℓ_2-distance between $T(p)$ and $T(\widetilde{p})$, where $\widetilde{p}$ is an ϵ-corruption of p (under additive, subtractive, or general contamination). For general contamination, we can write

$$b_T(p, \epsilon) = \sup\{\|T(p) - T(\widetilde{p})\|_2 \mid d_{\mathrm{TV}}(\widetilde{p}, p) \le \epsilon\}\,.$$

The breakdown point $\epsilon^*(p)$ is defined as the minimum fraction of corruptions that can drive the maximum bias to infinity.

Definition 1.22 (Breakdown Point) For a fixed distribution p, the breakdown point $\epsilon^*(T, p)$ of the estimator T on p is defined to be the infimum value of ϵ such that the maximum ϵ-bias of T on p is unbounded. For general contamination, we can write $\epsilon^*(T, p) = \inf\{\epsilon \mid b_T(p, \epsilon) = \infty\}$. For a family of distributions $\mathcal{D}$, the breakdown point of an estimator T on $\mathcal{D}$ is the worst breakdown point for any distribution $p \in \mathcal{D}$, that is, $\epsilon^*(T, \mathcal{D}) = \inf\{\epsilon^*(T, p), p \in \mathcal{D}\}$.

While the notion of breakdown point can be quite informative in certain settings, it is generally not sufficiently precise to quantify the robustness of an estimator in high dimensions. We provide a few illustrative examples for the problem of robust mean estimation when the inlier distribution is an identity covariance Gaussian, that is, when the family $\mathcal{D}$ is $\{\mathcal{N}(\mu, I), \mu \in \mathbf{R}^d\}$. A first observation is that the empirical mean has a breakdown point of 0. In particular, arbitrarily small corruptions (in total variation distance) to the underlying distribution can produce arbitrarily large errors in the mean. This agrees with the intuition that the empirical mean is highly nonrobust in the presence of outliers. A second example is that of the coordinate-wise median. It turns out that the coordinate-wise median has breakdown point of $1/2$ (which is the maximum possible) in any dimension d. This may suggest that the coordinate-wise median is *the most robust* mean estimator in high dimensions. On the other hand,

https://doi.org/10.1017/9781108943161.001 Published online by Cambridge University Press

it is not difficult to construct examples where the coordinate-wise median will have ℓ_2-distance of $\Omega(\epsilon\sqrt{d})$ from the true mean. A third example is that of the Tukey median. Recall that the Tukey median is known to have ℓ_2-error of $O(\epsilon)$ from the true mean (which is information-theoretically best possible). On the other hand, for Gaussians in $d \geq 2$ dimensions and additive contamination, its breakdown point can be shown to be equal to $1/3$ (see Exercise 1.8). In particular, it would be considered inferior to the coordinate-wise median with respect to this criterion. In essence, the breakdown point is a measure of how many corruptions an estimator can deal with before it becomes totally useless. However, if one cares about the size of the errors that one incurs (more precisely than simply knowing whether or not they are finite), the breakdown point will be an insufficient measure of robustness.

1.7 Exercises

1.1 (Definitions of Total Variation Distance) Prove that the different formulations of total variation distance given in Definition 1.4 are equivalent.

1.2 (Contamination Models) In this exercise, we will compare three contamination models – the strong contamination model, the total variation distance model, and the Huber contamination model – in terms of the difficulty they impose on a learner. In particular, we say that error model A can *simulate* error model B for sample size N if for every strategy the adversary for error model B can employ to corrupt a set of N samples, an adversary for error model A can employ a corresponding strategy, so that the distributions over sets of samples received by the algorithm are close in total variation distance.

(a) Show that if error model A can simulate error model B for sample size N, then any learning algorithm that works against corruptions of type A will also work against corruptions of type B. In particular, this shows that corruptions of type B are weaker than corruptions of type A.

(b) Show that for any $\epsilon' > \epsilon > 0$, the strong contamination model with the ability to corrupt an ϵ'-fraction of samples can simulate the ϵ-total variation distance error model over N samples, for any N a sufficiently large function of ϵ, ϵ'.

(c) Show that the ϵ-error total variation distance contamination model can simulate the ϵ-error Huber contamination model for any number of samples.

https://doi.org/10.1017/9781108943161.001 Published online by Cambridge University Press

1.3 (Precise Limits of Robust Learning) Here we will show that Proposition 1.7 is tight in the following sense: Let X and Y be two given probability distributions with $d_{\mathrm{TV}}(X, Y) = \delta$, for some $\delta > 0$. Let D be a distribution known to be either X or Y. An algorithm is given corrupted samples from D and is asked to determine whether $D = X$ or $D = Y$. Show that it can reliably make this determination with a bounded number of samples if:

(a) The algorithm is given samples with ϵ-additive contamination and $\delta > \epsilon/(1 - \epsilon)$.

(b) The algorithm is given samples with ϵ-subtractive contamination and $\delta > \epsilon$.

(c) The algorithm is given samples with ϵ-general contamination and $\delta > 2\epsilon$.

(Hint: Note that there is a set S so that $|X(S) - Y(S)| = \delta$. Consider the fraction of samples that lie in S.)

1.4 (Robustness of the Median)

(a) We showed in this chapter that the median is a robust mean estimator for $\mathcal{N}(\mu, 1)$ with error $O(\epsilon)$. What is the optimal constant factor in the $O(\cdot)$ for small ϵ?

(b) A distribution D on $\mathbf{R}$ with mean $\mu \in \mathbf{R}$ is called (s, ϵ)-smooth, where $\epsilon > 0$ and $s = s(\epsilon)$, if it satisfies $\Pr_{X \sim D}[X \geq \mu + s] \leq 1/2 - \epsilon$ and $\Pr_{X \sim D}[X \leq \mu - s] \leq 1/2 - \epsilon$. Show that given a sufficiently large ϵ'-corrupted set T of samples from D (for some $\epsilon > \epsilon' > 0$), the median of T is a robust estimator of the mean μ with error at most s.

(c) Construct a one-dimensional distribution D with sub-Gaussian tails such that the median of D does not perform well as a robust mean estimator. What about a symmetric distribution D?

1.5 (Robust Mean Estimation Under Bounded kth Moment) Let D be a distribution on $\mathbf{R}$ with bounded kth moment, for some $k \geq 2$. That is, D satisfies $\mathbf{E}_{X \sim D}[|X - \mu|^k] \leq \sigma^k$, for some known parameter $\sigma > 0$ and positive integer k, where μ is the mean of D.

(a) Show that the truncated mean of D is a robust mean estimator of the mean μ with error $O(\sigma\epsilon^{1-1/k})$.

(b) Show that the bound from part (a) is minimax optimal by proving Lemma 1.12. In particular, show that if an algorithm is given an ϵ-corrupted set of samples (in the Huber model) from a distribution D guaranteed to have bounded kth moments in the above sense (and

https://doi.org/10.1017/9781108943161.001 Published online by Cambridge University Press

for which nothing else is known), it is information-theoretically impossible to learn the mean of D within error better than $\Omega(\sigma\epsilon^{1-1/k})$ with more than $2/3$ probability of success.

1.6 (Robust Mean Estimation for Log-Concave Distributions) Let D be a log-concave distribution on $\mathbf{R}$ with standard deviation at most 1. A standard result about such distributions tells us that D has sub-exponential tails, in the sense that the probability that a sample from D is at distance more than t from its mean is $O(\exp(-\Omega(t)))$.

(a) Show that the truncated mean of D is a robust mean estimator of the mean μ with error $O(\epsilon \log(1/\epsilon))$.

(b) Show that the bound from part (a) is minimax optimal by proving Lemma 1.10. In particular, show that if an algorithm is given an ϵ-corrupted set of samples (even in the Huber model) from a distribution D guaranteed to be log-concave with variance at most 1 (and for which nothing else is known), it is information-theoretically impossible to learn the mean of D within error better than $\Omega(\epsilon \log(1/\epsilon))$ with more than $2/3$ success probability.

1.7 (Obliviousness to Contamination Parameter) Note that, in contrast to the median, the truncated mean requires a priori knowledge of the contamination parameter $\epsilon > 0$. In this problem, we will explore to what extent this can be avoided.

(a) Let D be a distribution on $\mathbf{R}$ with variance at most $\sigma > 0$, where σ is a known parameter. Consider the following estimator for the mean μ of D: Draw n ϵ-corrupted points from D, where $n \gg 1/\epsilon^2$. Let $X_1 \leq X_2 \leq \cdots \leq X_n$ be an ordering of these points. Find the minimum $1 \leq a \leq n/2$ such that the subsequence $X_a \leq X_{a+1} \leq \cdots \leq X_{n+1-a}$ has empirical variance at most 3σ. Output the empirical mean of $\{X_a, X_{a+1}, \ldots, X_{n+1-a}\}$. Show that this gives a robust estimator of μ with error $O(\sigma\sqrt{\epsilon})$.

(b) Let D be a distribution on $\mathbf{R}$ with variance at most $\sigma > 0$, where σ is *unknown*. Show that it is information-theoretically impossible to robustly estimate the mean of D without a priori knowledge of the contamination parameter $\epsilon > 0$. In particular, show that even given an unlimited number of samples, no algorithm that does not know either ϵ or σ can learn the mean of D to error $O(\sigma\sqrt{\epsilon})$ with probability $2/3$.

(c) Let D be a distribution on $\mathbf{R}$ with bounded kth moment, for some $k \geq 2$ in the sense of Problem 1.5. Design an algorithm that learns

https://doi.org/10.1017/9781108943161.001 Published online by Cambridge University Press

the mean of D to error $O(\sigma\epsilon^{1-1/k})$ in the presence of an ϵ-fraction of outliers without knowing ϵ.

1.8 (Breakdown Point Computations)

(a) Show that the breakdown point of the median of a continuous, one-dimensional distribution is $1/2$.

(b) Show that the breakdown point of the Tukey median of a two-dimensional Gaussian with additive contamination is at most $1/3$.

(c) Show that the breakdown point of the Tukey median of any symmetric, continuous distribution with respect to additive contamination is at least $1/3$.

(d) Show that the breakdown point of the Tukey median of any symmetric, continuous distribution with respect to total variation contamination is at most $1/4$.

1.9 (Estimation Accuracy with Corruption Rate Close to 1/2) In this exercise, we examine what happens to the error rates for robust mean estimation problems when the fraction of outliers ϵ is close to $1/2$ (note that when *equals* $1/2$, mean estimation is usually impossible, by Proposition 1.7).

(a) Let $X = \mathcal{N}(\mu, 1) \in \mathbf{R}$ be a Gaussian with unknown mean μ. Show that if one is given sufficiently many samples from X with ϵ-general contamination for some $\epsilon < 1/2$, the empirical median estimates μ to error $O(\sqrt{\log(1/(1/2-\epsilon))})$ with high probability.

(b) Show that the bound in part (a) is best possible in the sense that no algorithm given such ϵ-corrupted samples can reliably learn μ to error $o(\sqrt{\log(1/(1/2-\epsilon))})$ as ϵ approaches $1/2$.

(c) Let $X \in \mathbf{R}$ be a distribution with variance at most 1 and unknown mean μ. Show that if one is given sufficiently many samples from X with ϵ-general contamination for some $\epsilon < 1/2$, an appropriate truncated mean can approximate μ to error $O(1/\sqrt{1/2-\epsilon})$ with high probability.

(d) Show that the bound in part (c) is best possible in the sense that no algorithm given such ϵ-corrupted samples can reliably learn μ to error $o(1/\sqrt{1/2-\epsilon})$, as ϵ approaches $1/2$.

1.10 (High Probability Mean Estimation) Estimation problems for heavy-tailed distributions exhibit many of the same difficulties that problems of estimating with adversarial noise do. These issues become particularly clear when we want to construct estimators with very small probability of error. In this exercise, we explore some of these connections.

https://doi.org/10.1017/9781108943161.001 Published online by Cambridge University Press

(a) Consider the sample mean as an estimator of the mean of a one-dimensional Gaussian $\mathcal{N}(\mu, \sigma^2)$. Show that for any $\delta \in (0, 1)$, given n i.i.d. samples from $\mathcal{N}(\mu, \sigma^2)$, with probability $1 - \delta$, the sample mean has distance $O(\sigma \sqrt{\log(1/\delta)/n})$ from the true mean.

(b) Show that the sample mean does not work well for general distributions with bounded variance. In particular, for $n \in \mathbf{Z}_+$, $\delta \in (0, 1)$, show that there is a one-dimensional distribution X with standard deviation at most σ such that with probability at least δ the empirical mean of X computed from n i.i.d. samples differs from the true mean by at least $\Omega(\sigma \sqrt{1/(n\delta)})$.

(c) Show that the rate from part (b) can be improved by taking an appropriate truncated mean. In particular, given n a positive integer and $\delta \in (0, 1)$, design an estimator that given n i.i.d. samples from a distribution X with standard deviation at most σ produces an estimator that is within distance $O(\sigma \sqrt{\log(1/\delta)/n})$ of the true mean of X with probability at least $1 - \delta$.

Hint: You may need to make use of Bernstein's Inequality (Theorem A.7) in order to prove this.

(d) Show that the estimator from part (c) can be made robust to contamination. In particular, in the presence of an ϵ-fraction of adversarial errors, this estimator can be modified to achieve error $O(\sigma \sqrt{\log(1/\delta)/n} + \sigma \sqrt{\epsilon})$ with probability at least $1 - \delta$.

1.11 (Robustness of Geometric Median) For a finite set $S \subset \mathbf{R}^d$, define its *geometric median* to be the point $x \in \mathbf{R}^d$ minimizing $\sum_{y \in S} \|x - y\|_2$. Let S be an ϵ-corrupted set of samples from $\mathcal{N}(\mu, I) \in \mathbf{R}^d$ of sufficiently large size.

(a) Show that the geometric median of S has ℓ_2-distance $O(\epsilon \sqrt{d})$ from μ with high probability.

(b) Show that this upper bound is tight for a worst-case adversary.

1.12 (Sample-Efficient Robust Estimation) Use the methodology we introduced to establish Proposition 1.20 to obtain robust (and computationally inefficient) estimators for the following tasks:

(a) Estimating the mean of a distribution $X \in \mathbf{R}^d$ with bounded kth moments. In particular, show that if the kth moment of X is at most σ^k in any direction, there is an estimator that approximates the mean of X to error $O(\epsilon^{1-1/k}\sigma)$ from ϵ-corrupted samples with high probability.

(b) Sparse mean estimation of $\mathcal{N}(\mu, I)$. Here the goal is to estimate the mean μ under the assumption that it is k-sparse, that is, if μ is

https://doi.org/10.1017/9781108943161.001 Published online by Cambridge University Press

supported on an unknown subset of k coordinates. The sample complexity should depend polynomially on k, but only logarithmically on the underlying dimension.

(c) Estimating the covariance of $\mathcal{N}(0, \Sigma)$ under the assumption that $\Sigma \preceq I$. Specifically, find a $\widehat{\Sigma}$ that is close to Σ in spectral norm. What about Frobenius norm?

1.8 Discussion and Related Work

The traditional approach in statistics is to design estimators that perform well under the assumption that the underlying observations are i.i.d. samples from a model of interest. Robust statistics aims to design estimators that are insensitive or stable against small deviations from this classical assumption. That is, a small change in the underlying distribution should result in a small change in the performance of the estimator. Two closely related approaches of quantifying the deviation from the standard i.i.d. assumption involve outlying observations or model misspecification.

As a subfield of Statistics, Robust Statistics was initiated in the pioneering works of [140], [3], and [95]. The latter work introduced the contamination model of Definition 1.1. More general contamination models, with respect to other metrics, were studied in [83]. The reader is referred to some early introductory textbooks from the statistics community [85, 97]. The quote by Peter Huber given in the introduction of this chapter is from Chapter 8 of [96].

Early work in the robust statistics community focused on the sample complexity of robust estimation and on the notion of the breakdown point [72, 73, 84]. Interestingly, recent work in robust statistics [26] advocates that achieving robustness under Huber contamination is more general than achieving large breakdown point, and provides a unified way of studying robustness.

The Tukey median was defined by [141]. It is known that in the presence of ϵ-contamination, when the inlier data is drawn from an unknown mean and identity covariance Gaussian, the Tukey median achieves the optimal robustness of $O(\epsilon)$. The same guarantee holds for other symmetric distributions as well; see, for example, [26]. Several other depth functions have been studied in the relevant statistics literature [26, 133, 135]. Unfortunately, the Tukey median is NP-hard to compute in general [103] and the many heuristics proposed to approximate it degrade in the quality of their approximation as the dimension scales. Similar hardness results have been shown [15, 86] for essentially all known classical estimators in robust statistics.

In recent years, learning in the presence of outliers has become a pressing challenge in a number of high-dimensional data analysis applications. These

https://doi.org/10.1017/9781108943161.001 Published online by Cambridge University Press

include the analysis of biological datasets, where natural outliers are common [116, 124, 132] and can contaminate the downstream statistical analysis, and *data poisoning attacks* in machine learning [14], where even a small fraction of fake data (outliers) can substantially degrade the quality of the learned model [19, 137]. In the following chapters of this book, we develop a general algorithmic theory that leads to computationally efficient estimators for a wide range of high-dimensional estimation tasks, including the mean estimation task considered in this chapter. These efficient estimators have led to practical improvements in the analysis of genetic data [46] and in adversarial machine learning [44, 88, 139].

https://doi.org/10.1017/9781108943161.001 Published online by Cambridge University Press