2

Efficient High-Dimensional Robust Mean Estimation

2.1 Introduction

In Chapter 1, we analyzed some standard efficient robust estimators for the one-dimensional setting and discussed the information-theoretic aspects of basic robust statistics problems in any dimension. Unfortunately, in high dimensions, the methods discussed in that chapter are inherently unsatisfactory. In particular, these approaches either incur runtime exponential in the dimension or lead to error that scales polynomially in the dimension. In fact, over several decades, this dichotomy persisted in all known algorithms for even the most basic high-dimensional unsupervised problems in the presence of adversarial outliers. The first algorithmic progress in this direction was made in the context of high-dimensional robust mean estimation for Gaussians and other well-behaved distributions. These developments form the basis for essentially all algorithms in this book. Thus, it is natural for our discussion on algorithmic high-dimensional robust statistics to begin there.

Recall that in order for robust mean estimation to be at all possible, one needs to make some assumptions on the behavior of the inlier distribution X. As we will see, these assumptions usually amount to certain concentration properties. While many of the algorithms we present work for distributions with only weak assumptions of this form (e.g., bounded covariance), the basic case of Gaussians with identity covariance (i.e., distributions of the form $X \sim \mathcal{N}(\mu, I)$, for some unknown mean μ) is particularly illuminating. As such, many of our motivating examples will be specific to this case.

2.1.1 Key Difficulties and High-Level Intuition

Arguably, the most natural attempt at robustly estimating the mean of a distribution would be to identify the outliers and output the empirical mean of

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

the remaining points. The key difficulty in high dimensions is the fact that the outliers cannot be identified at an individual level, even when they move the mean significantly. In a number of cases, we can easily identify the "extreme outliers" via a pruning procedure exploiting the concentration properties of the inliers. Alas, such naive approaches typically do not suffice to obtain nontrivial error guarantees.

The simplest example illustrating this difficulty is that of a high-dimensional spherical Gaussian. Typical samples will be at ℓ_2-distance approximately $\Theta(\sqrt{d})$ from the true mean, where d is the dimension. Given this, we can apply a kind of basic, "naive filtering" by removing all points at Euclidean distance more than $10\sqrt{d}$ from the coordinate-wise median. It is not hard to see that only a tiny fraction of inliers will be removed by this procedure, while all of the sufficiently extreme outliers will be. Unfortunately, it is difficult to remove much else by this kind of procedure. In particular, since any point at distance approximately $\sqrt{d}$ from the mean is just as likely to appear as any other, none of them can safely be eliminated without risking the removal of inliers as well. However, if an ϵ-fraction of outliers are placed at distance $\sqrt{d}$ *in roughly the same direction* from the unknown mean (see Figure 2.1), an adversary can corrupt the sample mean by as much as $\Omega(\epsilon\sqrt{d})$.

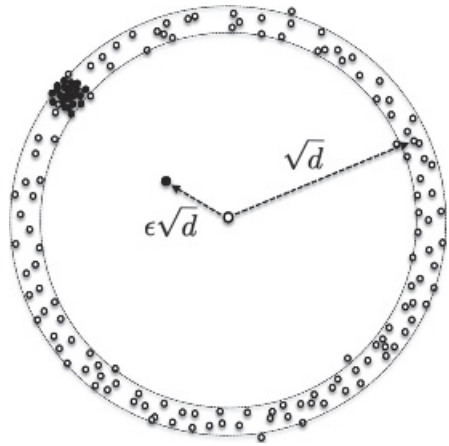

Figure 2.1 A hard instance for naive filtering. Note that the inlier samples (white) for a high-dimensional spherical Gaussian are concentrated in a spherical shell of distance approximately $\sqrt{d}$ from the mean. If the outliers (black) are placed within this shell, they will be difficult to detect. Moreover, if the outliers are all placed in roughly the same location in the shell, they can corrupt the mean by as much as $\epsilon\sqrt{d}$.

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

This leaves the algorithm designer with a dilemma of sorts. On the one hand, potential outliers at distance $\Theta(\sqrt{d})$ from the unknown mean could lead to large ℓ_2-error, scaling polynomially with d. On the other hand, if the adversary places outliers at distance approximately $\Theta(\sqrt{d})$ from the true mean in *random directions*, it may be information-theoretically impossible to distinguish them from the inliers. The way out is the realization that, in order to obtain a robust estimate of the mean, *it is in fact not necessary to detect and remove all outliers*. It is only required that the algorithm can detect the "consequential outliers," that is, the ones that can significantly impact our estimates of the mean.

So how can we make progress? To begin with, let us assume that there are no extreme outliers (as these can be removed via naive filtering). Then we claim that *the only way that the empirical mean can be far from the true mean is if there is a "conspiracy" of many outliers, all producing errors in approximately the same direction.* Intuitively, if our corrupted points are at distance $O(\sqrt{d})$ from the true mean in random directions, their contributions will on average cancel out, leading to a small error in the sample mean. In conclusion, it suffices to be able to detect these kinds of conspiracies of outliers.

The next key insight is simple and powerful. Let T be an ϵ-corrupted set of points drawn from $\mathcal{N}(\mu, I)$. If such a conspiracy of outliers substantially moves the empirical mean μ_T of T, it must move μ_T in some direction. That is, there is a unit vector v such that these outliers cause $v \cdot (\mu_T - \mu)$ to be large. For this to happen, it must be the case that these outliers are on average far from μ in the v-direction. In particular, if an ϵ-fraction of corrupted points in T move the sample average of $v \cdot (U_T - \mu)$, where U_T is the uniform distribution on T, by more than δ (δ should be thought of as small, but substantially larger than ϵ), then on average these corrupted points x must have $v \cdot (x - \mu)$ at least δ/ϵ, as shown in Figure 2.2. This in turn means that these corrupted points will have a contribution of at least $\epsilon \cdot (\delta/\epsilon)^2 = \delta^2/\epsilon$ to the variance of $v \cdot U_T$. Fortunately, this condition can actually be algorithmically detected! In particular, by computing the top eigenvector of the sample covariance matrix, we can efficiently determine whether or not there is any direction v for which the variance of $v \cdot U_T$ is abnormally large.

The aforementioned discussion leads us to the overall structure of the algorithms we will describe in this chapter. Starting with an ϵ-corrupted set of points T (perhaps weighted in some way), we compute the sample covariance matrix and find the eigenvector v^* with largest eigenvalue λ^*. If λ^* is not much larger than it should be (in the absence of outliers), by the above discussion, the empirical mean is close to the true mean, and we can return that as an answer. Otherwise, we have obtained a particular direction v^* for which we know

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

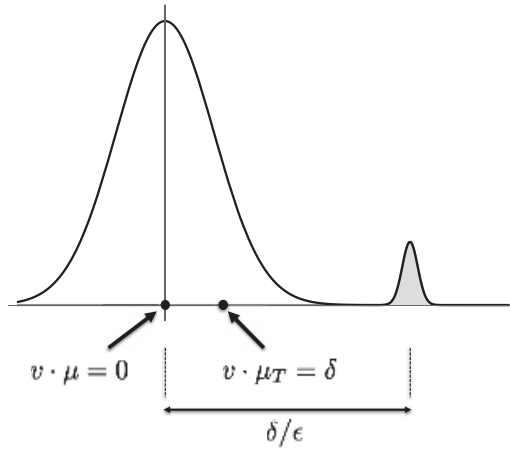

Figure 2.2 An example of an ϵ-fraction of outliers changing the empirical mean of T by δ in the v-direction. The graph represents the projections of the samples onto the v-direction. Notice that the errors must on average have $v \cdot x$ at least δ/ϵ-far from $v \cdot \mu$. This means that they must contribute at least δ^2/ϵ to the variance of $v \cdot T$.

that the outliers play an unusual role, that is, the outliers behave significantly differently than the inliers. The distribution of the points projected in the v^*-direction can then be used to perform some sort of outlier removal. As to how exactly to perform this outlier removal step, there are several different techniques that we will discuss, some of which depend on particular features of the inliers.

2.2 Stability and Robust Mean Estimation

In the strong contamination model, we begin by drawing a set S of n independent samples from the true distribution. We will typically call these uncorrupted sample points *inliers*. The adversary can then select up to an ϵ-fraction of these points, changing them arbitrarily and giving our algorithm a new dataset T to work with.

For our algorithm to succeed, we want it to satisfy the fairly strong requirement that with high probability over the set S of inliers, no matter what corruptions the adversary decides to make, our algorithm when run on T will output a good approximation to the target parameter. To prove such a statement, we typically want to define a deterministic condition on S under which our algorithm is guaranteed to succeed. In particular, we will require a condition on the set of uncorrupted samples such that:

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

1. A sufficiently large collection S of independent samples from our inlier distribution satisfies this condition with high probability.
2. If the set S of inliers satisfies this condition, our algorithm will succeed when run on T no matter what corruptions the adversary chooses to apply.

Toward this goal, we introduce a condition called *stability* (Definition 2.1), which will form the core of our conditions on the uncorrupted samples. In particular, we will show that if the uncorrupted samples are assumed to be stable, then there is an efficiently checkable condition on the ϵ-corrupted dataset that will imply that the true mean is close to the sample mean of the corrupted (i.e., including the outliers) dataset (see Lemma 2.6). Although some algorithms presented later in this chapter (and in this book) may require stronger conditions on their inliers in order to be effective, some version of this stability condition will almost always be involved.

The robust mean estimation algorithms in this chapter will depend heavily on computing means and covariances of various sets of samples. Although additive corruptions will always have the power to induce large changes in the sample mean and covariance, we will at least want to know that any large set of inliers has close to the right value. It is this requirement that makes up the core of the stability condition.

Definition 2.1 (Stability Condition) Fix $0 < \epsilon < 1/2$ and $\delta \geq \epsilon$. A finite set $S \subset \mathbf{R}^d$ is (ϵ, δ)-*stable* (with respect to a vector μ or a distribution X with $\mu_X := \mathbf{E}[X] = \mu$) if for every unit vector $v \in \mathbf{R}^d$ and every $S' \subseteq S$ with $|S'| \geq (1-\epsilon)|S|$, the following conditions hold:

1. $\left|\frac{1}{|S'|}\sum_{x\in S'} v \cdot (x-\mu)\right| \leq \delta$, and
2. $\left|\frac{1}{|S'|}\sum_{x\in S'} (v\cdot(x-\mu))^2 - 1\right| \leq \delta^2/\epsilon$.

Similarly, we say that a *distribution* X on $\mathbf{R}^d$ is (ϵ, δ)-stable with respect to a vector μ if for every unit vector $v \in \mathbf{R}^d$ and distribution X' obtained from X by ϵ-subtractive contamination, the following conditions hold:

1. $|\mathbf{E}[v \cdot (X' - \mu)]| \leq \delta$, and
2. $\left|\mathbf{E}[(v\cdot(X'-\mu))^2] - 1\right| \leq \delta^2/\epsilon$.

Some comments are in order. The first condition (for the finite set stability condition) is equivalent to $\|\mu_{S'} - \mu\|_2 \leq \delta$, where $\mu_{S'}$ is the empirical mean of S'. The second condition is equivalent to $\|\bar{\Sigma}_{S'} - I\|_2 \leq \delta^2/\epsilon$, where $\bar{\Sigma}_{S'} = (1/|S'|)\sum_{x\in S'}(x-\mu)(x-\mu)^\top$ is the empirical second moment matrix of S' with respect to μ. Since μ is close to $\mu_{S'}$ by the first condition, this is equivalent (up to changing δ by a constant factor) to saying that $\|\mathbf{Cov}[S'] - I\|_2 = O(\delta^2/\epsilon)$.

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

In other words, *removing any ϵ-fraction of the points will not change the mean by more than δ nor the variance in any direction by more than δ^2/ϵ.*

It is also worth nothing that Definition 2.1 is intended for distributions X with covariance $\Sigma_X \preceq I$. If one wants to perform robust mean estimation for distributions X with other covariance matrices, one can usually reduce to this case by applying the linear transformation $x \to \Sigma_X^{-1/2} x$ to the data.

Finally, it is worth comparing the notions of stability for finite sets and distributions. While these definitions are fairly similar, we believe it is important to include both, as it is sometimes more convenient to work with one or the other. The close relationship between these two definitions will also be important to us, and can be made rigorous via the following simple lemma.

Lemma 2.2 *If S is a set of points in $\mathbf{R}^d$ and $\delta > \epsilon > 0$ with $\epsilon|S|$ an integer, then S is (ϵ, δ)-stable with respect to some vector μ if and only if the uniform distribution over S is (ϵ, δ)-stable with respect to μ.*

Proof The "only if" part here is immediate, since if S' is a subset $S' \subseteq S$ with $|S'| \geq (1-\epsilon)|S|$, then the uniform distribution over S' can be obtained from the uniform distribution over S by ϵ-subtractive contamination. To show the reverse, we note that for a specific choice of unit vector v, if one wants to find a distribution X' for which one of conditions 1 or 2 above does not hold, one will want to remove the ϵ-fraction of the distribution on which $v \cdot (X - \mu)$ or $(v \cdot (X' - \mu))^2 - 1$ takes its most extreme values. This is equivalent to throwing away some $\epsilon|S|$ points from the support, but if S is stable, this will be insufficient to change the mean or variance by enough. □

Although Lemma 2.2 only applies when $\epsilon|S|$ is an integer, by combining it with the results of Exercise 2.1, we find that, so long as $|S| \geq 1/\epsilon$, S is (ϵ, δ)-stable if and only if the uniform distribution over S is $(\epsilon, \Theta(\delta))$-stable.

The fact that the conditions of Definition 2.1 must hold *for every* large subset S' of S might make it unclear if they can hold with high probability. It can in fact be shown that these conditions are satisfied for various distribution classes with appropriate concentration properties. Morally speaking, if a distribution X is stable, then we would expect a large enough set S of i.i.d. samples from X to be stable (with comparable parameters) with high probability.

2.2.1 Sample Complexity Bounds for the Stability Condition

Before we explain how to leverage stability for the design of computationally efficient algorithms, we show that for some natural distributions the stability of the set of inliers can be achieved with high probability given a reasonable

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

number of i.i.d. samples. The sample complexity bounds presented in this section are intentionally rough; the reader interested in more precise bounds is referred to Section 3.2.

We start with the class of sub-Gaussian distributions. Recall that a distribution on $\mathbf{R}^d$ is sub-Gaussian if any univariate projection has sub-Gaussian tails. For this distribution class, we can show:

Proposition 2.3 *If N is at least a sufficiently large degree polynomial in d/ϵ, then a set of N i.i.d. samples from an identity covariance sub-Gaussian distribution in $\mathbf{R}^d$ is $(\epsilon, O(\epsilon\sqrt{\log(1/\epsilon)}))$-stable with high probability.*

In order to see why this is the correct value of δ, we note that the Gaussian distribution, $X = \mathcal{N}(\mu, I)$, is $(\epsilon, O(\epsilon\sqrt{\log(1/\epsilon)}))$-stable with respect to μ. This is because removing an ϵ-fraction of the mass will have the greatest impact on $\mathbf{E}[v\cdot X]$ or $\mathbf{Var}[v\cdot X]$ if we remove the ϵ-tails of $v\cdot X$. A simple calculation shows that this affects the mean by $O(\epsilon\sqrt{\log(1/\epsilon)})$ and the variance by $O(\epsilon\log(1/\epsilon))$. This is because the ϵ-tails of the distribution are $O(\sqrt{\log(1/\epsilon)})$ far from the mean.

To help formalize this intuition, we provide a proof sketch of Proposition 2.3 here. It turns out that the optimal sample complexity in Proposition 2.3 is $\tilde{\Theta}(d/\epsilon^2)$. The reader is referred to Section 3.2 for the proof of this optimal bound.

Proof Sketch. An easy way to prove this result is by noting that it suffices for our dataset S to have the empirical distribution of $v \cdot S := \{v \cdot x, x \in S\}$ mimic the real distribution $v \cdot X$ for all unit vectors v. To formalize this, we consider thresholds. In particular, we would like it to hold for every vector v and every threshold $t \in \mathbf{R}$ that

$$|\mathbf{Pr}_{x\sim_u S}[v \cdot x > t] - \mathbf{Pr}_{x\sim X}[v \cdot x > t]| \tag{2.1}$$

should be small. By the VC Inequality (Theorem A.12), the error in Equation (2.1) is never more than η with high probability, as long as N is at least a sufficiently large constant multiple of d/η^2.

Note that the average value of $v \cdot (x - \mu)$ or $(v \cdot (x - \mu))^2$ can be computed from these tail probabilities as

$$\frac{1}{|S|}\sum_{x\in S} v \cdot (x - \mu) = \int_{v\cdot\mu}^{\infty} \mathbf{Pr}_{x\sim_u S}[v \cdot x > t]dt - \int_{-\infty}^{v\cdot\mu} \mathbf{Pr}_{x\sim_u S}[v \cdot x < t]dt,$$

and

$$\frac{1}{|S|}\sum_{x\in S}(v \cdot (x - \mu))^2 = \int_0^{\infty} 2\,t\,\mathbf{Pr}_{x\sim_u S}[|v \cdot x - v \cdot \mu| < t]dt.$$

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

Knowing that each probability above is within $O(\eta)$ of the corresponding probability for $x \sim X$ is *almost* sufficient to show that the mean and covariance of S are close to μ and $\mathbf{Cov}[X] = I$, respectively. A slight technical difficulty, however, comes from the fact that these integrals have infinite range of t, and thus an $O(\eta)$ error for each given t produces an infinite error overall. We can fix this slight glitch by noting that with high probability each $x \in S$ satisfies $\|x - \mu\|_2 < O(\sqrt{d \log(dN)})$ (for example, because each coordinate of $x - \mu$ is at most $O(\sqrt{\log(dN)})$). This observation allows us to truncate these integrals to ones of finite length and show that $\|\mu_S - \mu\|_2 = O(\eta \sqrt{d \log(dN)})$ and $\|\mathbf{Cov}[S] - I\|_2 = O(\eta d \log(dN))$.

Having established good bounds on the mean and covariance of the full set S, we next need to prove a stronger statement. We actually need to bound these quantities for S', where S' is *any* $(1 - \epsilon)$-dense subset of S. To that end, we note that

$$\left|\mathbf{Pr}_{x \sim_u S}[v \cdot x > t] - \mathbf{Pr}_{x \sim_u S'}[v \cdot x > t]\right| \leq \min\{\mathbf{Pr}_{x \sim_u S}[v \cdot x > t], O(\epsilon)\}.$$

This inequality holds because removing elements can decrease a tail probability by ϵ, but cannot decrease it to less than 0. This allows us to bound the differences between the averages over S and S'. For example, we have that

$$\begin{aligned}
&\left|\frac{1}{|S|}\sum_{x \in S} v \cdot (x - \mu) - \frac{1}{|S'|}\sum_{x \in S'} v \cdot (x - \mu)\right| \\
&\leq \int_0^{O(\sqrt{d \log(dN)})} \min\{\mathbf{Pr}_{x \sim_u S}[v \cdot (x - \mu) > t],\ O(\epsilon)\} dt \\
&\quad + \int_{-O(\sqrt{d \log(dN)})}^{0} \min\{\mathbf{Pr}_{x \sim_u S}[v \cdot (x - \mu) < t],\ O(\epsilon)\} dt \\
&\leq \int_{-O(\sqrt{d \log(dN)})}^{O(\sqrt{d \log(dN)})} \min\{\exp(-\Omega(t^2)) + O(\eta),\ O(\epsilon)\} dt \\
&\leq O(\eta \sqrt{d \log(dN)}) + \int_{-O(\sqrt{\log(1/\epsilon)})}^{O(\sqrt{\log(1/\epsilon)})} O(\epsilon) dt + \int_{|t| \gg \sqrt{\log(1/\epsilon)}} \exp(-\Omega(t^2)) dt \\
&\leq O(\eta \sqrt{d \log(dN)}) + O(\epsilon \sqrt{\log(1/\epsilon)}).
\end{aligned}$$

This is $O(\epsilon \sqrt{\log(1/\epsilon)})$, assuming that η is sufficiently small. A similar argument can be used to bound the covariance term, and this completes our proof. □

Note that the proof of Proposition 2.3 essentially boiled down to an argument about the tail bounds of the distribution X. Morally speaking, if X is an

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

identity covariance distribution where the ϵ-tails in any direction contribute no more than δ to the mean and δ^2/ϵ to the variance in that direction, sufficiently many samples from X will be (ϵ, δ)-stable with high probability (see Exercise 2.4).

A more general setting considers inlier distributions with bounded and unknown covariance matrix. For this more general class of bounded covariance distributions, one can show the following.

Proposition 2.4 *Let S be a multiset of N i.i.d. samples from a distribution with covariance $\Sigma \preceq I$, where N is at least a sufficiently large degree polynomial in d/ϵ. With high probability, there exists a subset $S' \subseteq S$ of cardinality $|S'| \geq (1-\epsilon)|S|$ such that S' is $(\epsilon, O(\sqrt{\epsilon}))$-stable.*

It is worth pointing out a qualitative difference between Proposition 2.4 and its analogue Proposition 2.3 (for identity covariance sub-Gaussian distributions). For the bounded covariance case, a sufficiently large set of i.i.d. samples S from the inlier distribution is *not* guaranteed to be stable. On the other hand, Proposition 2.4 shows that there exists a $(1-\epsilon)$ density, stable subset S' (this still suffices for our purposes, as T, the set of corrupted samples, will be an $O(\epsilon)$-corruption of S'). This relaxation is necessary as there are simple examples where the proposition fails if we do not consider such subsets (see Exercise 2.3).

To see why Proposition 2.4 holds, we note that in order for a set S to be $(\epsilon, O(\sqrt{\epsilon}))$-stable with respect to μ, it suffices to check that $\|\mu_S - \mu\|_2 = O(\sqrt{\epsilon})$ and $\mathbf{Cov}[S] = O(I)$. We note that all but an ϵ-fraction of the mass of a bounded covariance distribution is within distance $O(\sqrt{d/\epsilon})$ of its mean μ. Moreover, if we throw away the points further away, this does not affect the mean by much. Letting S' be the set of samples not too far from the mean μ will have roughly the correct mean and covariance matrix with high probability.

The reader is referred to Section 3.2 for a proof of this result with the optimal sample complexity, which turns out to be $\tilde{\Theta}(d/\epsilon)$.

Remark 2.5 Analogous bounds can be shown for identity covariance distributions with bounded higher central moments. For example, if our distribution has identity covariance and its kth central moment, where $k \geq 4$, is bounded from above by a constant, it can be shown that a set of $\Omega(d \log(d)/\epsilon^{2-2/k})$ samples contains a large subset that is $(\epsilon, O(\epsilon^{1-1/k}))$-stable with high probability.

2.2.2 Stability and Algorithm Design

We now return to the use of the stability condition in algorithm design. In particular, we show how one can certify – under certain conditions – that the

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

sample mean of an ϵ-corrupted version of a stable set is a good approximation to the true mean. This is perhaps the most important property of stability for us and can be quantified in the following lemma.

Lemma 2.6 (Certificate for Empirical Mean) *Let S be an (ϵ, δ)-stable set with respect to a vector μ, for some $\delta \geq \epsilon > 0$ and $\epsilon \leq 1/3$. Let T be an ϵ-corrupted version of S. Let μ_T and Σ_T be the empirical mean and covariance of T. If the largest eigenvalue of Σ_T is at most $1 + \lambda$, for some $\lambda \geq 0$, then $\|\mu_T - \mu\|_2 \leq O(\delta + \sqrt{\epsilon\lambda})$.*

This lemma states that if our set of inliers S is stable and our set of corrupted samples T has bounded covariance, then the empirical mean of T is certifiably close to the true mean.

Lemma 2.6 follows by applying the following slightly more general statement to the uniform distribution over S.

Lemma 2.7 (Certificate for Empirical Mean, Strong Version) *Let X be an (ϵ, δ)-stable distribution with respect to a vector μ, for some $\delta \geq \epsilon > 0$ and $\epsilon \leq 1/3$. Let Y be a distribution with $d_{\mathrm{TV}}(X, Y) \leq \epsilon$ (i.e., Y is an ϵ-corrupted version of X). Denote by μ_Y and Σ_Y the mean and covariance of Y. If the largest eigenvalue of Σ_Y is at most $1 + \lambda$, for some $\lambda \geq 0$, then $\|\mu_Y - \mu\|_2 \leq O(\delta + \sqrt{\epsilon\lambda})$.*

Proof of Lemma 2.7 Let $Y = (1-\epsilon)X' + \epsilon E$ for some distribution X' obtained from X by ϵ-subtractive contamination. Let $\mu_X, \mu_{X'}, \mu_E$ and $\Sigma_X, \Sigma_{X'}, \Sigma_E$ denote the means and covariances of X, X', and E, respectively. A simple calculation gives

$$\Sigma_Y = (1-\epsilon)\Sigma_{X'} + \epsilon\Sigma_E + \epsilon(1-\epsilon)(\mu_{X'} - \mu_E)(\mu_{X'} - \mu_E)^\top.$$

Let v be the unit vector in the direction of $\mu_{X'} - \mu_E$. We have

$$\begin{aligned} 1 + \lambda \geq v^\top \Sigma_Y v &= (1-\epsilon)v^\top \Sigma_{X'} v + \epsilon v^\top \Sigma_E v + \epsilon(1-\epsilon)v^\top(\mu_{X'} - \mu_E)(\mu_{X'} - \mu_E)^\top v \\ &\geq (1-\epsilon)(1 - \delta^2/\epsilon) + \epsilon(1-\epsilon)\|\mu_{X'} - \mu_E\|_2^2 \\ &\geq 1 - O(\delta^2/\epsilon) + (\epsilon/2)\|\mu_{X'} - \mu_E\|_2^2, \end{aligned}$$

where we used the variational characterization of eigenvalues, the fact that Σ_E is positive semidefinite, and the second stability condition for X. By rearranging, we obtain $\|\mu_{X'} - \mu_E\|_2 = O(\delta/\epsilon + \sqrt{\lambda/\epsilon})$. Therefore, we can write

$$\begin{aligned} \|\mu_Y - \mu\|_2 &= \|(1-\epsilon)\mu_{X'} + \epsilon\mu_E - \mu\|_2 = \|\mu_{X'} - \mu + \epsilon(\mu_E - \mu_{X'})\|_2 \\ &\leq \|\mu_{X'} - \mu\|_2 + \epsilon\|\mu_{X'} - \mu_E\|_2 = O(\delta) + \epsilon \cdot O(\delta/\epsilon + \sqrt{\lambda/\epsilon}) \\ &= O(\delta + \sqrt{\lambda\epsilon}), \end{aligned}$$

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

where we used the first stability condition for X and our obtained upper bound on $\|\mu_{X'} - \mu_E\|_2$. □

Remark 2.8 It is worth noting that the proof of Lemma 2.7 only used the lower bound part in the second condition of Definition 2.1, namely, that the variance of X' in each direction is *at least* $1 - \delta^2/\epsilon$. Although this is sufficient for certifying that our mean is close, the corresponding upper bound will be crucially used in the design and analysis of our robust mean estimation algorithms in the following sections.

Lemma 2.6 says that if our input set of points T is an ϵ-corrupted version of any stable set S and has bounded covariance, the sample mean of T closely approximates the true mean of the original distribution. This lemma, or a variant thereof, is a key result in all known robust mean estimation algorithms.

Unfortunately, we are not always guaranteed that the set T we are given has this property. In particular, if the corrupted set T includes some large outliers, or many outliers in the same direction, there may well be directions of large variance. In order to deal with this, we will want to compute a subset, T', of T such that T' has bounded covariance and large intersection with S. If we can achieve this, then since T' will be a corrupted version of S with bounded covariance, we can apply Lemma 2.6 to show that $\|\mu_{T'} - \mu\|_2$ is small.

For some of the algorithms presented, it will be convenient to find a probability distribution over T rather than a subset. For these cases, we can use Lemma 2.7 applied to the appropriate distribution on T.

For the more general outlier removal procedure, we are given our initial ϵ-corrupted set T, and we will attempt to find a distribution W supported on T such that the "weighted" covariance matrix Σ_W has no large eigenvalues. For such a solution, the weight $W(x)$ of an $x \in T$ can be thought of as quantifying our belief about whether point x is an inlier or an outlier. It will also be important for us to ensure that W is close to the uniform distribution over S in total variation distance. This is complicated by the fact that we must be able to guarantee this closeness without knowing exactly what the set S is. Intuitively, we can do this by ensuring that W is obtained by removing at most ϵ mass from the uniform distribution over T.

More concretely, the following general framework can be used for robust mean estimation.

Definition 2.9 For a finite set T and $\epsilon \in (0, 1)$, we will denote by Δ^T the set of all probability distributions W supported on T, whose probability mass function $W(x)$ satisfies $W(x) \leq \frac{1}{|T|(1-\epsilon)}$, for all $x \in T$.

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

Lemma 2.10 *Let S be a $(3\epsilon, \delta)$-stable set with respect to μ and let T be an ϵ-corrupted version of S for some $\epsilon < 1/6$. Given any $W \in \Delta^T$ such that $\|\Sigma_W\|_2 \le 1 + \lambda$, for some $\lambda \ge 0$, we have $\|\mu - \mu_W\|_2 = O(\delta + \sqrt{\epsilon\lambda})$.*

Proof We note that *any* distribution in Δ^T differs from U_S, the uniform distribution on S, by at most 3ϵ. Indeed, for $\epsilon \le 1/3$, we have

$$\begin{aligned} d_{\mathrm{TV}}(U_S, W) &= \sum_{x \in T} \max\{W(x) - U_S(x), 0\} \\ &= \sum_{x \in S \cap T} \max\{W(x) - 1/|T|, 0\} + \sum_{x \in T \setminus S} W(x) \\ &\le \sum_{x \in S \cap T} \frac{\epsilon}{|T|(1-\epsilon)} + \sum_{x \in T \setminus S} \frac{1}{|T|(1-\epsilon)} \\ &\le |T| \left(\frac{\epsilon}{|T|(1-\epsilon)} \right) + \epsilon |T| \left(\frac{1}{|T|(1-\epsilon)} \right) \\ &= \frac{2\epsilon}{1-\epsilon} \le 3\epsilon. \end{aligned}$$

Therefore, by Lemma 2.7 we have $\|\mu - \mu_W\|_2 = O(\delta + \sqrt{\epsilon\lambda})$. □

Lemma 2.10 provides us with a clear plan for how to perform robust mean estimation. Given a set T (promised to be an ϵ-corruption of a $(3\epsilon, \delta)$-stable set), we merely need to find a $W \in \Delta^T$ with bounded covariance matrix.

A natural first question is whether such a distribution W exists. Fortunately, this can be easily guaranteed. In particular, if we take W to be W^*, the uniform distribution over $S \cap T$, the largest eigenvalue is at most $1 + \delta^2/\epsilon$ by the stability of S. Thus, for this choice of W, we can take $\lambda = \delta^2/\epsilon$, and we have $\|\mu - \mu_{W^*}\|_2 = O(\delta)$.

At this point, we have an *inefficient* algorithm for approximating μ: Find *any* $W \in \Delta^T$ with Σ_W bounded above by $(1 + \delta^2/\epsilon)I$ and return its mean. The remaining question is *how we can efficiently find* such a W. There are two basic algorithmic techniques to achieve this, which we present in the subsequent sections.

The first algorithmic technique we will describe is based on convex programming. We will call this *the unknown convex programming method*. Note that Δ^T is a convex set and that finding a point in Δ^T that has bounded covariance is *almost* a convex program. It is not quite a convex program because the variance of $v \cdot W$, for fixed v, is not a convex function of W. However, one can show that given a W with variance in some direction significantly larger than $1 + \delta^2/\epsilon$, we can efficiently construct a hyperplane separating W from W^* (the uniform distribution over $S \cap T$). This method works naturally under only the

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

stability condition. On the other hand, as it relies on the ellipsoid algorithm, it is quite slow (although polynomial time). See Section 2.3 for more details.

Our second technique, which we will call *(iterative) filtering*, is an iterative outlier removal method that is typically faster, as it relies only on spectral techniques. The main idea of the method is the following: If Σ_W does not have large eigenvalues, then the empirical mean is close to the true mean. Otherwise, there is some unit vector v such that $\mathbf{Var}[v \cdot W]$ is substantially larger than it should be. This can only be the case if W assigns substantial mass to elements of $T \setminus S$ that have values of $v \cdot x$ very far from the true mean of $v \cdot \mu$. This observation allows us to perform some kind of outlier removal, in particular by removing (or down-weighting) the points x that have $v \cdot x$ inappropriately large.

An important conceptual point here is that one cannot afford to remove only outliers. However, it is possible to ensure that more outliers are removed than inliers. Given a W where Σ_W has a large eigenvalue, one filtering step gives a new distribution $W' \in \Delta^T$ that is closer to W^* than W was. Repeating the process eventually gives a W with no large eigenvalues. The filtering method and its variations are discussed in Section 2.4.

2.3 The Unknown Convex Programming Method

Given an ϵ-corruption T of a stable set S, we would like to estimate the mean of the corresponding distribution X. To achieve this, by Lemma 2.10, it suffices to find a distribution $W \in \Delta^T$ such that Σ_W has no large eigenvalues. We note that this condition *almost* defines a convex program. This is because Δ^T is a convex set of probability distributions and the bounded covariance condition says that $\mathbf{Var}[v \cdot W] \leq 1 + \lambda$ for all unit vectors v. Unfortunately, the variance $\mathbf{Var}[v \cdot W] = \mathbf{E}[|v \cdot (W - \mu_W)|^2]$ is not quite linear in W. (If we instead had $\mathbf{E}[|v \cdot (W - \mu_0)|^2]$, for some fixed vector μ_0, this *would* be linear in W.) However, we will show that a unit vector v for which $\mathbf{Var}[v \cdot W]$ is too large can still be used to obtain a separation oracle, that is, a linear function L for which $L(W) > L(W^*)$, where W^* is the uniform distribution over $S \cap T$.

In particular, suppose that we identify a unit vector v such that $\mathbf{Var}[v \cdot W] = 1 + \lambda$, where $\lambda > C(\delta^2/\epsilon)$ for a sufficiently large universal constant $C > 0$. Applying Lemma 2.10 to the one-dimensional projection $v \cdot W$ gives

$$|v \cdot (\mu_W - \mu_X)| \leq O(\delta + \sqrt{\epsilon\lambda}) = O(\sqrt{\epsilon\lambda}).$$

For a probability distribution Y, let $L(Y) := \mathbf{E}[|v \cdot (Y - \mu_W)|^2]$. Note that L is a linear function of the probability distribution Y with $L(W) = 1 + \lambda$. We can write

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

$$\begin{aligned} L(W^*) &= \mathbf{E}_{W^*}[|v \cdot (W^* - \mu_W)|^2] = \mathbf{Var}[v \cdot W^*] + |v \cdot (\mu_W - \mu_{W^*})|^2 \\ &\le 1 + \delta^2/\epsilon + 2|v \cdot (\mu_W - \mu_X)|^2 + 2|v \cdot (\mu_{W^*} - \mu_X)|^2 \\ &\le 1 + O(\delta^2/\epsilon + \epsilon\lambda) < 1 + \lambda = L(W). \end{aligned}$$

In summary, we have an explicit convex set Δ^T of probability distributions from which we want to find one with eigenvalues bounded by $1 + O(\delta^2/\epsilon)$. Given any $W \in \Delta^T$ which does not satisfy this condition, we can produce a linear function L that separates W from W^*. In fact, it is not hard to see that L also separates W from some small neighborhood R of W^*. Using the ellipsoid algorithm, we obtain the following general theorem.

Theorem 2.11 *Let S be a $(3\epsilon, \delta)$-stable set with respect to a distribution X for some $\epsilon > 0$ sufficiently small. Let T be an ϵ-corrupted version of S. There exists a polynomial time algorithm which given ϵ, δ, and T returns $\widehat{\mu}$ such that $\|\widehat{\mu} - \mu_X\|_2 = O(\delta)$.*

Proof Sketch. Simply run the ellipsoid algorithm with the above separation oracle. At each stage one of two things happens. On the one hand, we may have found a $W \in \Delta^T$ with $\mathbf{Cov}[W] \preceq (1 + O(\delta^2/\epsilon))\, I$. In this case, $\mathbf{E}[W]$ is an appropriate approximation of μ_X by Lemma 2.10. Otherwise, we find a separation oracle L, separating W from R. This lets us find a smaller ellipsoid containing R. As the volume of this ellipsoid decreases by a $(1 - \mathrm{poly}(1/d))$-factor at every iteration, after at most a polynomial number of rounds the ellipsoid will be smaller than R. This shows that we must reach the first case after at most a polynomial number of iterations, and thus our algorithm will run in polynomial time. □

Implications for Concrete Distribution Families Combining Theorem 2.11 with corresponding stability bounds, we obtain concrete applications for various distribution families of interest. Using Proposition 2.3, we obtain:

Corollary 2.12 (Identity Covariance Sub-Gaussian Distributions) *Let T be a set of N ϵ-corrupted samples from an identity covariance sub-Gaussian distribution X on $\mathbf{R}^d$, where N is at least a sufficiently large polynomial in d/ϵ. There exists a polynomial time algorithm which given ϵ and T returns $\widehat{\mu}$ such that with high probability $\|\widehat{\mu} - \mu_X\|_2 = O(\epsilon\sqrt{\log(1/\epsilon)})$.*

We note that Corollary 2.12 can be immediately adapted for identity covariance distributions satisfying weaker concentration assumptions. For example, if X satisfies subexponential concentration in each direction, we obtain an efficient robust mean estimation algorithm with ℓ_2-error of $O(\epsilon \log(1/\epsilon))$. If X has identity covariance and bounded kth central moments, $k \ge 2$, we obtain

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

error $O(\epsilon^{1-1/k})$. As shown in Chapter 1, these error bounds are information-theoretically optimal up to constant factors.

For distributions with unknown and bounded covariance, using Proposition 2.4 we obtain:

Corollary 2.13 (Unknown Bounded Covariance Distributions) *Let T be a set of N ϵ-corrupted samples from a distribution X on $\mathbf{R}^d$ with unknown covariance $\Sigma_X \preceq \sigma^2 I$, for some known $\sigma > 0$, where N is at least a sufficiently large polynomial in d/ϵ. There exists a polynomial time algorithm which given ϵ, σ, and T returns $\widehat{\mu}$ such that with high probability $\|\widehat{\mu} - \mu_X\|_2 = O(\sigma\sqrt{\epsilon})$.*

Similarly, as shown in Chapter 1, this error bound is information-theoretically optimal up to constant factors.

2.4 The Filtering Method

As in the unknown convex programming method, the goal of the filtering method is to find a distribution $W \in \Delta^T$ such that Σ_W has bounded eigenvalues. Given a $W \in \Delta^T$, Σ_W either has bounded eigenvalues (in which case the weighted empirical mean works) or there is a direction v in which $\mathbf{Var}[v \cdot W]$ is too large. In the latter case, the projections $v \cdot W$ must behave very differently from the projections $v \cdot S$ or $v \cdot X$. In particular, since an ϵ-fraction of outliers are causing a much larger increase in the standard deviation, this means that the distribution of $v \cdot W$ will have many "extreme points" – more than one would expect to find in $v \cdot S$. This fact allows us to identify a nonempty subset of extreme points, the majority of which are outliers. These points can then be removed (or down-weighted) in order to "clean up" our sample. Formally, given a $W \in \Delta^T$ without bounded eigenvalues, we can efficiently find a $W' \in \Delta^T$ such that W' is closer to W^* than W was. Iterating this procedure eventually terminates giving a W with bounded eigenvalues.

While it may be conceptually useful to consider the above scheme for general distributions W over points, in most cases it suffices to consider only W given as the uniform distribution over some set of points. The filtering step in this case consists of replacing the set T by some subset $T' = T \setminus R$, where $R \subset T$. To guarantee progress toward W^* (the uniform distribution over $S \cap T$), it suffices to ensure that at most a third of the elements of R are also in S, or equivalently that at least two-thirds of the removed points are outliers (perhaps in expectation). The algorithm will terminate when the current set of points T' has bounded empirical covariance, and the output will be the empirical mean of T'.

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

Before we proceed with a more detailed technical discussion, we note that there are several possible ways to implement the filtering step, and that the method used has a significant impact on the analysis. In general, a filtering step removes all points that are "far" from the sample mean in a large variance direction. However, the precise way that this is quantified can vary in important ways.

2.4.1 Tail-Bound-Based Filtering

In this section, we present a filtering method that yields efficient robust mean estimators with optimal error bounds for identity covariance (or, more generally, known covariance) distributions whose univariate projections satisfy appropriate tail bounds. For the purposes of this section, we will restrict ourselves to the Gaussian setting. We note however that this method immediately extends to distributions with weaker concentration properties, for example, subexponential or even inverse polynomial concentration, with appropriate modifications.

We note that the filtering method presented here requires an additional condition on our set of inlier samples, on top of the stability condition. This is quantified in the following definition.

Definition 2.14 A set $S \subset \mathbf{R}^d$ is *tail-bound-good (with respect to $X = \mathcal{N}(\mu_X, I)$)* if for every unit vector v and every $t > 0$, we have

$$\mathbf{Pr}_{x \sim_u S} \left[|v \cdot (x - \mu_X)| > 2t + 2 \right] \leq e^{-t^2/2}. \tag{2.2}$$

Since any univariate projection of X is distributed like a standard Gaussian, Condition (2.2) should hold if the uniform distribution over S were replaced by X. It can be shown that this condition holds with high probability if S is a set of i.i.d. samples from X of a sufficiently large size. Unfortunately, the sample size required for this condition to hold can be exponential in the dimension. In the rest of this section, to avoid cluttering in the relevant expressions, we develop and analyze our filtering algorithm under this condition. We will then explain (see Remark 2.16) how a simple modification to Definition 2.14 suffices for our algorithm to work and will be satisfied with a polynomial sample size.

Intuitively, the additional tail condition of Definition 2.14 means that the univariate projections of our inlier set satisfy strong tail bounds. If we can find a direction in which one of these tails bounds are substantially violated, we will know that most of the extreme points in this direction must be outliers. Formally, we have the following:

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

Lemma 2.15 *Let $\epsilon > 0$ be a sufficiently small constant. Let $S \subset \mathbf{R}^d$ be both $(2\epsilon, \delta)$-stable and tail-bound-good with respect to $X = \mathcal{N}(\mu_X, I)$, with $\delta = C\epsilon\sqrt{\log(1/\epsilon)}$, for $C > 0$ a sufficiently large constant. Let $T \subset \mathbf{R}^d$ be such that $|T \cap S| \geq (1 - 2\epsilon)\max(|T|, |S|)$ and assume we are given a unit vector $v \in \mathbf{R}^d$ for which $\mathbf{Var}[v \cdot T] > 1 + 2\delta^2/\epsilon$ and $\mathbf{Var}[v \cdot T] > \|\mathbf{Cov}[T]\|_2 - \epsilon$. There exists a polynomial-time algorithm that returns a subset $R \subset T$ satisfying $|R \cap S| < |R|/3$.*

To see why Lemma 2.15 suffices for our purposes, note that by replacing T by $T' = T \setminus R$, we obtain a less noisy version of S than T was. In particular, it is easy to see that the size of the symmetric difference between S and T' is strictly smaller than the size of the symmetric difference between S and T. From this it follows that the hypothesis $|T \cap S| \geq (1 - 2\epsilon)\max(|T|, |S|)$ still holds when T is replaced by T', allowing us to iterate this process until we are left with a set with small variance.

Proof Let $\mathbf{Var}[v \cdot T] = 1 + \lambda$. Our goal will be to compute some threshold L such that the substantial majority of the samples x with $|v \cdot (x - \mu_T)| > L$ are outliers, as is shown in Figure 2.3. This ought to be possible since by assumption

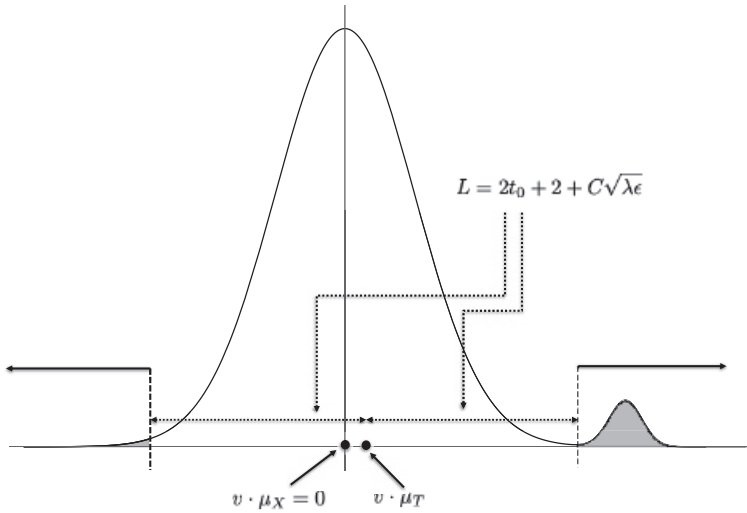

Figure 2.3 Illustration of tail-bound-based filtering. The figure shows the graph of $v \cdot x$ for samples x, with the bump on the right representing the error distribution. The grayed out portions represent the points with $|v \cdot (x - \mu_T)| > L$ that are removed by the filtering algorithm. Notice that the majority of these points are outliers.

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

the inliers are well-concentrated about the mean. On the other hand, we must have many faraway outliers in order to cause $\mathbf{Var}[v \cdot T]$ so large.

We know that since the set S is tail-bound-good, the univariate projection $v \cdot S$ is well-concentrated about $v \cdot \mu_X$. Unfortunately, the algorithm only knows μ_T. However, applying Lemma 2.6 to the set T (and noting that $\|\mathbf{Cov}[T]\|_2 \leq 1 + O(\lambda)$), we get that $|v \cdot \mu_X - v \cdot \mu_T| \leq C\sqrt{\lambda\epsilon}$. Thus, by Condition (2.2),

$$\mathbf{Pr}_{x \sim_u S}\left[|v \cdot (x - \mu_T)| > 2t + 2 + C\sqrt{\lambda\epsilon}\right] \leq e^{-t^2/2}.$$

We claim that there exists a threshold t_0 such that

$$\mathbf{Pr}_{x \sim_u T}\left[|v \cdot (x - \mu_T)| > 2t_0 + 2 + C\sqrt{\lambda\epsilon}\right] > 4e^{-t_0^2/2}. \tag{2.3}$$

Given this claim, the set $R = \{x \in T : |v \cdot (x - \mu_T)| > 2t_0 + 2 + C\sqrt{\lambda\epsilon}\}$ will satisfy the conditions of the lemma.

To prove our claim, we analyze the variance of $v \cdot T$ and note that much of the excess must be due to points in $T \setminus S$. In particular, by our assumption on the variance in the v-direction, we have that

$$\sum_{x \in T} |v \cdot (x - \mu_T)|^2 = |T|\mathbf{Var}[v \cdot T] = |T|(1 + \lambda),$$

where $\lambda > 2\delta^2/\epsilon$. The contribution from the points $x \in S \cap T$ is at most

$$\begin{aligned}\sum_{x \in S} |v \cdot (x - \mu_T)|^2 &= |S|\left(\mathbf{Var}[v \cdot S] + |v \cdot (\mu_T - \mu_S)|^2\right) \leq |S|(1 + \delta^2/\epsilon + 2C^2\lambda\epsilon) \\ &\leq |T|(1 + 2C^2\lambda\epsilon + 3\lambda/5),\end{aligned}$$

where the first inequality uses the stability of S, and the last inequality uses that $|T| \geq (1 - 2\epsilon)|S|$. If ϵ is sufficiently small relative to C, it follows that $\sum_{x \in T \setminus S} |v \cdot (x - \mu_T)|^2 \geq |T|\lambda/3$. On the other hand, by definition we have that

$$\sum_{x \in T \setminus S} |v \cdot (x - \mu_T)|^2 = |T| \int_0^\infty 2t\mathbf{Pr}_{x \sim_u T}\left[|v \cdot (x - \mu_T)| > t, x \notin S\right] dt. \tag{2.4}$$

Assume for the sake of contradiction that there is no t_0 for which Condition (2.3) is satisfied. Then the RHS of (2.4) is at most

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

$$
\begin{aligned}
&|T|\Bigg(\int_0^{2+C\sqrt{\lambda\epsilon}+10\sqrt{\log(1/\epsilon)}} 2t\mathbf{Pr}_{x\sim_u T}[x\notin S]dt \\
&\qquad + \int_{2+C\sqrt{\lambda\epsilon}+10\sqrt{\log(1/\epsilon)}}^{\infty} 2t\mathbf{Pr}_{x\sim_u T}\left[|v\cdot(x-\mu_T)|>t\right]dt\Bigg) \\
&\le |T|\left(\epsilon(2+C\sqrt{\lambda\epsilon}+10\sqrt{\log(1/\epsilon)})^2 + \int_{5\sqrt{\log(1/\epsilon)}}^{\infty} 16(2t+2+C\sqrt{\lambda\epsilon})e^{-t^2/2}dt\right) \\
&\le |T|\left(O(C^2\lambda\epsilon^2+\epsilon\log(1/\epsilon)) + O(\epsilon^2(\sqrt{\log(1/\epsilon)}+C\sqrt{\lambda\epsilon}))\right) \\
&\le |T|O(C^2\lambda\epsilon^2+(\delta^2/\epsilon)/C) < |T|\lambda/3,
\end{aligned}
$$

which is a contradiction. Therefore, the tail bounds and the concentration violation together imply the existence of such a t_0 (which can be efficiently computed by simple enumeration). □

Remark 2.16 We note that although exponentially many samples are required to ensure that Condition (2.2) holds with high probability, one can carefully weaken this condition so that it can be achieved with polynomially many samples without breaking the aforementioned analysis. Specifically, it suffices to add an inverse polynomially small slack term in the right-hand side to account for the difference between the empirical and population values of the corresponding probability. Using the VC Inequality (Theorem A.12), one can show that this weaker condition holds for the uniform distribution over S with high probability, where S is a set of i.i.d. samples from X of a sufficiently large polynomial size. This slightly alters the analysis, as one needs to add this slack term to all of the relevant probability integrals. However, these integrals can still be truncated to cover only a polynomial range (using the fact that likely no inliers will be too far from the true mean), and thus the total integral of this additional error will remain small.

2.4.2 Randomized and Weighted Filtering

The filtering method described in Section 2.4.1 works by guaranteeing that (assuming the set of inliers is stable and tail-bound-good) each filtering step removes more outliers than inliers. For some of the more general settings one instead requires a randomized filtering method that merely removes more outliers *in expectation*. In this section, we will develop the general theory of such randomized filters. This will then be applied in Section 2.4.3, where we produce a specific randomized filter that works assuming only the stability of the set of inliers.

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

Randomized Filtering The tail-bound-based filtering method of the previous section is deterministic, relying on the violation of a concentration inequality satisfied by the inliers. In some settings (such as robust estimation of the mean of a bounded covariance distribution), deterministic filtering seems to fail to give optimal results, and we require the filtering procedure to be randomized.

The main idea of randomized filtering is simple: Suppose we can identify a nonnegative function $f(x)$, defined on the samples x, for which (under some high probability condition on the inliers) it holds that $\sum_T f(x) \geq 2\sum_S f(x)$, where T is an ϵ-corrupted set of samples and S is the corresponding set of inliers. Then we can create a randomized filter by removing each sample point $x \in T$ with probability proportional to $f(x)$. This ensures that the *expected* number of outliers removed is at least the *expected* number of inliers removed. The analysis of such a randomized filter is slightly more subtle, so we will discuss it in the following paragraphs.

The key property the above randomized filter ensures is that the sequence of random variables

$$(\#\text{ Inliers removed}) - (\#\text{ Outliers removed})$$

(where "inliers" are points in S and "outliers" points in $T \setminus S$) across iterations is a supermartingale. Since the total number of outliers removed across all iterations accounts for at most an ϵ-fraction of the total samples, this means that with probability at least $2/3$, at no point does the algorithm remove more than a 2ϵ-fraction of the inliers. A formal statement follows.

Theorem 2.17 (Randomized Filtering) *Let $S \subset \mathbf{R}^d$ be a $(4\epsilon, \delta)$-stable set with respect to some distribution X, for $\epsilon < 1/12$, and let T be an ϵ-corrupted version of S. Suppose that given any $T' \subseteq T$ with $|T' \cap S| \geq (1 - 4\epsilon)|S|$ for which $\mathbf{Cov}[T']$ has an eigenvalue bigger than $1 + \lambda$, for some $\lambda \geq 0$, there is a polynomial-time algorithm that computes a nonzero function $f : T' \to \mathbf{R}_+$ such that $\sum_{x \in T'} f(x) \geq 2\sum_{x \in T' \cap S} f(x)$. Then there exists a polynomial-time randomized algorithm that given T computes a vector $\widehat{\mu}$ that with probability at least $2/3$ satisfies $\|\widehat{\mu} - \mu_X\|_2 = O(\delta + \sqrt{\epsilon\lambda})$.*

The algorithm is described in pseudocode below:

Algorithm `Randomized-Filtering`

1. Compute $\mathbf{Cov}[T]$ and its largest eigenvalue ν.
2. If $\nu \leq 1 + \lambda$, return $\mu_T = (1/|T|)\sum_{x \in T} x$.
3. Else
   - Compute f as guaranteed in the theorem statement.
   - Remove each $x \in T$ with probability $f(x)/\max_{x \in T} f(x)$ and return to Step 1 with the new set T.

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

Proof of Theorem 2.17 First, it is easy to see that this algorithm runs in polynomial time. Indeed, as the point $x \in T$ attaining the maximum value of $f(x)$ is definitely removed in each filtering iteration, each iteration reduces $|T|$ by at least one. To establish correctness, we will show that, with probability at least $2/3$, it holds throughout the algorithm that $|S \cap T| \geq (1-4\epsilon)|S|$. Assuming this claim, Lemma 2.6 implies that our final error will be as desired.

To prove the desired claim, we consider the sequence of random variables

$$d(T_i) := |(S \cap T) \setminus T_i| + |T_i \setminus S|,$$

where T_i denotes the version of T after the ith iteration of our algorithm. Note that $d(T_i)$ is essentially the number of remaining outliers plus the number of inliers that our algorithm has removed so far. We note that, initially, $d(T_0) \leq \epsilon|S|$ and that $d(T_i)$ cannot drop below 0. Finally, we note that at each stage of the algorithm $d(T_i)$ increases by (# Inliers removed) – (# Outliers removed), and that the expectation of this quantity is

$$\sum_{x \in S \cap T_i} f(x) - \sum_{x \in T_i \setminus S} f(x) = 2 \sum_{x \in S \cap T_i} f(x) - \sum_{x \in T_i} f(x) \leq 0.$$

This means that the sequence of random variables $d(T_i)$ is a supermartingale (at least until we reach a point where $|S \cap T| \leq (1-4\epsilon)|S|$). However, if we set a stopping time at the first occasion where this condition fails, we note that the expectation of $d(T_i)$ is at most $\epsilon|S|$. Since it is always at least 0, Proposition A.5 implies that with probability at least $2/3$ it is never more than $3\epsilon|S|$, which in turn implies that $|S \cap T| \geq (1-4\epsilon)|S|$ throughout the algorithm. If this is the case, the inequality $|T' \cap S| \geq (1-4\epsilon)|S|$ will continue to hold throughout our algorithm, thus eventually yielding such a set with the variance of T' bounded. By Lemma 2.6, the mean of this subset T' will be a suitable estimate for the true mean, completing the proof of Theorem 2.17. □

Methods of Point Removal The randomized filtering method described above only requires that each point x is removed with probability $f(x)/\max_{x \in T} f(x)$, without any assumption of independence. Therefore, given an f, there are several ways to implement this scheme. A few natural ones are given here:

- *Randomized Thresholding:* Perhaps the easiest method for implementing our randomized filter is generating a uniform random number y in the interval $[0, \max_{x \in T} f(x)]$ and removing all points $x \in T$ for which $f(x) \geq y$. This method is practically useful in many applications. Finding the set of such points is often fairly easy, as this condition may well correspond to a simple threshold.

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

- *Independent Removal:* Each $x \in T$ is removed independently with probability $f(x)/\max_{x \in T} f(x)$. This scheme has the advantage of leading to less variance in $d(T)$. A careful analysis of the random walk involved allows one to reduce the failure probability to $\exp(-\Omega(\epsilon|S|))$ (see Exercise 2.11).
- *Deterministic Reweighting:* Instead of removing points, this scheme allows for weighted sets of points. In particular, each point will be assigned a weight in $[0, 1]$, and we will consider weighted means and covariances. Instead of removing a point x with probability proportional to $f(x)$, we can multiplicatively reduce the weight assigned to x by a quantity proportional to $f(x)$. This ensures that the appropriate weighted version of $d(T)$ is definitely nonincreasing, implying deterministic correctness of the algorithm.

Weighted Filtering The last of the aforementioned methods being deterministic is useful in some settings, and so the algorithm is worth explicitly stating. To begin, for a weight function $w\colon T \to \mathbf{R}_+$, we define the weighted mean and covariance of T by

$$\mu_w[T] := \frac{1}{\|w\|_1} \sum_{x \in T} w_x x,$$
$$\mathbf{Cov}_w[T] := \frac{1}{\|w\|_1} \sum_{x \in T} w_x (x - \mu_w)(x - \mu_w)^\top.$$

One can observe that these quantities are simply the mean and covariance of the probability distribution on T that assigns each point $x \in T$ probability of $w_x/\|w\|_1$.

With this setup, we have the following theorem, a direct analogue of Theorem 2.17.

Theorem 2.18 (Weighted Filtering) *Let $S \subset \mathbf{R}^d$ be a $(4\epsilon, \delta)$-stable set with respect to some distribution X, for $\epsilon < 1/12$, and let T be an ϵ-corrupted version of S. Suppose that for any weight vector $w\colon T \to \mathbf{R}_+$ for which the corresponding probability distribution is (3ϵ)-close to the uniform distribution on S in total variation distance and for which $\mathbf{Cov}_w[T]$ has an eigenvalue larger than $1 + \lambda$, for some $\lambda \geq 0$, there is a polynomial-time algorithm that computes a nonzero function $f\colon T \to \mathbf{R}_+$ such that $\sum_{x \in S \cap T} w_x f(x) \leq (1/2) \sum_{x \in T} w_x f(x)$. Then there exists a polynomial-time algorithm that outputs a vector $\widehat{\mu}$ which with probability at least $2/3$ satisfies $\|\widehat{\mu} - \mu_X\|_2 = O(\delta + \sqrt{\epsilon\lambda})$.*

Proof The algorithm is described in pseudocode below.

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

Algorithm `Weighted-Filtering`

1. Set $t = 1$ and $w_x^{(1)} = 1/|T|$ for all $x \in T$.
2. While $\mathbf{Cov}_{w^{(t)}}[T]$ has an eigenvalue larger than $1 + \lambda$:
1. Compute a weight function $f(x)$ as described above.
2. Let $f_{\max}$ be the maximum value of $f(x)$ over $x \in T$ with $w_x^{(t)} \neq 0$.
3. Let $w_x^{(t+1)} = w_x^{(t)}(1 - f(x)/f_{\max})$. Set t to $t + 1$.
3. Return $\mu_{w^{(t)}}$.

To analyze this algorithm we make the following observations. First, at each iteration, the support of w decreases by at least 1, as $w_x^{(t+1)} = 0$ for x with $f(x) = f_{\max}$. This implies that the algorithm will terminate after polynomially many iterations.

To prove correctness, as long as the distribution defined by $w^{(t)}$ is close to the uniform distribution on S, we have that

$$\sum_{x \in S \cap T} w_x^{(t+1)} = \sum_{x \in S \cap T} [w_x^{(t)} - w_x^{(t)} f(x)/f_{\max}] = \sum_{x \in S \cap T} w_x^{(t)} - (1/f_{\max}) \sum_{x \in S \cap T} w_x^{(t)} f(x)$$
$$\geq \sum_{x \in S \cap T} w_x^{(t)} - \frac{1}{2}(1/f_{\max}) \sum_{x \in T} w_x^{(t)} f(x),$$

where the first equality follows from the definition of $w_x^{(t+1)}$ and the inequality follows from the definition of f. On the other hand, we can write

$$\sum_{x \in T} w_x^{(t+1)} = \sum_{x \in T} [w_x^{(t)} - w_x^{(t)} f(x)/f_{\max}] = \sum_{x \in T} w_x^{(t)} - (1/f_{\max}) \sum_{x \in T} w_x^{(t)} f(x).$$

This means that in each iteration the weight function $w_x^{(t)}$ decreases half as much over S as it does over T as a whole. Thus, the amount that $w_x^{(t)}$ decreases on $S \cap T$ is at most the amount it decreases on $T \setminus S$. Since initially we have that $\sum_{x \in T \setminus S} w_x^{(1)} = |T \setminus S|/|T| \leq \epsilon$, this means that at every stage t of the algorithm the following holds

$$\sum_{x \in S \cap T} w_x^{(t)} \geq 1 - 2\epsilon.$$

This implies that the distribution defined by $w^{(t)}$ remains (3ϵ)-close to the uniform distribution on S, even at the end of the algorithm when $\mathbf{Cov}_{w^{(t)}}[T] \leq (1 + \lambda) I$. Thus, by Lemma 2.7, we have that $\|\mu_{w^{(t)}} - \mu_S\|_2 = O(\delta + \sqrt{\epsilon\lambda})$, completing our proof. □

Practical Considerations While the aforementioned point removal methods have similar theoretical guarantees, recent implementations suggest that they have different practical performance on real datasets. The deterministic

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

reweighting method is somewhat slower in practice as its worst-case runtime and its typical runtime are comparable. In more detail, one can guarantee termination by setting the constant of proportionality so that at each step at least one of the nonzero weights is set to zero. However, in practical circumstances, we will not be able to do better. That is, the algorithm may well be forced to undergo $\epsilon|S|$ iterations. On the other hand, the randomized versions of the algorithm are likely to remove several points of T at each filtering step.

Another reason why the randomized versions may be preferable has to do with the quality of the results. The randomized algorithms only produce bad results when there is a chance that $d(T_i)$ ends up being very large. However, since $d(T_i)$ is a supermartingale, this will only ever be the case if there is a corresponding possibility that $d(T_i)$ will be exceptionally small. Thus, although the randomized algorithms may have a probability of giving worse results some of the time, this will only happen if a corresponding fraction of the time they also give *better* results than the theory guarantees. This consideration suggests that the randomized thresholding procedure might have advantages over the independent removal procedure, precisely because it has a higher probability of failure. This has been observed experimentally: In real datasets poisoned with a constant fraction of adversarial outliers, the number of iterations of randomized filtering is typically bounded by a small constant.

2.4.3 Universal Filtering

In this section, we show how to use randomized filtering to construct a universal filter that works under only the stability condition (Definition 2.1) – not requiring the tail-bound condition of the tail-bound filter (Lemma 2.15). To do this, we construct an appropriate score function f, as in the statement of Theorem 2.17. Formally, we show the following.

Proposition 2.19 *Let $S \subset \mathbf{R}^d$ be a $(2\epsilon, \delta)$-stable set for $\epsilon, \delta > 0$ sufficiently small constants with δ at least a sufficiently large multiple of ϵ. Let T be an ϵ-corrupted version of S. Suppose that $\mathbf{Cov}[T]$ has largest eigenvalue $1 + \lambda > 1 + 8\delta^2/\epsilon$. Then there exists a polynomial time algorithm that, on input ϵ, δ, T, computes a nonzero function $f\colon T \to \mathbf{R}_+$ satisfying $\sum_{x\in T} f(x) \geq 2 \sum_{x\in T\cap S} f(x)$.*

By combining Theorem 2.17 and Proposition 2.19, we obtain a randomized filtering algorithm establishing Theorem 2.11.

Proof of Proposition 2.19. The algorithm to construct f is the following. We start by computing the sample mean of T, μ_T, and the top (unit) eigenvector v of $\mathbf{Cov}[T]$. For $x \in T$, we define the function

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

$$g(x) = (v \cdot (x - \mu_T))^2.$$

Let L be the set of $\epsilon \cdot |T|$ elements of T on which $g(x)$ is largest. We define f to be

$$f(x) = \begin{cases} 0 & x \notin L, \\ g(x) & x \in L. \end{cases} \tag{2.5}$$

Our basic plan of attack is as follows: First, we note that the sum of $g(x)$ over $x \in T$ is the variance of $v \cdot T$, which is substantially larger than the sum of $g(x)$ over $x \in S$, which is approximately the variance of $v \cdot S$. Therefore, the sum of $g(x)$ over the $\epsilon|S|$ elements of $T \setminus S$ must be quite large. In fact, using the stability condition, we can show that the latter quantity must be larger than the sum of the largest $\epsilon|S|$ values of $g(x)$ over $x \in S$. However, since $|T \setminus S| \leq |L|$, we have that $\sum_{x \in T} f(x) = \sum_{x \in L} g(x) \geq \sum_{x \in T \setminus S} g(x) \geq 2 \sum_{x \in S} f(x)$.

We now proceed with the detailed analysis. First, note that

$$\sum_{x \in T} g(x) = |T| \mathbf{Var}[v \cdot T] = |T|(1 + \lambda).$$

Moreover, for any $S' \subseteq S$ with $|S'| \geq (1 - 2\epsilon)|S|$, we have that

$$\sum_{x \in S'} g(x) = |S'|(\mathbf{Var}[v \cdot S'] + (v \cdot (\mu_T - \mu_{S'}))^2). \tag{2.6}$$

By the second stability condition, we have that $|\mathbf{Var}[v \cdot S'] - 1| \leq \delta^2/\epsilon$. Furthermore, the stability condition and Lemma 2.6 give

$$\|\mu_T - \mu_{S'}\|_2 \leq \|\mu_T - \mu_X\|_2 + \|\mu_X - \mu_{S'}\|_2 = O(\delta + \sqrt{\epsilon\lambda}).$$

Since $\lambda \geq 8\delta^2/\epsilon$, combining the above gives

$$\sum_{x \in T \setminus S} g(x) \geq \sum_{x \in T} g(x) - \sum_{x \in S} g(x) \geq (2/3)|S|\lambda.$$

Moreover, since $|L| \geq |T \setminus S|$ and g takes its largest values on points $x \in L$, we have

$$\sum_{x \in T} f(x) = \sum_{x \in L} g(x) \geq \sum_{x \in T \setminus S} g(x) \geq (16/3)|S|\delta^2/\epsilon.$$

Comparing the results of Equation (2.6) for $S' = S$ and $S' = S \setminus L$, we find that

$$\begin{aligned} \sum_{x \in S \cap T} f(x) &= \sum_{x \in S \cap L} g(x) = \sum_{x \in S} g(x) - \sum_{x \in S \setminus L} g(x) \\ &= |S|(1 \pm \delta^2/\epsilon + O(\delta^2 + \epsilon\lambda)) - |S \setminus L|(1 \pm \delta^2/\epsilon + O(\delta^2 + \epsilon\lambda)) \\ &\leq 2|S|\delta^2/\epsilon + |S|O(\delta^2 + \epsilon\lambda). \end{aligned}$$

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

The latter quantity is at most $(1/2)\sum_{x\in T} f(x)$ when δ and ϵ/δ are at most sufficiently small constants. This completes the proof of Proposition 2.19. □

Remark 2.20 One can straightforwardly obtain a weighted version of Proposition 2.19 (essentially by replacing subsets by "weighted subsets"), which provides the function f required in the statement of Theorem 2.18. By doing so, we obtain a weighted filtering algorithm establishing Theorem 2.11.

2.5 Exercises

2.1 (Scaling Stability) Show that if the set $S \subset \mathbf{R}^d$ is (ϵ, δ)-stable with respect to μ, and if $\epsilon' > \epsilon$ is less than a sufficiently small constant, then S is $(\epsilon', O(\delta\epsilon'/\epsilon))$-stable with respect to μ.

2.2 (Resilience) Suppose that S is a set of points in $\mathbf{R}^d$ such that for every $S' \subseteq S$ with $|S'| \geq (1-2\epsilon)|S|$ we have $\|\mu_{S'} - \mu_S\|_2 \leq \delta$. (Note that this is the first condition in the definition of $(2\epsilon, \delta)$-stability, but not the second.) Show that if one is given a set T obtained by adversarially corrupting an ϵ-fraction of the points in S, it is information-theoretically possible to find a 2δ-approximation of the mean of S.

Remark 2.21 This condition was referred to as *resilience* by [136]. That work showed that it is information-theoretically sufficient to robustly learn to error $O(\delta)$, and computationally sufficient to learn to error $O(\delta/\sqrt{\epsilon})$. Although robust learning is possible with this weaker condition information-theoretically, it is believed that obtaining error $O(\delta)$ is computationally intractable without additional assumptions.

2.3 (Stability for Bounded Covariance) Recall that in Proposition 2.4 we needed to restrict to a subset of the sample points to ensure that the resulting subset is stable with high probability. Show that this assumption is necessary. In particular, show that for any positive integers N, d and real $\epsilon > 0$ sufficiently small, there is a distribution X on $\mathbf{R}^d$ with $\mathbf{Cov}[X] \preceq I_d$ such that, with probability at least $1/2$, the empirical distribution of N samples from X is *not* $(\epsilon, \sqrt{d\epsilon/2})$-stable with respect to X.
(Hint: Produce a distribution that has a $1/N$ probability of returning a very large vector.)

2.4 (Generic Stability Bound) Suppose that X is a probability distribution in $\mathbf{R}^d$ with mean μ with $\|X - \mu\|_2 \leq R$ almost surely, and such that no ϵ-fraction of the mass of X contributes more than δ^2/ϵ to the expectation of $(v \cdot (X-\mu))^2$ for any unit vector v. Prove that, for some $N = \mathrm{poly}(Rd/\epsilon)$, a set of N i.i.d. samples from X is $(\epsilon, O(\delta))$-stable with respect to μ with high probability.

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

(Hint: Use the VC Inequality, Theorem A.12, to show that with high probability the empirical distribution satisfies tail bounds similar to those that X does.)

2.5 (Other Tail-Bound-Based Filters) Devise filtering algorithms along the lines of the tail-bound-based filter for Gaussians that work for the following inlier distributions X:

(a) X is isotropic and logconcave. [Here you should be able to achieve error $O(\epsilon \log(1/\epsilon))$.]

(b) X is isotropic and has $\mathbf{E}[|v \cdot (X - \mu_X)|^k] \leq M$ (for some constants M and $k > 2$). [Here you should be able to achieve error $O_k(M^{1/k}\epsilon^{1-1/k})$.]

(c) X is an arbitrary distribution with $\mathbf{Cov}[X] \leq I$. [Although one can get sample sets that are $(\epsilon, O(\sqrt{\epsilon}))$-stable here, it seems impossible to achieve error $O(\sqrt{\epsilon})$ with a filter of this type. Show that it is possible to get error $O(\sqrt{\epsilon}\log(d/\epsilon))$.]

2.6 (Dimension Halving) Another approach for robust mean estimation of spherical Gaussians uses a dimension-halving technique. This method proceeds as follows:

(a) Use a naive filter to remove all points at distance more than roughly $\sqrt{d}$ from the mean.

(b) Compute the sample covariance matrix. Let V be the subspace spanned by eigenvalues larger than $1 + \Omega(\epsilon)$.

(c) Use the sample mean as an estimate for the projection of the true mean onto $V^{\perp}$, and recursively approximate the projection of the mean onto V.

Show that an algorithm along these lines can be used to obtain error $O(\epsilon\sqrt{\log(d)})$ with polynomial time and sample complexity.

(Hint: Show that $\dim(V) \leq d/2$.)

Remark 2.22 The dimension-halving technique was developed in [114].

2.7 (Robust Estimation in Other ℓ_p-Norms) Let $1 \leq p < 2$ and let $1/p + 1/q = 1$. Suppose that S is a set of points such that $\mathbf{Var}[v \cdot S] \leq 1$ for all v with $\|v\|_q \leq 1$. Show that there is an algorithm that given p, ϵ, and T, an ϵ-corrupted version of S, computes in polynomial time an estimate $\widehat{\mu}$ such that with high probability $\|\widehat{\mu} - \mu_S\|_p = O(\sqrt{\epsilon})$.

(Hint: Show that it suffices to find a large (weighted) subset T' of T for which the variance of $v \cdot T'$ is $O(1)$ for any v with $\|v\|_q \leq 1$. In order to

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

find such a subset, you may need the following result of [123]:
For any positive-definite matrix A, the following holds

$$\sup_{\|v\|_q=1} v^\top A v = \Theta\left(\sup_{Y\succeq 0, \|\mathrm{Diag}(Y)\|_{q/2}\le 1} \mathrm{tr}(AY) \right).$$

This is particularly convenient, as the right-hand side can be efficiently computed using convex programming.)

2.8 (Learning from Untrusted Batches) In the *learning from untrusted batches* problem, one is attempting to learn a distribution p over a finite domain $[n] := \{1, 2, \ldots, n\}$, in a distributed setting where many samples are partitioned across a few servers, but a constant fraction of the servers may be corrupted. More precisely, we are given m i.i.d. samples from p divided into *batches* of k samples each. However, an ϵ-fraction of these batches are allowed to be adversarially corrupted (usually in the Huber sense). The goal is to learn a distribution $\widehat{p}$ that is close to p in total variation distance.

(a) Show that for $k = 1$ one cannot learn p to error better than ϵ, no matter how large m is.

(b) Show that for any subset $S \subseteq [n]$, there is a polynomial-time algorithm to estimate the probability $p(S)$ that p assigns to S within error of $O(\epsilon/\sqrt{k})$. Use this to devise an inefficient algorithm to estimate p to error $O(\epsilon/\sqrt{k})$ in total variation distance.

(c) Show that the learning from untrusted batches problem is equivalent to estimating the mean of a multinomial distribution to small ℓ_1 error, given access to ϵ-corrupted samples. Show that the algorithm from Exercise 2.7 can be used to efficiently learn p to ℓ_1-error $O(\sqrt{\epsilon/k})$.

(d) One can actually do somewhat better than the above. The idea is to find sets $S \subset [n]$ such that the empirical variance of the number of samples in a batch from S is substantially larger than the variance over just the good batches, and using this set to filter. This can be done by comparing the sample covariance matrix to an approximation of the true covariance, and using a known result that gives a polynomial-time algorithm for the following task: Given a matrix M, compute a vector v with $\|v\|_\infty \le 1$ and $v^\top M v \gg \sup_{\|w\|_\infty \le 1} w^\top M w$. Give a polynomial-time algorithm that estimates p to error $O(\epsilon\sqrt{\log(1/\epsilon)/k})$ in total variation distance.

Remark 2.23 The learning from untrusted batches problem was introduced by [128] and subsequently studied in a sequence of works [27, 28, 99, 100, 101].

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

2.9 (Robust Mean Estimation for Balanced Product Distributions) Let X be a balanced product distribution on $\{0,1\}^d$. Namely, X_i is 1 with probability p_i and 0 otherwise, for some $1/3 \le p_i \le 2/3$, and the coordinates X_i are independent of one another. Note that $\Sigma := \mathbf{Cov}[X]$ is a diagonal matrix with entries $p_i(1-p_i)$. Give an efficient algorithm to estimate the mean of X to ℓ_2-error $O(\epsilon\sqrt{\log(1/\epsilon)})$ from a polynomial number of ϵ-corrupted samples.

(Hint: Compute an approximation to Σ and find a way to adjust for the fact that Σ is not close to I.)

2.10 (Achieving Breakdown Point of 1/2) The algorithms presented in this chapter all require that the fraction of corruptions ϵ is at most a sufficiently small positive constant. Adaptations of these algorithms can be made to work for ϵ approaching $1/2$. (For the more challenging setting when $\epsilon > 1/2$, see Chapter 5). Show that for all $0 < \epsilon < 1/2$ there is an algorithm that takes $\mathrm{poly}(d/(1/2-\epsilon))$ samples from $X = \mathcal{N}(\mu, I)$ in $\mathbf{R}^d$, runs in polynomial time, and with high probability computes an estimate $\widehat{\mu}$ with $\|\widehat{\mu} - \mu\|_2 \le f(\epsilon)$, for some function f.

(Hint: Some version of a filter should work, though you may need to be more careful either about the ratio of inliers versus outliers removed or about the properties that you can assume for $S \cap T$.)

2.11 (High-Probability Guarantees in Randomized Filtering) Consider the version of the randomized filter where each sample is removed independently with probability $f(x)/f_{\max}$. Show that if this algorithm is given a set T, which is an ϵ-corruption of a set S, the probability that the algorithm ever reaches a state where more than $3\epsilon|S|$ samples have been removed from S is at most $\exp(-\Omega(\epsilon|S|))$.

(Hint: Consider the expectation $\mathbf{E}\left[\exp(\eta(2|T_i \backslash S| + |(S \cap T)\backslash T_i|))\right]$ for $\eta > 0$ some sufficiently small constant.)

2.12 (Different Scores for the Universal Filter) Recall that for our universal filter we let $g(x) = |v \cdot (x - \mu_T)|^2$ and defined our scores to be $g(x)$ if x was in the top ϵ-fraction of values and 0 otherwise.

(a) Show that if instead $g(x)$ is used directly as the score function, this may throw away more good samples than bad ones, unless $\delta \gg \sqrt{\epsilon}$.

(b) Let m be an $O(1)$-additive approximation to $v \cdot \mu_S$ (for example, the median of $v \cdot T$ often works). Let $g(x) = |v \cdot x - m|^2$ and

$$f(x) := \begin{cases} g(x) & \text{if } g(x) > C(\delta/\epsilon)^2, \\ 0 & \text{otherwise} \end{cases}$$

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

for $C > 0$ some suitably large constant. Show that this score function works. Namely, show that if T is an ϵ-corruption of S, S is (ϵ, δ)-stable and if $\mathbf{Var}[v \cdot T] > 1 + C'\delta^2/\epsilon$, for some sufficiently large C', then $\sum_{x \in S \cap T} f(x) < \frac{1}{2} \sum_{x \in T} f(x)$.

2.6 Discussion and Related Work

The first computationally efficient algorithms for high-dimensional robust mean estimation with dimension-independent error guarantees were obtained in [45]. The same work introduced both the unknown convex programming and filtering techniques described in this chapter. The filtering technique was further refined in [46], specifically for the class of bounded covariance distributions. In this chapter, we gave a simplified and unified presentation of these techniques. In more detail, the stability condition of Definition 2.1 first appeared in [50], although a special case was implicitly used in [45]. Similarly, the universal filtering method that succeeds under the stability condition first appeared in [50].

The idea of removing outliers by projecting on the top eigenvector of the empirical covariance goes back to [110], who used it in the context of learning linear separators with malicious noise. That work [110] used a "hard" filtering step which only removes outliers, and consequently leads to errors that scale logarithmically with the dimension. Subsequently, the work of [5] employed a soft-outlier removal step in the same supervised setting as [110], to obtain improved bounds for that problem. It should be noted that the soft-outlier method of [5] is similarly insufficient to obtain dimension-independent error bounds for the unsupervised setting.

Contemporaneously with [45], [114] developed a recursive dimension-halving technique for high-dimensional robust mean estimation. Their technique leads to error $O(\epsilon \sqrt{\log d})$ for Gaussian robust mean estimation in Huber's contamination model. The algorithm of [114] begins by removing extreme outliers from the input set of ϵ-corrupted samples. This ensures that, after this basic outlier removal step, the empirical covariance matrix has trace $d(1 + \tilde{O}(\epsilon))$, which in turn implies that the $d/2$ smallest eigenvalues are all at most $1 + \tilde{O}(\epsilon)$. This allows [114] to show, using techniques akin to Lemma 2.6, that the projections of the true mean and the empirical mean onto the subspace spanned by the corresponding (small) eigenvectors are close. The [114] algorithm then uses this approximation for this projection of the mean, projects the remaining points onto the orthogonal subspace, and recursively finds the mean of the other projection. See Exercise 2.6 for more details.

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

In addition to robust mean estimation, [45, 114] developed efficient robust learning algorithms for a number of more complex statistical tasks, including robust covariance estimation, robust density estimation for mixtures of spherical Gaussians and binary product distributions (see Exercise 2.9), robust independent component analysis (ICA), and robust singular value decomposition (SVD). Building on the techniques of [45], a line of works [34, 35, 66] gave robust parameter estimation algorithms for Bayesian networks (with known graph structure) and Ising models. Another extension of these results was given in [136], who obtained an efficient algorithm for robust mean estimation with respect to all ℓ_p-norms (see Exercise 2.7 for more details).

The algorithmic approaches described in this chapter robustly estimate the mean of a spherical Gaussian within ℓ_2 error $O(\epsilon\sqrt{\log(1/\epsilon)})$ in the strong contamination model. A more sophisticated filtering technique that achieves the optimal error of $O(\epsilon)$ in the *additive* (adaptive) contamination model was developed in [47]. This method will be described and analyzed in Chapter 3. Very roughly, this algorithm proceeds, by using a novel filtering method, to remove corrupted points if the empirical covariance matrix has *many* eigenvalues of size $1+\Omega(\epsilon)$. Otherwise, the algorithm uses the empirical mean to estimate the mean on the space spanned by small eigenvectors, and then uses brute-force to estimate the projection onto the few principal eigenvectors. For the total variation contamination model (and, therefore, the strong contamination model), [63] gave evidence (in the form of Statistical Query lower bounds) that any improvement on the $O(\epsilon\sqrt{\log(1/\epsilon)})$ error requires super-polynomial time. These developments will be described in Chapter 8.

The focus of this chapter was on developing efficient robust mean estimation algorithms in high dimensions that succeed if the fraction of outliers is $\epsilon < \epsilon_0$, where $\epsilon_0 > 0$ is a sufficiently small universal constant. In principle, it is possible to do better than this, in particular to, obtain efficient robust mean estimators with breakdown point of $1/2$. This goal can be achieved by conceptually simple adaptations of the filtering method. The reader is referred to [39, 94, 144] and Exercise 2.10.

A related problem is that of *high probability mean estimation*. If one is given independent samples from a Gaussian (with no corruptions), the empirical mean gives a good estimate of the true mean, and furthermore one can show that this estimate is accurate with high probability. However, if the underlying distribution is replaced by a heavy-tailed distribution (such as, one with merely bounded covariance), these high probability bounds may no longer hold without a more sophisticated estimator. A sequence of works in the mathematical statistics community determined the optimal sample complexity of

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press

heavy-tailed mean estimation both without outliers [120] and in the strong contamination model [121]. (See also [119] for a related survey.)

Interestingly, there is a connection between high probability mean estimation and robust mean estimation, obtained by treating the extreme points from the heavy-tailed distribution (which make the high-probability estimation task challenging) as outliers; see, for example, [126]. In particular, [59] showed that robust mean estimation techniques could be used to obtain essentially optimal high probability mean estimation algorithms.

The first sample-optimal and polynomial-time algorithm for heavy-tail mean estimation (without outliers) was developed in [90]. Subsequent works [37, 42] developed simpler algorithms with significantly improved asymptotic runtime that also succeed with additive contamination. More recently, the work of [59] showed that any robust mean estimation algorithm that succeeds under the stability condition when combined with a simple preprocessing step achieves optimal rates for finite covariance distributions and works even in the strong contamination model. The latter work also establishes the sample complexity bounds stated in Remark 2.5 for identity covariance distributions with bounded kth central moments.

https://doi.org/10.1017/9781108943161.002 Published online by Cambridge University Press