ROBUST LEARNING: INFORMATION THEORY AND ALGORITHMS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Jacob Steinhardt
September 2018

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Percy Liang) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(John Duchi)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Gregory Valiant)

Approved for the Stanford University Committee on Graduate Studies

Preface

This thesis provides an overview of recent results in robust estimation due to myself and my collaborators. The key question is the following: given a dataset, some fraction of which consists of arbitrary outliers, what can be learned about the non-outlying points? This is a classical question going back at least to Tukey (1960). However, this question has recently received renewed interest for a combination of reasons. First, many of the older results do not give meaningful error bounds in high dimensions (for instance, the error often includes an implicit $\sqrt{d}$-factor in d dimensions). This calls for a renewed study as machine learning increasingly works with high-dimensional models. Second, in Charikar et al. (2017) we established connections between robust estimation and other problems such as clustering and learning of stochastic block models. Currently, the best known results for clustering mixtures of Gaussians are via these robust estimation techniques (Diakonikolas et al., 2018b; Kothari and Steinhardt, 2018; Hopkins and Li, 2018). Finally, high-dimensional biological datasets with structured outliers such as batch effects (Johnson et al., 2007; Leek et al., 2010), together with security concerns for machine learning systems (Steinhardt et al., 2017), motivate the study of robustness to worst-case outliers from an applied direction.

Recent research has shown encouraging progress on these questions, but the rapid progress has led to an opaque literature. Most papers are complex in isolation, but are in fact comprised of variations on a few key themes. This thesis aims to provide an accessible introduction to the area that highlights the major techniques. In Chapter 1, we introduce the basic problem of robust estimation, provide algorithms in the 1-dimensional case that foreshadow our later algorithms in high dimensions, and explain the basic difficulty of the high-dimensional setting. In Chapters 2 and 3, we focus on information-theoretic robustness—When is it possible (ignoring computational cost) to recover good parameter estimates in the presence of outliers? The picture here is pleasingly simple, based on a property called *resilience* that measures the influence of small sub-populations of points. Interestingly, resilience allows us to recover an estimate of the mean even when the *majority* of the points are outliers, assuming that we are allowed to output multiple guesses (the so-called *list-decodable setting* first introduced by Balcan et al. (2008)). This fundamental fact underlies the connection between robust learning and clustering, as we can think of each individual cluster as a population of "good" points and then regard the points from the remaining clusters as outliers.

In Chapter 4, we turn our attention to computationally efficient algorithms. Assuming the good points have bounded covariance, we can recover an estimate of the mean with an error that grows only with the largest eigenvalue of the covariance matrix (which is often independent of the dimension). The basic idea is that outliers that shift the mean by more than a small amount must create directions of large variation in the data, which can be detected by an eigendecomposition of the empirical covariance. We show how to extend this mean estimation result to general M-estimators as long as the gradient of the loss function has bounded covariance. Finally, in Chapter 5 we introduce an alternate computational approach based on duality. Using this approach, we can find approximate minimizers to a large family of saddle point problems in the presence of outliers. This allows us to recover similar mean estimation results as in Chapter 4, with the advantage that the results hold even when the majority of points are outliers. This yields algorithms for clustering that currently give the best known bounds. However, the techniques in Chapters 4 and 5 are both under active development. Both techniques are likely to enjoy stronger results even over the next year.

In summary, we will see a relatively complete information-theoretic perspective on robustness, as well as two approaches for designing efficient algorithms. These approaches are presented in general terms such that many key results in the field follow as simple corollaries, often requiring only about a page of algebra to check the conditions. We hope that by exposing the structure behind the arguments, we will enable new researchers to both apply and extend these results.

Acknowledgments

This thesis is dedicated in memory of Michael Cohen, who tragically passed away last September from undiagnosed type-1 diabetes. Michael was a bright-shining star in his field whose flame was extinguished far too early. While Michael's many significant research contributions are in the public record, his humility and generosity are not. On multiple occasions Michael discussed research with me and contributed important ideas, some of which would have merited co-authorship if Michael had not been too humble to accept. Three of these contributions touch upon the present work: first, Michael provided an alternate characterization of the mean estimation algorithm in Section 5.3.2 that substantially reduced its running-time; second, Michael pointed out the generalization of the BSS-based pruning argument appearing in Steinhardt et al. (2018); this argument plays a similar role to Theorem 3.1 for bounded-covariance distributions. Michael also pointed out the connection between strongly-convex norms and the concepts of type and co-type from functional analysis. Finally, while not on the topic of this thesis, Michael contributed substantially to one of my earlier papers on the exponentiated gradient algorithm. Many other researchers could compose similar lists of Michael's anonymous contributions. He will be sorely missed.

I also enclose an incomplete account of the many people who contributed to this work, either directly through ideas or discussions, or indirectly by providing support and feedback. Thanks especially to my advisor, Percy Liang, and to Greg Valiant and Moses Charikar, who spent enough time with me that they could well have been co-advisors. John Duchi and Chris Ré helped build the machine learning group almost from scratch and consistently provided helpful feedback on talks and papers. Tatsu Hashimoto, Michael Kim, Steve Mussmann, Arun Chaganty, Omer Reingold, Chris De Sa, Steven Bach, Paris Siminelakis, Roy Frostig, Xinkun Nie, Sorathan Chaturapruek, Brian Axelrod, and others diligently read paper drafts and acted as good citizens of the Stanford CS community. Daniel Selsam, Tudor Achim, Michael Webb, and Sida Wang were kindred spirits who regularly offered interesting perspectives on work and on life. Holden Karnofsky, Dario Amodei, Chris Olah, and Nick Beckstead challenged my ideas and pushed me to be better. Aditi Raghunathan and Pang Wei Koh showed extreme patience as my first junior collaborators, and multiplied my productivity in the last years of graduate school. Philippe Rigollet asked probing questions that led to the concept of

resilience. Pravesh Kothari and Tselil Schramm explained sum-of-squares programming to me many times until I finally understood it; Dan Kane, Jerry Li, and Gautam Kamath showed similar patience explaining their filtering algorithm. Thanks to these and other collaborators, including Zachary Lipton, Alistair Stewart, Ilias Diakonikolas, Jonathan Huggins, and David Steurer for putting up with me near deadlines. Alex Zhai, Paul Christiano, and Michael Cohen could well have been co-authors but were too humble to accept credit. Outside of research, many friends supported me throughout graduate school. Those not already mentioned above include Sindy Li, Nike Sun, Cathy Wu, Michela Meister, Mikaela Provost, Tim Telleen-Lawton, Jared Kaplan, Chris Roberts, Marta Shocket, Jon Losh, Michael Poon, Sasha Targ, Jeffrey Wu, Yang Hong, Rosa Cao, Yan Zhang, Danqi Chen, Will Fithian, Joy Zhang, Hamsa Sridhar Bastani, Osbert Bastani, Armin Pourshafeie, and many others. Finally, thanks to my family—my parents Suezette and Allan, my sister Emma, and my grandmother Sophia, who is no longer with us but helped raise me and spent many evenings solving word puzzles.

Contents

List of Algorithms

List of Figures

Chapter 1

Introduction

Our goal is to understand learning in the presence of **outliers**. As an example, consider the following two datasets:

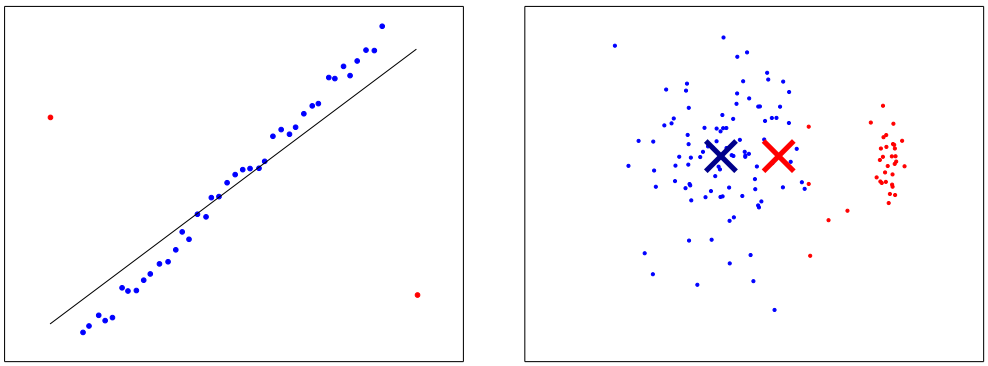

Figure 1.1: Two datasets with outliers.

For the regression dataset (left), the two outliers indicated in red cause the least squares line to differ from a more robust trend line that ignores these outliers. Similarly, for the mean estimation dataset (right), the red outliers skew the empirical mean.

The present work concerns methods whose **worst-case performance** in the presence of arbitrary outliers is good. We will see how to design and analyze such methods. In addition to worst-case performance, a large literature also studies average-case performance under statistical assumptions on the outliers. However, we do not study this scenario here.

1.1 Formal Setting

To formalize the notion of outliers, consider the following learning setting:

- We observe n data points $x_1, \ldots, x_n$.
- An unknown subset S of αn points are "good" (e.g. drawn from some "nice" distribution p^*).
- The remaining $(1-\alpha)n$ points are arbitrary outliers.

In particular, the outliers could be chosen by an adversary that has full knowledge of the good points and of whatever learning algorithm we choose to use.

In many settings, the fraction α of good data will be close to 1, in which case we often use the alternative notation $\epsilon = 1 - \alpha$ to refer to the fraction of outliers.

Typically, we are interested in approximately estimating some statistic of p^*, such as its mean, its best fit line (in the case of regression), a separating hyperplane (in the case of classification), and so on. We will first focus on mean estimation for simplicity, and later discuss how to generalize to more complex learning problems.

1.2 Robust Mean Estimation

In robust mean estimation, the n data points $x_1, \ldots, x_n$ lie in $\mathbb{R}^d$, and the αn good points are drawn from a distribution p^* with some unknown mean μ. Our goal is to output some estimate $\hat{\mu}$ of μ such that $\|\hat{\mu} - \mu\|$ is small in a given norm $\|\cdot\|$.

1.2.1 1-dimensional example

Consider the following histogram of a 1-dimensional dataset, where the height of each bar represents the number of points with a given value:

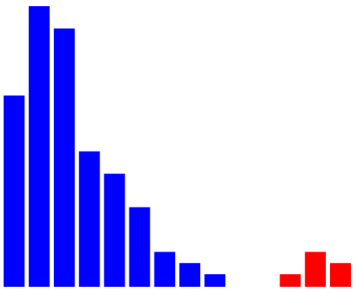

Figure 1.2: Histogram of a dataset with outliers.

Are the red points outliers? Or part of the real data? Depending on the conclusion, the estimated mean could vary substantially. Without further assumptions on the data-generating distribution p^*, we cannot rule out either case. This brings us to an important principle:

With no assumptions on the distribution p^*, robust estimation is impossible.

In particular, we must make assumptions that are strong enough to reject sufficiently extreme points as outliers, or else even a small fraction of such points can dominate the estimate of the mean.

1.2.2 First-pass assumption: bounded variance

We now consider one possible assumption that allows us to estimate the mean in the presence of outliers. For now, take p^* to be the empirical distribution over the $(1-\epsilon)n$ good points, so that we can ignore issues of finite-sample concentration (we will attend to these issues in Chapter 3).

Suppose that p^* is a distribution on the real line $\mathbb{R}$, with bounded variance in the sense that $\mathbb{E}_{x\sim p^*}[(x-\mu)^2] \le \sigma^2$, where μ is the (unknown) true mean of p^*. Then, given an ϵ-fraction of outliers, we can estimate μ to within error $\mathcal{O}(\sigma\sqrt{\epsilon})$. Indeed, consider the following procedure:

Algorithm 1 `TrimmedMean`

1: Remove the smallest and largest $2\epsilon n$ points (so $4\epsilon n$ points are removed in total).
2: Return the mean of the remaining points.

In the remainder of this subsection we will show that this algorithm succeeds at estimating the mean.

Analyzing Algorithm 1. We will make use of a strengthened version of Chebyshev's inequality, which we recall here (see Section A.1 for a proof):

Lemma 1.1. *Suppose that p has mean μ and variance σ^2. Then, $\mathbb{P}_{X\sim p}[X \ge \mu + \sigma/\sqrt{\delta}] \le \delta$. Moreover, if E is any event with probability at least δ, then $|\mathbb{E}_{X\sim p}[X \mid E] - \mu| \le \sigma\sqrt{\frac{2(1-\delta)}{\delta}}$.*

The first part, which is the standard Chebyshev inequality, says that it is unlikely for a point to be more than a few standard deviations away from μ. The second part says that any large population of points must have a mean close to μ. This second property, which is called *resilience*, is central to robust estimation, and will be studied in more detail in Chapter 2.

With Lemma 1.1 in hand, we can prove the following fact about Algorithm 1:

Proposition 1.2. *Assume the fraction ϵ of outliers is at most $\frac{1}{8}$. Then the output $\hat{\mu}$ of Algorithm 1 satisfies $|\hat{\mu} - \mu| \le 8\sigma\sqrt{\epsilon}$.*

Proof. First note that all outliers which exceed the ϵ-quantile of p^* are removed by Algorithm 1. Therefore, all non-removed outliers lie within $\frac{\sigma}{\sqrt{\epsilon}}$ of the mean μ by Chebyshev's inequality.

On the other hand, we remove at most $4\epsilon n$ good points (since we remove $4\epsilon n$ points in total), which accounts for at most a $\frac{4\epsilon}{1-\epsilon}$ fraction of the good points. Applying Lemma 1.1 with $\delta = 1 - \frac{4\epsilon}{1-\epsilon}$, the mean of the remaining good points lies within $\sigma\sqrt{\frac{8\epsilon}{1-5\epsilon}}$ of μ.

Now let ϵ' be the fraction of remaining points which are bad, and note that $\epsilon' \le \frac{\epsilon}{1-4\epsilon}$. The mean of all the remaining points differs from μ by at most $\epsilon' \cdot \sigma\sqrt{\frac{1}{\epsilon}} + (1-\epsilon') \cdot \sigma\sqrt{\frac{8\epsilon}{1-5\epsilon}}$, which is at most $\frac{4\sqrt{\epsilon}}{1-4\epsilon}\sigma$. This is in turn at most $8\sigma\sqrt{\epsilon}$ assuming that $\epsilon \le \frac{1}{8}$. □

Remark 1.3. The key fact driving the proof of Proposition 1.2 is that any $(1-\epsilon)$-fraction of the good points has mean at most $\mathcal{O}(\sigma\sqrt{\epsilon})$ away from the true mean due to Chebyshev's inequality (Lemma 1.1), which makes use of the bound σ^2 on the variance. Any other bound on the deviation from the mean would yield an analogous result. For instance, if p^* has bounded kth moment, then the $\mathcal{O}(\sigma\sqrt{\epsilon})$ in Lemma 1.1 can be improved to $\mathcal{O}(\tilde{\sigma}\epsilon^{1-1/k})$, where $\tilde{\sigma}^k$ is a bound on the kth moment; in this case Algorithm 1 will estimate μ with a correspondingly improved error of $\mathcal{O}(\tilde{\sigma}\epsilon^{1-1/k})$.

1.2.3 An alternative procedure: comparing mean and variance

We next analyze a more complicated procedure that will generalize to yield efficient algorithms beyond the one-dimensional setting. For instance, many of the efficient algorithms discussed in Chapter 4 follow a similar template.

This alternative procedure compares the variance of the data to some known upper bound σ^2 on what the variance "should be". If the variance is not too large, it outputs the empirical mean; otherwise, it down-weights points that are far away from the mean and re-estimates the variance.

Algorithm 2 `Filter1D`

1: Initialize weights $c_1, \ldots, c_n = 1$.
2: Compute the empirical mean $\hat{\mu}_c$ of the data, $\hat{\mu}_c \stackrel{\text{def}}{=} (\sum_{i=1}^n c_i x_i)/(\sum_{i=1}^n c_i)$.
3: Compute the empirical variance $\hat{\sigma}_c^2 \stackrel{\text{def}}{=} \sum_{i=1}^n c_i \tau_i / \sum_{i=1}^n c_i$, where $\tau_i = (x_i - \hat{\mu}_c)^2$.
4: If $\hat{\sigma}_c^2 \le 16\sigma^2$, output $\hat{\mu}_c$.
5: Otherwise, update $c_i \leftarrow c_i \cdot (1 - \tau_i/\tau_{\max})$, where $\tau_{\max} = \max_i \tau_i$.
6: Go back to line 2.

The intuition is as follows: if the empirical variance $\hat{\sigma}_c^2$ is much larger than the variance σ^2 of the good data, then the bad points must on average be very far away from the empirical mean (i.e., τ_i must be large on average). This is depicted in the diagram below:

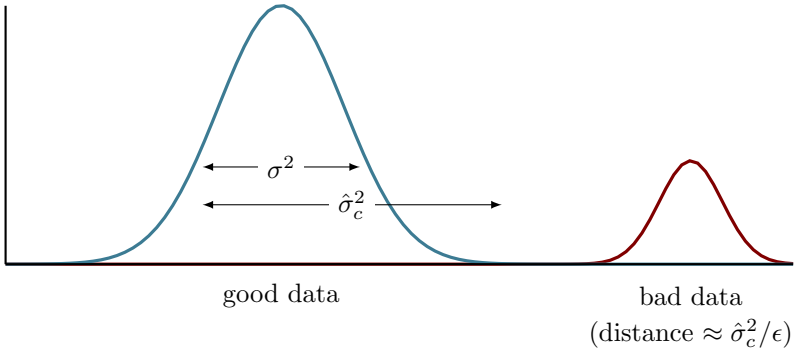

Figure 1.3: Intuition behind Algorithm 2. Because there is only an ϵ-fraction of bad data, it must lie far away to increase the variance by a constant factor.

The variables τ_i help to formalize this; one can calculate that

$$\frac{1}{|S|}\sum_{i\in S} c_i\tau_i \le \frac{1}{|S|}\sum_{i\in S}(x_i-\hat{\mu}_c)^2 = \sigma^2 + (\mu-\hat{\mu}_c)^2, \tag{1.1}$$

so that if $\sigma^2 \ll \hat{\sigma}_c^2$ then most of the contribution to the overall sum $\sum_{i=1}^n c_i\tau_i$ must come from the ϵ-fraction of bad points. The main complication is that the mean $\hat{\mu}_c$ may differ from the true mean μ of the good points, which leads to the $(\mu-\hat{\mu}_c)^2$ term in (1.1).

To bound $(\mu-\hat{\mu}_c)^2$, we will inductively assume that the following invariant holds:

$$\sum_{i\in S}(1-c_i) \le \frac{1-\epsilon}{2}\sum_{i\notin S}(1-c_i) \tag{$\mathcal{I}$}$$

The invariant ($\mathcal{I}$) posits that the amount of mass removed from the good points is smaller than that removed from the bad points. This ensures that the weights c_i are biased towards the good points, which intuitively should imply that $\hat{\mu}_c$ is close to μ. We formalize this with Lemma 1.4:

Lemma 1.4. *Suppose that the invariant ($\mathcal{I}$) holds. Then $|\mu-\hat{\mu}_c| \le \sigma\sqrt{\frac{2\epsilon}{2-\epsilon}} + \hat{\sigma}_c\sqrt{\frac{2\epsilon}{1-\epsilon}}$.*

Suppose further that $\hat{\sigma}_c^2 \ge 16\sigma^2$ and $\epsilon \le \frac{1}{12}$. Then we have

$$\sum_{i\in S} c_i\tau_i \le \frac{1-\epsilon}{3}\hat{\sigma}_c^2 n, \text{ while } \sum_{i\notin S} c_i\tau_i \ge \frac{2}{3}\hat{\sigma}_c^2 n. \tag{1.2}$$

The second part of Lemma 1.4 states that τ_i is large across the bad points and small across the good points; in particular, the sum of $c_i\tau_i$ across $[n]\backslash S$ is more than twice as large as the sum across S. Note that this means the *average* of $c_i\tau_i$ across $[n]\backslash S$ is roughly $\frac{2}{\epsilon}$ times larger than the average across S (since there are only an ϵ-fraction of bad points). The proof of Lemma 1.4 consists of straightforward but tedious calculation and is deferred to Section A.2.

Intuitively, Lemma 1.4 should give us the power to separate good points from bad points, by removing points for which τ_i is large. The difficulty is that Lemma 1.4 only controls the τ_i on average rather than pointwise. By appropriately downweighting (rather than removing) points as in line 5 of Algorithm 2, we can make use of this "on-average" information. The following lemma formalizes this, asserting that we remove bad points much more quickly than good points.

Lemma 1.5. *Suppose that τ_i is any quantity such that $\sum_{i\in S} c_i\tau_i \le \frac{1-\epsilon}{2}\sum_{i\notin S} c_i\tau_i$. Then, the update $c_i \leftarrow c_i(1-\tau_i/\tau_{\max})$ in Algorithm 2 preserves the invariant ($\mathcal{I}$), meaning that if ($\mathcal{I}$) holds before the update, it will continue to hold after the update.*

We prove Lemma 1.5 later in this section. The remaining analysis of Algorithm 2 now proceeds by induction, as Lemma 1.5 provides the necessary inductive step. We will show the following:

Proposition 1.6. *Suppose that $\epsilon \le \frac{1}{12}$ and that the variance of the good points is at most σ^2. Then, the output $\hat{\mu}_c$ of Algorithm 2 satisfies $|\hat{\mu}_c - \mu| \le \mathcal{O}(\sigma\sqrt{\epsilon})$.*

Proof. Our inductive hypothesis is the invariant ($\mathcal{I}$). This hold initially since $\sum_{i\in S}(1-c_i) = \sum_{i\notin S}(1-c_i) = 0$. Turning to the inductive step, until we output $\hat{\mu}_c$ we have $\hat{\sigma}_c^2 \geq 16\sigma^2$, so we can apply Lemma 1.4. In particular, the condition of Lemma 1.5 holds due to the conclusion (1.2) of Lemma 1.4. Therefore, the invariant ($\mathcal{I}$) is preserved, which completes the induction.

To conclude, note that $\hat{\sigma}_c \leq 4\sigma$ whenever we output $\hat{\mu}_c$. Apply Lemma 1.4 once more to obtain $|\hat{\mu}_c - \mu| \leq \sigma\sqrt{\frac{2\epsilon}{2-\epsilon}} + \hat{\sigma}_c\sqrt{\frac{2\epsilon}{1-\epsilon}} = \mathcal{O}(\sigma\sqrt{\epsilon})$, as was to be shown. □

We end this subsection by proving Lemma 1.5.

Proof of Lemma 1.5. Let $c_i' = c_i \cdot (1 - \tau_i/\tau_{\max})$. Then for any set I we have

$$\sum_{i\in I}(1-c_i') = \sum_{i\in I}(1-c_i) + \sum_{i\in I}(c_i - c_i') \tag{1.3}$$

$$= \sum_{i\in I}(1-c_i) + \frac{1}{\tau_{\max}}\sum_{i\in I} c_i\tau_i. \tag{1.4}$$

Applying this with $I = S$ and $I = [n]\backslash S$, we note that $\sum_{i\in S}(1-c_i) \leq \frac{1-\epsilon}{2}\sum_{i\notin S}(1-c_i)$ by the assumed invariant ($\mathcal{I}$), while $\sum_{i\in S} c_i\tau_i \leq \frac{1-\epsilon}{2}\sum_{i\notin S} c_i\tau_i$ by the assumption of the lemma. Therefore, the invariant ($\mathcal{I}$) continues to hold for the c_i'. □

1.3 The Challenge: High Dimensions

In the previous section, we saw two procedures for robustly estimating the mean of a 1-dimensional dataset, assuming the true data had bounded variance. These procedures work by removing data points that are too far away from the mean, and then returning the mean of the remaining points.

It is tempting to apply this same idea in higher dimensions—for instance, removing points that are far away from the mean in ℓ_2-distance. Unfortunately, this incurs large error in high dimensions.

To see why, consider the following simplified example. The distribution p^* over the true data is an isotropic Gaussian $\mathcal{N}(\mu, I)$, with unknown mean μ and independent variance 1 in every coordinate. In this case, the typical distance $\|x_i - \mu\|_2$ of a sample x_i from the mean μ is roughly $\sqrt{d}$, since there are d coordinates and x_i differs from μ by roughly 1 in every coordinate. (In fact, $\|x_i - \mu\|_2$ can be shown to concentrate around $\sqrt{d}$ with high probability.) This means that the outliers can lie at a distance $\sqrt{d}$ from μ without being detected, thus shifting the mean by $\Theta(\epsilon\sqrt{d})$; Figure 1.4 depicts this. Therefore, filtering based on ℓ_2 distance will necessarily incur an error of at least $\epsilon\sqrt{d}$. This dimension-dependent $\sqrt{d}$ factor often renders bounds meaningless.

In fact, the situation is even worse; not only are the bad points no further from the mean than the good points in ℓ_2-distance, they actually have the same probability density under the true data-generating distribution $\mathcal{N}(\mu, I)$. This means that there is no procedure that measures each point in isolation and can avoid the $\sqrt{d}$ factor in the error.

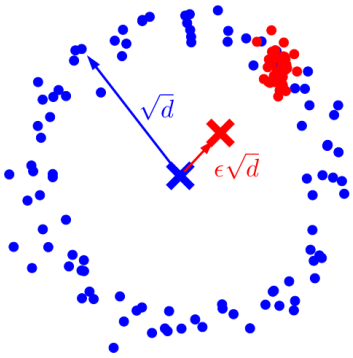

Figure 1.4: The outliers can lie at distance $\sqrt{d}$ without being detected, skewing the mean by $\epsilon\sqrt{d}$.

This leads us to an important take-away: *In high dimensions, outliers can substantially perturb the mean while individually looking innocuous.* To handle this, we will instead need to analyze entire populations of outliers at once. We will see in later chapters that this is possible, and that we can avoid dimension-dependent error through more nuanced strategies.

1.4 Learning with Majority Outliers

In robust statistics, the *breakdown point* refers to the maximum fraction of outliers that a procedure can tolerate before incurring arbitrarily high error. For instance, our analysis of `TrimmedMean` and `Filter1D` established that they have breakdown points of at least $\epsilon = \frac{1}{8}$ and $\frac{1}{12}$, respectively.

It would appear that the best possible breakdown point is $\epsilon = \frac{1}{2}$. Indeed, beyond this point a majority of the data are outliers, so it is difficult to distinguish the good data from potential outliers. For instance, if $\epsilon = \frac{1}{3}$ we might observe the following dataset consisting of 3 identical clusters:

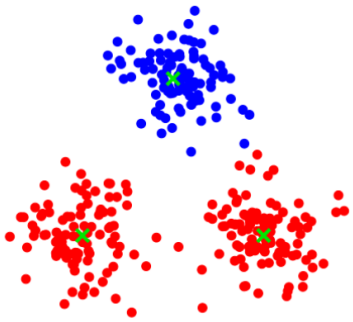

Figure 1.5: Three clusters, two of which consist of outliers.

Donoho (1982) formalizes this argument and shows that no translation-equivariant estimator can achieve a breakdown point better than $\frac{1}{2}$. However, we *can* surpass this barrier if we change the model slightly to account for ambiguities such as the one depicted above. For instance, we can adopt

the *list-decodable learning model* (Balcan et al., 2008), which allows us to output multiple candidate answers:

Definition 1.7 (List-decodable model)**.** In the list-decodable model, we are allowed to output m estimates $\hat{\theta}_1, \dots, \hat{\theta}_m$ of a target parameter θ^*, and succeed if at least one of the $\hat{\theta}_j$ is close to θ^*.

In the example from Figure 1.5, we can output 3 candidate means (the 3 green `x`'s) and be sure that at least one is close to the true mean. A fuller treatment of the list-decodable model will appear in Chapter 2. For now, we will show that in one dimension, the same bounded variance assumption from before is sufficient to enable robust learning.

More specifically, assume that an α-fraction of the data is good data that has variance at most σ^2 around its mean μ. Before, we took α to be $1-\epsilon$ to emphasize that $\alpha \approx 1$, but now we are interested in the case where α is small and potentially less than $\frac{1}{2}$. We have the following result:

Theorem 1.8. *Consider 1-dimensional data lying on the real line. Suppose that an α-fraction of the data is good, while the remaining data consists of arbitrary outliers. Also suppose that the good data has variance at most σ^2 around its mean μ. Then it is possible to output $m \le \lfloor \frac{1}{\alpha(1-\delta)} \rfloor$ candidate means $\hat{\mu}_1, \dots, \hat{\mu}_m$ such that $\min_{j=1}^m |\hat{\mu}_j - \mu| = \mathcal{O}(\sigma/\sqrt{\delta})$.*

Note that the $\frac{1}{\alpha}$ factor in the bound on m is optimal, since an adversary can always create $\frac{1}{\alpha}$ identical clusters with different means.

Proof of Theorem 1.8. Let a and b denote the $\delta/2$ and $1-\delta/2$ quantiles of the good data, respectively. By Lemma 1.1 we know that a and b are both within $2\sigma/\sqrt{\delta}$ of μ. Now split the data into $m+1 = \lceil \frac{1}{\alpha(1-\delta)} \rceil$ intervals each containing at least a $\alpha(1-\delta)$-fraction of the points. Let $\hat{\mu}_1, \dots, \hat{\mu}_m$ be the m boundaries of these $m+1$ intervals:

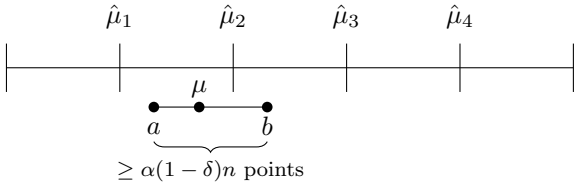

Figure 1.6: Illustration of the $m+1$ intervals and the m candidate means $\hat{\mu}_j$.

At least one of these boundaries must lie between a and b (as otherwise a and b would lie in the same interval, which is impossible because there are $\alpha(1-\delta)$ points between a and b). Therefore, one of the $\hat{\mu}_j$ lies between a and b, and in particular is within $\mathcal{O}(\sigma/\sqrt{\delta})$ of μ, as claimed. □

Remark 1.9. The $1/\sqrt{\delta}$ dependence can be improved to $\delta^{-1/k}$ if instead of bounded variance we have bounded kth moments. This roughly reflects that events of probability δ can differ from the

true mean by up to $\delta^{-1/k}$. In contrast, in Proposition 1.2 we cared about events of probability $1-\delta$, which can differ from the true mean by $\delta^{1-1/k}$ (see Remark 1.3).

Remark 1.10. The lack of dependence on α is specific to the 1-dimensional setting. In higher dimensions the best error guarantee assuming bounded covariance is $\mathcal{O}(1/\sqrt{\alpha})$ (Charikar et al., 2017), which has a matching lower bound proved in Proposition 5.4 of Diakonikolas et al. (2018b).

1.4.1 Connection: Agnostic Learning of Mixtures

Theorem 1.8 is interesting because it shows that there are procedures with breakdown points better than $\frac{1}{2}$ (and indeed arbitrarily close to 1) if we redefine the problem via the list-decodable model. Beyond its conceptual interest, list-decoding has important implications for the problem of learning mixture models.

Classically, the problem of learning mixtures is the following: there are k distributions p_1^*, …, p_k^*, and we observe samples from $w_1 p_1^* + \cdots + w_k p_k^*$, where the w_j are weights summing to 1. The goal is to disentangle the different mixture components and estimate statistics of each of the p_j^*. For instance, a common assumption is that each of the p_j^* is a Gaussian, and the goal is to estimate each of their means and variances.

List-decodable learning provides one way of learning mixtures; by setting $\alpha = \min_{j=1}^{k} w_j$, we can think of each of the mixture components as the "good" data, whence Theorem 1.8 guarantees that the means of each of the mixture components is close to at least one of the $\hat{\mu}_j$. This reduction is illustrated below:

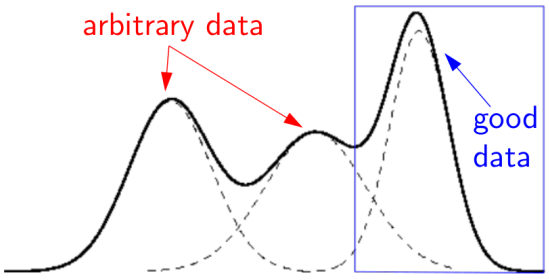

Figure 1.7: Reduction from learning mixtures to list-decodable robust learning.

The connection between learning mixtures and robust learning becomes most interesting when we move beyond the 1-dimensional case. For instance, clustering is NP-hard even in 2 dimensions (Mahajan et al., 2009), but generalizations of Theorem 1.8 will yield algorithms for efficient clustering in high dimensions under appropriate assumptions. Moreover, these algorithms work in the *robust agnostic setting* where the underlying model may be mis-specified (e.g. the data may not actually be Gaussian) and where a large fraction of outlier points do not come from any mixture component.

1.5 Beyond Mean Estimation

We have mainly focused on mean estimation so far, but robust learning applies to many other settings as well (some of which will be discussed in the sequel). For instance, we also discussed robust clustering above, and will later discuss stochastic block models, which are related to clustering on graphs. We can also consider robust *classification*, where some small fraction of data are arbitrarily misclassified. Another problem is *item frequency estimation*, where we observe samples from a discrete distribution π and wish to determine which items are most frequent. Finally, we can consider general M-estimation or stochastic optimization, where each observation is a (typically convex) loss function corresponding to an individual data point, and the goal is to find an approximate minimizer of the average loss over the good points. We will treat many of these in detail in the sequel.

1.6 History

Robust estimation was first systematically studied by Tukey and his students (Tukey, 1960), with other contributions by Box (1953), Huber (1964), and Hampel (1968). There were even earlier investigations going back to at least Newcomb (1886). Huber and Ronchetti (2009) and Hampel et al. (2011) provide two recent surveys. These investigations study properties including breakdown point (introduced in Section 1.4) and influence (the extent to which a single datum affects the answer), but were largely focused on low or moderate rather than high dimensions. The learning theory community has also studied learning with errors (Kearns and Li, 1993), although often under more restrictive assumptions on the outliers.

High-dimensional robust estimation was studied by Maronna (1976), Donoho (1982), and Donoho and Gasko (1992) (among others). These works noted obstacles to obtaining robust estimates in high dimensions, while Donoho (1982) analyzed estimators with favorable breakdown points, including the Tukey median (Tukey, 1975). Despite a good breakdown point, most of these estimators still incur error growing with the dimension. An exception is the Tukey median, which has been characterized as an optimally robust estimator, but this is imprecise–it overlooks the fact that the Tukey median is only consistent under strong symmetry assumptions, and is NP-hard to compute (Johnson and Preparata, 1978). Diakonikolas et al. (2016) show that a number of natural approaches fail in high dimensions, and provide a higher-dimensional analog of the `Filter1D` procedure, which we will cover in later chapters.

The list-decodable learning model was first introduced in Balcan et al. (2008) and later studied by others including Balcan et al. (2009) and Kushagra et al. (2016). It was first applied in the context of robust learning by Charikar et al. (2017), which built on earlier work in Steinhardt et al. (2016).

1.7 Exercises

Robust estimators under higher moment bounds

1. [1] Show that if the kth moment $\mathbb{E}_{x\sim p^*}[|x-\mu|^k]^{1/k}$ is bounded by σ_k, then the error of the `TrimmedMean` procedure is $\mathcal{O}(\sigma_k \epsilon^{1-1/k})$.

2. [1] Suppose that p^* is sub-Gaussian, meaning that $\mathbb{E}_{x\sim p^*}[\exp(\lambda(x-\mu))] \leq \exp(\frac{1}{2}\lambda^2\sigma^2)$ for all $\lambda \in \mathbb{R}$. Show that the error of `TrimmedMean` is $\mathcal{O}(\sigma\epsilon\sqrt{\log(2/\epsilon)})$.

3. [1+] Suppose that p^* has bounded 4th moments: $\mathbb{E}_{x\sim p^*}[(x-\mu)^4]^{1/4} \leq \sigma_4$. Devise a variant of `Filter1D` that achieves error $\mathcal{O}(\sigma_4\epsilon^{3/4})$.

Better breakdown point via resilience

Call a set $T \subseteq \mathbb{R}$ of points (σ, ϵ)-resilient if every subset of $(1-\epsilon)|T|$ of the points has mean within σ of the overall mean.

4. [1+] If a set T has variance bounded by σ^2, show that it is $(\mathcal{O}(\sigma\sqrt{\epsilon}), \epsilon)$-resilient for $\epsilon < \frac{1}{2}$.

5. [2+] Let $\epsilon < \frac{1}{2}$. Suppose that we are given a set T of $(1-\epsilon)n$ good points that is $(\sigma, \frac{\epsilon}{1-\epsilon})$-resilient with mean μ, together with ϵn arbitrary outliers. Let T' be any $(\sigma, \frac{\epsilon}{1-\epsilon})$-resilient subset of the n points with $|T'| \geq (1-\epsilon)n$. Show that the mean of T' is within 2σ of μ. *(Hint: consider the mean of $T \cap T'$.)*

Median and Tukey median

We say that p^* is (s, ϵ)-stable if $\mathbb{P}_{x\sim p^*}[x \geq \mu + s] < \frac{1}{2} - \epsilon$ and $\mathbb{P}_{x\sim p^*}[x \leq \mu - s] < \frac{1}{2} - \epsilon$. Let $s(\epsilon)$ denote the minimum s for which p^* is (s, ϵ)-stable (or the infimum if the minimum does not exist).

6. [1] Show that $s(0) = 0$ if and only if the median is unique and equals the mean.

7. [1] Show that the median estimates the mean with error at most $s(\frac{\epsilon}{2-2\epsilon})$ in the presence of an ϵ-fraction of outliers.

8. [2] Show that a Gaussian with variance σ^2 is $(\mathcal{O}(\sigma\epsilon), \epsilon)$-stable for $\epsilon \leq \frac{1}{4}$.

9. [2] Show that if a distribution is (s, ϵ)-stable, then the empirical distribution of n i.i.d. samples from p^* will be $(s, \frac{\epsilon}{2})$-stable with probability at least $1 - 2\exp(-c\epsilon n)$ for some $c > 0$.

10. [3] Call a distribution on $\mathbb{R}^d$ (s, ϵ)-stable if it is (s, ϵ)-stable when projected onto any unit vector. Show that if a distribution on $\mathbb{R}^d$ is (s, ϵ)-stable, then the empirical distribution on n samples is $(2s, \frac{\epsilon}{2} - \mathcal{O}(\frac{d}{n}))$-stable with probability at least $1 - 2\exp(-c\epsilon n)$.

Given data $x_1, \ldots, x_n \in \mathbb{R}$, the *depth* of a point x_i is the minimum of the number of points to the left and to the right of x_i. For $x_1, \ldots, x_n \in \mathbb{R}^d$, the depth of a point x_i is the minimum depth along all 1-dimensional projections. The *Tukey median* is the point with maximum depth.

11. [1+] Let $S \subseteq \{x_1, \ldots, x_n\}$ be a set of $(1-\epsilon)n$ "good" points. Show that if there is a point of depth $\frac{1-\delta}{2}|S|$ in S, then there is a point of depth $\frac{1-\delta-\epsilon}{2}n$ in the overall data.

12. [1+] Suppose that the good data are $(s, c \cdot (\epsilon + \delta))$-stable in every direction for some sufficiently large constant c. Under the same assumptions as the previous problem, show that the Tukey median estimates the mean to ℓ_2-error s.

13. [2+] Show that given $(1-\epsilon)n$ samples from $\mathcal{N}(\mu, \sigma^2 I)$, and an ϵ-fraction of outliers, the Tukey median estimates μ with ℓ_2-error $\mathcal{O}(\sigma\epsilon)$ with high probability, assuming that ϵ is sufficiently small and $n \gg \frac{d}{\epsilon}$. *(Hint: use the results of the previous exercises.)*

Breaking common estimators

14. [2+] Given data points $x_1, \ldots, x_n \in \mathbb{R}^d$, the *geometric median* is the estimator $\hat{\mu}$ that minimizes $\sum_{i=1}^n \|x_i - \hat{\mu}\|_2$. Suppose that a $(1-\epsilon)$-fraction of points are drawn from $\mathcal{N}(\mu, I)$, while the remaining points are arbitrary outliers. Show that the geometric median can have error $\|\hat{\mu} - \mu\|_2 = \Omega(\epsilon\sqrt{d})$.

15. [2] Consider data points $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$. A $(1-\epsilon)$-fraction of good points are generated as follows: $x \sim \mathcal{N}(0, I)$, and $y = \langle w^*, x \rangle + v$, where $v \sim \mathcal{N}(0, 1)$ and $w^* \in \mathbb{R}^d$ is a parameter we wish to estimate. The remaining points are arbitrary outliers.

    Consider an estimator $\hat{w}$ that first discards all points i for which $\|x_i\|_2 > 2\sqrt{d}$, and then runs least squares regression on the remaining points. Show that $\hat{w}$ can have error $\|\hat{w} - w^*\|_2 = \Omega(\epsilon\sqrt{d})$.

Chapter 2

Information-Theoretic Results

In the previous chapter we mainly focused on robust estimation in one dimension, while discussing some difficulties in obtaining good estimates in higher dimensions. In this chapter we will handle the higher-dimensional setting. Our results will be information-theoretic in nature, with efficient algorithms presented in Chapters 4 and 5. As before, we will restrict out attention to mean estimation in some norm. The exposition in this chapter closely follows that in Steinhardt et al. (2018).

2.1 Resilience

The key to our results is the property of *resilience*. A set S is resilient if the the mean of every large subset of S is close to the mean of all of S. More formally, for a norm $\|\cdot\|$, resilience is defined as follows:

Definition 2.1 (Resilience)**.** A set of points $\{x_i\}_{i\in S}$ lying in $\mathbb{R}^d$ is (σ,ϵ)-*resilient* in a norm $\|\cdot\|$ if, for all subsets $T \subseteq S$ of size at least $(1-\epsilon)|S|$,

$$\left\| \frac{1}{|T|} \sum_{i\in T} (x_i - \mu) \right\| \leq \sigma, \tag{2.1}$$

where μ is the mean of S. More generally, a distribution p is said to be (σ,ϵ)-resilient if $\|\mathbb{E}[x-\mu \mid E]\| \leq \sigma$ for every event E of probability at least $1-\epsilon$.

As the most basic example, an isotropic Gaussian is resilient:

Example 2.2 (Isotropic Gaussian)**.** Let $p = \mathcal{N}(\mu, I)$ be an isotropic Gaussian distribution. Then p is $(\mathcal{O}(\epsilon\sqrt{\log(2/\epsilon)}),\epsilon)$-resilient in the ℓ_2-norm.

Proof. Let $\hat{\mu}$ be the mean of the worst-case event E (i.e. the event E with $\mathbb{P}[E] \geq 1-\epsilon$ for which $\|\hat{\mu}-\mu\|_2$ is largest). By rotational symmetry of p, we can assume that $\hat{\mu}-\mu$ lies in the direction of

the vector $(1, 0, 0, \ldots)$. But then it is clear that the worst-case E consists of taking the $1 - \epsilon$ quantile of points whose first coordinate is largest, as shown below:

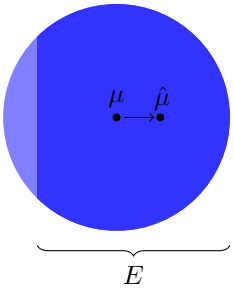

Figure 2.1: Worst-case shift in mean for an isotropic Gaussian.

Resilience thus reduces to the one-dimensional case and the question is how much the mean of a Gaussian shifts upon removing its ϵ-quantile of smallest values. Simple calculation reveals this to be $\mathcal{O}(\epsilon\sqrt{\log(2/\epsilon)})$, yielding the claimed result. □

More generally, if p has kth moments bounded by σ_k, then p is $(\mathcal{O}(\sigma_k \epsilon^{1-1/k}), \epsilon)$-resilient (see Example 2.7). We detail more examples where Definition 2.1 is satisfied in Section 2.2 and Section 3.1. As additional intuition for what resilient sets look like, consider the following diagram:

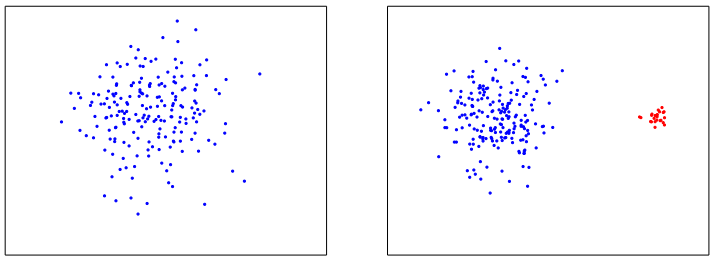

Figure 2.2: Resilient set (left) and non-resilient set (right).

The set on the left is resilient because all of the points are clustered around the mean, and hence removing any small population of points will not shift the mean by much. In contrast, the set on the right is not resilient (with any small parameter σ) due to the population of red points that are all far from the mean.

A key fact is that even in high dimensions, distributions can be resilient with a parameter σ that does not grow with the dimension, as in Example 2.2. This is despite the fact that (as shown in Section 1.3) all of the individual points are at distance $\sqrt{d}$ from the mean. Even though individual points are far away from the mean, they are far away *in different directions* and so the behavior of populations of points is substantially different than individual points. This is what allows us to circumvent the $\sqrt{d}$ barrier from Section 1.3—as we will see in the following subsection, (σ, ϵ)-resilience is sufficient to enable robust estimation with error $\mathcal{O}(\sigma)$.

2.1.1 Resilience Implies Robustness

Let $\tilde{S}$ be a set of data corrupted by outliers. Assuming that a subset $S \subseteq \tilde{S}$ of good points is (σ, ϵ)-resilient, we will see that there is a simple strategy for approximately recovering the mean of S—find *any* large (σ, ϵ)-resilient subset S' of the corrupted set $\tilde{S}$, and output the mean of S'. Thus, the mere existence of a resilient set S means that all resilient sets must have similar means.

The reason why is the following—since S' and S are both large, they must have large intersection, and so they must have similar means due to the condition (2.1) applied with $T = S \cap S'$. This argument is illustrated in Figure 2.3 below. We establish the claim formally in Proposition 2.3.

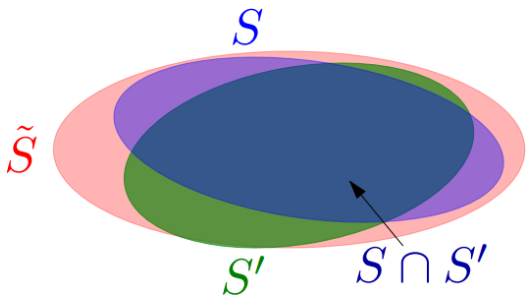

Figure 2.3: Large resilient sets have large overlap, and hence similar means.

Pleasingly, resilience reduces the question of handling outliers to a purely algorithmic question—that of finding any large resilient set. Rather than wondering whether it is even information-theoretically possible to estimate μ, we can instead focus on efficiently finding resilient subsets of $\tilde{S}$ (which is the purview of Chapters 4 and 5).

We next formalize the above argument to show that resilience is indeed information-theoretically sufficient for robust recovery of the mean μ. In what follows, we use $\sigma^*(\epsilon)$ to denote the smallest σ such that S is (σ, ϵ)-resilient.

Proposition 2.3. *Suppose that $\tilde{S} = \{x_1, \ldots, x_n\}$ contains a set S of size $(1-\epsilon)n$ that is resilient with mean μ (where S and μ are both unknown). Then if $\epsilon < \frac{1}{2}$, it is possible to recover a $\hat{\mu}$ such that $\|\hat{\mu} - \mu\| \le 2\sigma^*(\frac{\epsilon}{1-\epsilon})$.*

In other words, robustness to an ϵ fraction of outliers depends on resilience to a $\frac{\epsilon}{1-\epsilon}$ fraction of deletions. Thus, we can estimate μ in the presence of outliers as long as $\sigma^*(\frac{\epsilon}{1-\epsilon})$ is small.

Proof of Proposition 2.3. We prove Proposition 2.3 via a constructive (albeit exponential-time) algorithm. To prove the first part, suppose that the true set S is $(\sigma, \frac{\epsilon}{1-\epsilon})$-resilient with mean μ, and let S' be any set of size $(1-\epsilon)n$ that is $(\sigma, \frac{\epsilon}{1-\epsilon})$-resilient (with some potentially different mean μ'). We claim that μ' is sufficiently close to μ.

Indeed, let $T = S \cap S'$, which by the pigeonhole principle has size at least $(1-2\epsilon)n = \frac{1-2\epsilon}{1-\epsilon}|S| =$

$(1 - \frac{\epsilon}{1-\epsilon})|S|$. Therefore, by the definition of resilience,

$$\left\| \frac{1}{|T|} \sum_{i \in T} (x_i - \mu) \right\| \leq \sigma. \tag{2.2}$$

Thus, if we let $\mu_{S \cap S'}$ denote the mean of $S \cap S'$, we have $\|\mu - \mu_{S \cap S'}\| \leq \sigma$. But by the same argument, $\|\mu' - \mu_{S \cap S'}\| \leq \sigma$ as well. By the triangle inequality, $\|\mu - \mu'\| \leq 2\sigma$, which completes the proof. □

2.1.2 List-Decodable Learning with a Majority of Outliers

In Section 1.4, we saw that it is possible to estimate the mean even when there are a majority of outliers, as long as we measure success in the list-decodable model, where we can output m (typically $\mathcal{O}(1/\alpha)$) candidate answers. Here we generalize this observation to higher dimensions, and show that resilience is sufficient for learning in the list-decodable model.

The basic intuition is that we can cover the corrupted set $\tilde{S}$ by resilient sets $S'_1, \ldots, S'_{2/\alpha}$ of size $\frac{\alpha}{2}n$. Then by the pigeonhole principle, the resilient set S must have large overlap with at least one of the S', and hence have similar mean. This is captured in Figure 2.4 below:

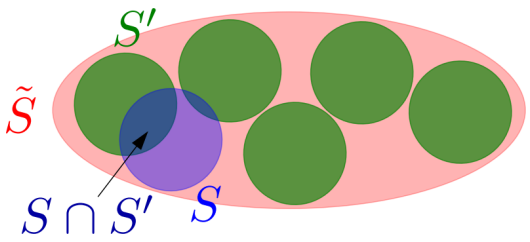

Figure 2.4: If we cover $\tilde{S}$ by resilient sets, at least one of the sets S' has large intersection with S.

The main difference is that S and S' may have relatively small overlap (in a roughly α-fraction of elements). We thus need to care about resilience when the subset T is small compared to S. The following lemma relates resilience on large sets to resilience on small sets:

Lemma 2.4. *For any $0 < \epsilon < 1$, a distribution/set is (σ, ϵ)-resilient if and only if it is $(\frac{1-\epsilon}{\epsilon}\sigma, 1-\epsilon)$-resilient.*

This result follows directly from the definition and is proved in Section B.1. With Lemma 2.4 in hand, we can prove an analog of Proposition 2.3 in the list-decoding model:

Proposition 2.5. *As in Proposition 2.3, suppose that a set $\tilde{S} = \{x_1, \ldots, x_n\}$ contains a resilient set S with mean μ. Then if $|S| \geq \alpha n$ (even if $\alpha < \frac{1}{2}$), it is possible to output $m \leq \frac{2}{\alpha}$ estimates $\hat{\mu}_1, \ldots, \hat{\mu}_m$ such that $\|\hat{\mu}_j - \mu\| \leq \frac{8}{\alpha}\sigma^*(\frac{\alpha}{4})$ for some j.*

Proof of Proposition 2.5. Given Lemma 2.4, the proof of Proposition 2.5 is similar to Proposition 2.3, but requires us to consider multiple resilient sets S_i rather than a single S'. Suppose S is $(\sigma, \frac{\alpha}{4})$-resilient around μ—and thus also $(\frac{4}{\alpha}\sigma, 1 - \frac{\alpha}{4})$-resilient by Lemma 2.4—and let $S_1, \ldots, S_m$ be a

maximal collection of subsets of $[n]$ such that:

1. $|S_j| \geq \frac{\alpha}{2}n$ for all j.
2. S_j is $(\frac{4}{\alpha}\sigma, 1-\frac{\alpha}{2})$-resilient (with mean μ_j).
3. $S_j \cap S_{j'} = \emptyset$ for all $j \neq j'$.

Clearly $m \leq \frac{2}{\alpha}$. We claim that S has large intersection with at least one of the S_j and hence μ_j is close to μ. By maximality of the collection $\{S_j\}_{j=1}^m$, it must be that $S_0 = S\backslash(S_1 \cup \cdots \cup S_m)$ cannot be added to the collection. First suppose that $|S_0| \geq \frac{\alpha}{2}n$. Then S_0 is $(\frac{4}{\alpha}\sigma, 1-\frac{\alpha}{2})$-resilient (because any subset of $\frac{\alpha}{2}|S_0|$ points in S_0 is a subset of at least $\frac{\alpha}{4}|S|$ points in S). This contradicts the maximality of $\{S_j\}_{j=1}^m$, so we must have $|S_0| < \frac{\alpha}{2}n$.

Now, this implies that $|S \cap (S_1 \cup \cdots \cup S_m)| \geq \frac{\alpha}{2}n$, so by pigeonhole we must have $|S \cap S_j| \geq \frac{\alpha}{2}|S_j|$ for some j. Letting $T = S \cap S_j$ as before, we find that $|T| \geq \frac{\alpha}{2}|S_j| \geq \frac{\alpha}{4}|S|$ and hence by resilience of S_j and S we have $\|\mu - \mu_j\| \leq 2 \cdot (\frac{4}{\alpha}\sigma) = \frac{8}{\alpha}\sigma$ by the same triangle inequality argument as before. □

Summary. We have now seen that resilience yields robustness both for a small fraction ϵ of outliers, and in the list-decodable setting when there is only a small fraction α of good points. Beyond robustness, this latter result provides a means for clustering data drawn from a mixture of distributions (see Section 1.4.1). We next examine several examples of resilient distributions and the implications of Proposition 2.3 and 2.5.

2.2 Examples of Resilient Distributions

Beyond isotropic Gaussians (Example 2.2), there are a number of distributional assumptions that imply resilience. First, a general characterization of resilience is that a distribution is resilient in a norm $\|\cdot\|$ if and only if it has *bounded tails* in the dual norm $\|\cdot\|_*$. (Recall that the dual norm to a norm $\|\cdot\|$ is defined via $\|v\|_* = \sup_{\|x\| \leq 1}\langle v, x\rangle$.)

Lemma 2.6. *For a fixed vector v, let $\tau_\epsilon(v)$ denote the ϵ-quantile of $\langle x-\mu, v\rangle$: $\mathbb{P}_{x\sim p}[\langle x-\mu, v\rangle \geq \tau_\epsilon(v)] = \epsilon$. Then, p is (σ, ϵ)-resilient in a norm $\|\cdot\|$ if and only if the ϵ-tail of p has bounded mean when projected onto any dual unit vector v:*

$$\mathbb{E}_p[\langle x-\mu, v\rangle \mid \langle x-\mu, v\rangle \geq \tau_\epsilon(v)] \leq \frac{1-\epsilon}{\epsilon}\sigma \text{ whenever } \|v\|_* \leq 1. \tag{2.3}$$

In particular, the ϵ-quantile satisfies $\tau_\epsilon(v) \leq \frac{1-\epsilon}{\epsilon}\sigma$.

In other words, if we project onto any unit vector v in the dual norm, the ϵ-tail of $x-\mu$ must have mean at most $\frac{1-\epsilon}{\epsilon}\sigma$. Thus, for instance, a distribution with variance at most σ_0^2 along every unit vector would have $\sigma = \mathcal{O}(\sigma_0\sqrt{\epsilon})$, since the ϵ-tail is bounded by $\mathcal{O}(\sigma_0/\sqrt{\epsilon})$ by Chebyshev's inequality. Lemma 2.6 is proved in Section B.2 (see also Section 2.3 for some intuition).

In the remainder of this section we give several more concrete examples, many of which leverage Lemma 2.6. We will focus on establishing resilience of the population distribution; establishing finite-sample concentration is more technical and is treated in detail in Chapter 3.

Example 2.7 (Bounded moments)**.** Suppose that a distribution p^* on $\mathbb{R}^d$ has bounded kth moments, in the sense that $\mathbb{E}[|\langle x-\mu, v\rangle|^k] \le \sigma_k^k \|v\|_2^k$ for all vectors $v \in \mathbb{R}^d$. Then, p^* is resilient in the ℓ_2-norm with $\sigma^*(\epsilon) \le 2\sigma_k \epsilon^{1-1/k}$ for $\epsilon \le \frac{1}{2}$, and $\sigma^*(1-\epsilon) \le \epsilon^{-1/k}\sigma_k$.

Proof. We will apply Lemma 2.6. Note that the dual of the ℓ_2-norm is again the ℓ_2-norm. Now consider any ℓ_2 unit vector v; we have

$$\mathbb{E}[\langle x-\mu, v\rangle \mid \langle x-\mu, v\rangle \ge \tau_\epsilon(v)] \le (\mathbb{E}[|\langle x-\mu, v\rangle|^k \mid \langle x-\mu, v\rangle \ge \tau_\epsilon(v)])^{1/k} \tag{2.4}$$

$$\le (\frac{1}{\epsilon}\mathbb{E}[|\langle x-\mu, v\rangle|^k])^{1/k} \le \epsilon^{-1/k}\sigma_k. \tag{2.5}$$

Therefore, by Lemma 2.6 we have $\sigma^*(\epsilon) \le \frac{\epsilon}{1-\epsilon} \cdot \epsilon^{-1/k}\sigma_k \le \frac{\epsilon^{1-1/k}}{1-\epsilon}\sigma_k$. Thus in particular $\sigma^*(\epsilon) \le 2\epsilon^{1-1/k}\sigma_k$ for $\epsilon \le \frac{1}{2}$, and by Lemma 2.4 we also have $\sigma^*(1-\epsilon) \le \epsilon^{-1/k}\sigma_k$. □

Example 2.8 (Item frequency estimation)**.** Let π be a distribution on $\{1, \dots, m\}$, and let $F_k(\pi)$ be the distribution on $[0,1]^m$ obtained by sampling k i.i.d. draws from π and taking the empirical frequency. (For instance, if $m = 5$ and $k = 3$, the samples $(2, 4, 2)$ would yield the frequency vector $(0, \frac{2}{3}, 0, \frac{1}{3}, 0)$.) Then $F_k(\pi)$ is resilient in the ℓ_1-norm with $\sigma^*(\epsilon) \le 6\epsilon\sqrt{\log(1/\epsilon)/k}$ for $\epsilon \le \frac{1}{2}$, and $\sigma^*(1-\epsilon) \le 3\sqrt{\log(1/\epsilon)/k}$.

Proof. As before, we will apply Lemma 2.6. The dual of the ℓ_1-norm is the ℓ_∞-norm, so we need to bound the tails of $F_k(\pi)$ when projected onto any ℓ_∞ unit vector v.

Now consider a sample $y \sim \pi$, interpreted as an indicator vector in $\{0,1\}^m$. We will analyze the moment generating function $\mathbb{E}[\exp(v^\top(y-\pi))]$ in order to get a tail bound on $y - \pi$. Provided $\|v\|_\infty \le 1$, the moment generating function is bounded as

$$\mathbb{E}[\exp(v^\top(y-\pi))] = \exp(-v^\top\pi)\sum_{j=1}^m \pi_j \exp(v_j) \tag{2.6}$$

$$\overset{(i)}{\le} \exp(-v^\top\pi)\sum_{j=1}^m \pi_j(1 + v_j + v_j^2) \tag{2.7}$$

$$= \exp(-v^\top\pi)\big(1 + \sum_{j=1}^m \pi_j(v_j + v_j^2)\big) \tag{2.8}$$

$$\overset{(ii)}{\le} \exp(-v^\top\pi)\exp(\sum_{j=1}^m \pi_j(v_j + v_j^2)) = \exp(\sum_{j=1}^m \pi_j v_j^2). \tag{2.9}$$

Here (i) uses the fact that $\exp(z) \le 1 + z + z^2$ for $|z| \le 1$, while (ii) is simply the inequality

$1 + z \le \exp(z)$. In particular, (2.9) implies that for any ℓ_∞ unit vector v and $c \in [-1, 1]$ we have $\mathbb{E}[\exp(cv^\top(y - \pi))] \le \exp(c^2 \sum_{j=1}^m \pi_j) = \exp(c^2)$.

Now, let x be an average of k independent samples from π. We claim that the moment generating function of x satisfies $\mathbb{E}[\exp(cv^\top(x - \pi)] \le \exp(2c^2/k)$. Indeed, the previous result implies that $\mathbb{E}[\exp(cv^\top(x-\pi))] \le \exp(c^2/k)$ for $c \in [-k, k]$. Then note that $v^\top(x-\pi) \le 2$, and so $\mathbb{E}[\exp(cv^\top(x-\pi))] \le \exp(2c^2/k)$ for all $|c| \ge k$ as well. Hence $\mathbb{E}[\exp(cv^\top(x - \pi))] \le \exp(2c^2/k)$ for all c.

Now, let E be any event of probability ϵ. We have

$$\mathbb{E}[v^\top(x - \pi) \mid E] \le \frac{1}{c} \log(\mathbb{E}[\exp(cv^\top(x - \pi)) \mid E]) \tag{2.10}$$

$$\le \frac{1}{c} \log(\frac{1}{\epsilon}\mathbb{E}[\exp(cv^\top(x - \pi))]) \tag{2.11}$$

$$\le \frac{1}{c} \log(\frac{1}{\epsilon} \exp(2c^2/k)) \tag{2.12}$$

$$= \frac{\log(1/\epsilon) + 2c^2/k}{c}. \tag{2.13}$$

Optimizing c yields $\mathbb{E}[v^\top(x - \pi) \mid E] \le \sqrt{8 \log(1/\epsilon)/k}$, and so (by Lemma 2.4) p is (σ, ϵ)-resilient around its mean in the ℓ_1-norm, with $\sigma = \frac{\epsilon}{1-\epsilon}\sqrt{8 \log(1/\epsilon)/k}$. The result follows by simple calculation together with Lemma 2.4. □

Our final example will be important to the study of stochastic block models, which are a common model of graph clustering. Let $\mathrm{Ber}(q)$ denote a Bernoulli distribution with parameter q (i.e., $X \sim \mathrm{Ber}(q)$ is 1 with probability q and 0 otherwise). We will see later (Section 3.3) that stochastic block models can be expressed in terms of product distributions where a γ-fraction of the coordinates have elevated mean.

Example 2.9 (Sparse product distributions)**.** Let p^* be a product distribution on $\{0, 1\}^d$, where γd of the coordinates are $\mathrm{Ber}(\frac{a}{d})$ and the remaining $(1 - \gamma)d$ are $\mathrm{Ber}(\frac{b}{d})$ for some $\gamma \le \frac{1}{2}$. Let $\|x\|_{(\gamma)}$ be the sum of the γd largest coordinates of x (in absolute value). Then p^* is resilient in the (γ)-norm with $\sigma^*(\epsilon) \le \mathcal{O}(\epsilon\sqrt{\gamma \max(a, b) \log(\frac{1}{\epsilon})} + \epsilon \log(\frac{1}{\epsilon}))$ for $\epsilon \le \frac{1}{2}$.

Proof. The dual of the (γ)-norm can be shown to be $\|v\|_* = \max(\|v\|_\infty, \frac{1}{\gamma d}\|v\|_1)$. Moreover, the unit ball in this norm is the convex hull of all $\{-1, 0, +1\}$-vectors with γd non-zero coordinates. Applying Lemma 2.6, we need to bound the tail of

$$Z(v) = \sum_{j=1}^{\gamma d} v_j \,\mathrm{Ber}(\frac{a}{d}) + \sum_{j=\gamma d+1}^{d} v_j \,\mathrm{Ber}(\frac{b}{d}) \tag{2.14}$$

for all such vectors v. Note that the variance of $\mathrm{Ber}(\frac{a}{d})$ is at most $\frac{a}{d}$, and similarly for $\mathrm{Ber}(\frac{b}{d})$, so that the sum of the variances (multiplied by v_j^2) is at most $\gamma \max(a, b)$. We recall the moment generating function form of Bernstein's inequality (proved in Section B.3):

Proposition 2.10. *Suppose that X is a random variable with mean μ such that $X \in [0,1]$ and $\mathrm{Var}[X] \le S$. Then $\mathbb{E}[\exp(\lambda(X-\mu))] \le \exp((e^\lambda - \lambda - 1)S)$ for all $\lambda \ge 0$.*

Using Proposition 2.10 to bound each individual term in the sum for $Z(v)$, we then obtain

$$\mathbb{E}[\exp(\lambda(Z(v) - \mathbb{E}[Z(v)]))] \le \exp((e^\lambda - \lambda - 1)\gamma \max(a,b)). \tag{2.15}$$

As in Example 2.8, the ϵ-tail is then bounded by $\frac{1}{\lambda}((e^\lambda - \lambda - 1)\gamma\max(a,b) + \log(1/\epsilon))$. By taking $\lambda = \min(1, \sqrt{\log(1/\epsilon)/\gamma\max(a,b)})$, we obtain a bound of $\mathcal{O}(\sqrt{\gamma \max(a,b)\log(\frac{1}{\epsilon})} + \log(\frac{1}{\epsilon}))$. By Lemma 2.6, we conclude that $\sigma^*(\epsilon) \le \mathcal{O}\Big(\epsilon\sqrt{\gamma\max(a,b)\log(\frac{1}{\epsilon})} + \epsilon\log(\frac{1}{\epsilon})\Big)$ for $\epsilon \le \frac{1}{2}$. □

2.3 Basic Properties and Dual Norm Perspective

Having seen several examples of resilient distributions, we now collect some basic properties of resilience, as well as a dual perspective that is often fruitful.

This dual perspective is already foreshadowed in Lemma 2.6, and is based on the following picture:

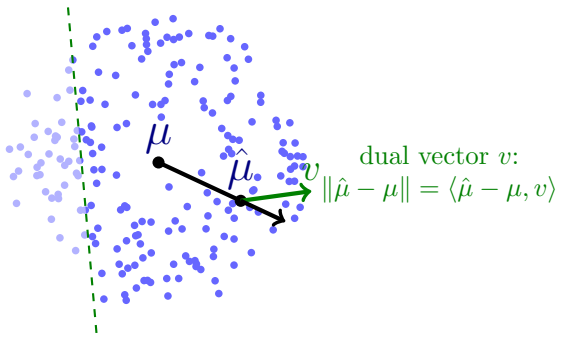

Figure 2.5: The optimal set T discards the smallest $\epsilon|S|$ elements projected onto a dual unit vector v.

Specifically, letting $\hat{\mu} = \mathbb{E}[X \mid E]$, if we have $\|\hat{\mu} - \mu\| = \sigma$, then there must be some dual norm unit vector v such that $\langle \hat{\mu} - \mu, v\rangle = \sigma$ and $\|v\|_* = 1$. Moreover, for such a v, $\langle \hat{\mu} - \mu, v\rangle$ will be largest when E consists of the $(1-\epsilon)$-fraction of points for which $\langle X - \mu, v\rangle$ is largest. Therefore, resilience reduces to a 1-dimensional problem along each of the dual unit vectors v. This is the basic idea behind Lemma 2.6, as well as Example 2.2.

A related result establishes that for $\epsilon = \frac{1}{2}$, resilience in a norm is equivalent to having bounded first moments in the dual norm (see Section B.4 for a proof):

Lemma 2.11. *Suppose that S is $(\sigma, \frac{1}{2})$-resilient in a norm $\|\cdot\|$, and let $\|\cdot\|_*$ be the dual norm. Then S has 1st moments bounded by 2σ: $\frac{1}{|S|}\sum_{i\in S}|\langle x_i - \mu, v\rangle| \le 2\sigma\|v\|_*$ for all $v \in \mathbb{R}^d$.*

Conversely, if S has 1st moments bounded by σ, it is $(2\sigma, \frac{1}{2})$-resilient.

We can also consider resilience around points μ_0 that differ from the mean μ of S, asking that $\|\sum_{i\in T}(x_i - \mu_0)\| \le \sigma$. This is useful especially in cases where the mean of a finite set might differ slightly from the population mean. It turns out that if S is resilient around any point μ_0, it is also resilient around its mean:

Lemma 2.12. *Suppose that S is (σ, ϵ)-resilient around a point μ_0, in the sense that $\left\|\frac{1}{|T|}\sum_{i\in T}(x_i - \mu_0)\right\| \le \sigma$ whenever $|T| \ge (1-\epsilon)|S|$. Let μ be the mean of S. Then S is $(2\sigma, \epsilon)$-resilient around μ.*

Conversely, if S is (σ, ϵ)-resilient around its mean μ, then it is $(\sigma + \|\mu - \mu_0\|, \epsilon)$-resilient around any other point μ_0.

See Section B.5 for a proof. Lemma 2.12 will also be useful in the following two sections where we are given some initial guess $\hat{\mu}$ of the mean and will want to establish an analog of Proposition 2.5.

2.4 Some Initial Algorithms

So far our focus on resilience has been information-theoretic. However, the information-theoretic picture already hints at algorithms: if we can efficiently find resilient sets (at least one of which overlaps S), then we can estimate the mean.

In this section we take this a step further. First, we give a class of norms (*finite* norms) for which resilient sets can be found efficiently assuming a good guess of the mean μ. This is not directly useful, as few norms satisfy the finiteness condition and approximating μ is the entire point of resilience! However, we can use this result to show that given a collection of candidate means (one of which is close to the true mean), we can always narrow down to at most $\mathcal{O}(1/\alpha)$ candidates, even if the norm is not finite. This latter result will see repeated use in the sequel, as many of our later efficient algorithms will output a list of many more than $1/\alpha$ candidates, which must then be narrowed down.

2.4.1 Efficient Algorithms for Finite Norms

We now show that we can find resilient sets efficiently, assuming that we have a good guess of the mean μ. We will assume that the norm is *finite*, in the sense that

$$\|x\| = \max_{j=1}^{N} |\langle x, v_j\rangle|, \tag{2.16}$$

for some finite collection of vectors $v_1, \ldots, v_N$. In this case, whenever a set S exists that is resilient around a known vector μ, we can find a set S' that is almost as large as S and that is resilient with slightly worse parameters:

Theorem 2.13. *Assume that $x_1, \ldots, x_n$ contains a set S of size αn that is $(\sigma, \frac{\alpha}{32\log(4/3\alpha)})$-resilient around a point μ (not necessarily its mean). Then if $\|\cdot\|$ takes the form (2.16), there is an $\mathcal{O}(Nn)$-time algorithm (Algorithm 3) for finding a set S' that is $(2\sigma, \frac{\alpha}{4})$-resilient around μ, with $|S \cap S'| \ge \frac{\alpha}{2}n$.*

Roughly, the degree of resilience of S' is worse by a $\mathcal{O}(\log(2/\alpha))$-factor compared to S. Algorithm 3 for producing S' is given below. The basic idea is to prune away points until the remaining set is resilient. Since the norm is finite, we can do this by checking resilience along each direction v_j; if we find a violation of resilience, then the update in line 9 removes bad points faster than it removes good points.

Algorithm 3 `FindResilientSet`

1: Input $x_1, \dots, x_n$ and μ.
2: Initialize $c_1, \dots, c_n = 1$.
3: Let $Z = \sum_{i=1}^n c_i$.
4: **for** $j = 1, \dots, N$ **do**
5: Let p denote the distribution placing mass $\frac{c_i}{Z}$ on $\langle x_i - \mu, v_j \rangle$.
6: Let σ_+ denote the mean of the $\frac{\alpha}{8}$-fraction of largest values under p.
7: **if** $\sigma_+ \geq 2\sigma$ **then**
8: Let $\tau_i = \max(\langle x_i - \mu, v_j \rangle - \sigma, 0)$.
9: Update $c_i \leftarrow c_i(1 - \tau_i/\tau_{\max})$, where $\tau_{\max} = \max_i \tau_i$.
10: Go back to line 3.
11: **end if**
12: Repeat lines 5-11 with v_j replaced by $-v_j$.
13: **end for**
14: Output $S' = \{i \mid c_i \geq \frac{1}{2}\}$.

The analysis of Algorithm 3 is similar to the filtering algorithm from Chapter 1. First, we will show that $\sum_i c_i \tau_i$ is small across S and large across $\{1, \dots, n\}$.

Lemma 2.14. *The weights τ_i satisfy*

$$\sum_{i \in S} c_i \tau_i \leq \frac{\alpha^2}{32 \log(4/3\alpha)} \sigma \cdot n, \tag{2.17}$$

$$\sum_{i=1}^n c_i \tau_i \geq \frac{\alpha}{8} \sigma \cdot \sum_{i=1}^n c_i. \tag{2.18}$$

See Section B.6 for a proof. The proof of (2.17) exploits resilience of S, while the proof of (2.18) uses the assumption that $\sigma_+ \geq 2\sigma$. As a result of Lemma 2.14, updating the c_i preserves an invariant similarly to Lemma 1.5 from Chapter 1:

Lemma 2.15. *The update $c_i \leftarrow c_i(1 - \tau_i/\tau_{\max})$ on line 9 preserves the invariant*

$$\sum_{i=1}^n c_i \leq n \exp\Big(- \frac{4 \log(4/3\alpha)}{\alpha n} \sum_{i \in S} (1 - c_i) \Big). \tag{$\mathcal{R}$}$$

In particular, $\sum_{i \in S}(1 - c_i) \leq \frac{\alpha}{4} n$ throughout the execution of Algorithm 3.

See Section B.7 for a proof. The invariant ($\mathcal{R}$) is more complicated than the invariant ($\mathcal{I}$) from Lemma 1.5. The reason for this is that while the upper bound (2.17) depends on n, the lower bound

(2.18) depends on $\sum_{i=1}^{n} c_i$, which gets smaller as the algorithm progresses. We therefore need a non-linear function (the exp function) to track the relative change in $\sum_i c_i$ as Algorithm 3 progresses. This is also the reason for the additional $\mathcal{O}(\log(2/\alpha))$-factor.

With Lemmas 2.14 and 2.15 in hand, we can now prove Theorem 2.13.

Proof of Theorem 2.13. Since $(\mathcal{R})$ is preserved (Lemma 2.15), we have that $\sum_{i\in S}(1-c_i) \le \frac{\alpha}{4}n$ at the end of Algorithm 3, and hence $|S \cap S'| \ge \frac{\alpha}{2}n$ (since at most $\frac{\alpha}{2}n$ elements of S can have $c_i < \frac{1}{2}$). In addition, any $\frac{\alpha}{4}$-fraction of elements in S' would have mass at least $\frac{\alpha}{8}$ under p, and p is $(2\sigma, \frac{\alpha}{8})$-resilient by construction (otherwise we would continue to update the c_i). Therefore, S' is $(2\sigma, \frac{\alpha}{4})$-resilient, as was to be shown. □

2.4.2 Corollary: $\mathcal{O}(\frac{1}{\alpha})$ Outputs Suffice

While the finite norm assumption in Theorem 2.13 may seem restrictive, it implies the following result for general norms: If S is resilient around its true mean μ and we are given candidates $\hat{\mu}_1, \dots, \hat{\mu}_m$ such that $\|\hat{\mu}_j - \mu\|$ is small for some j, then we can find a sublist L of at most $\frac{4}{\alpha}$ of the $\hat{\mu}_j$ such that $\|\hat{\mu}_j - \mu\|$ is small for some $j \in L$. More formally:

Corollary 2.16. *Suppose that $x_1, \dots, x_n$ contains a set S of size αn that is $(\sigma, \frac{\alpha}{32\log(4/3\alpha)})$-resilient around its mean μ. Let $\hat{\mu}_1, \dots, \hat{\mu}_m$ be candidate means such that $\min_{j=1}^{m} \|\hat{\mu}_j - \mu\| \le R$. Then there is an efficient procedure that outputs a list $L \subseteq \{1, \dots, m\}$ such that $|L| \le 4/\alpha$ and $\min_{j\in L} \|\hat{\mu}_j - \mu\| \le 5(R+\sigma)$.*

Corollary 2.16 will be important in Chapter 5, as we will obtain recursive algorithms that output a multiplicatively increasing number of candidate means, and need a way to narrow down to a smaller number at each stage to prevent exponential blow-up.

Proof. The basic idea is the following: for each $j \ne j'$, let $v_{jj'}$ be the unit vector in the dual norm such that $\langle \hat{\mu}_j - \hat{\mu}_{j'}, v_{jj'} \rangle = \|\hat{\mu}_j - \hat{\mu}_{j'}\|$. Then define the finite norm

$$\|x\|_u = \max_{j \ne j'} |\langle x, v_{jj'} \rangle|. \tag{2.19}$$

We have $\|x\|_u \le \|x\|$ and hence S is also $(\sigma, \frac{\alpha}{32\log(4/3\alpha)})$-resilient around μ under the norm $\|\cdot\|_u$. Moreover, if $\|\hat{\mu}_{j^*} - \mu\| \le R$ then S is $(R+\sigma, \frac{\alpha}{32\log(4/3\alpha)})$-resilient around $\hat{\mu}_{j^*}$ by Lemma 2.12.

Now, run Algorithm 3 for each of the candidates $\hat{\mu}_j$ using the norm $\|\cdot\|_u$. By Theorem 2.13, for the true candidate $\hat{\mu}_{j^*}$ we obtain a set S'_{j^*} that is $(2(R+\sigma), \frac{\alpha}{4})$-resilient around $\hat{\mu}_{j^*}$ and has size at least $\frac{\alpha}{2}n$. We may also obtain such sets S'_j for other $\hat{\mu}_j$ as well.

We can filter the S'_j to at most $4/\alpha$ elements using a modification of the argument from Proposition 2.5. Consider a maximal subcollection L of the S'_j such that $|S'_j \cap S'_{j'}| \le \frac{\alpha^2}{8}n$ for all $j, j' \in L$. By the pigeonhole principle, there are at most $\frac{4}{\alpha}$ sets in the subcollection. Moreover, S'_{j^*} must intersect

one of these sets S'_j in at least $\frac{\alpha^2}{8}n$ elements by maximality of L. By resilience of both S'_{j^*} and S'_j, we must then have $\|\mu_{j^*} - \mu_j\|_u \leq 4(R+\sigma)$. But $\|\mu_{j^*} - \mu_j\|_u \geq |\langle \mu_{j^*} - \mu_j, v_{j^*j}\rangle| = \|\mu_{j^*} - \mu\|$, so $\|\mu_{j^*} - \mu_j\| \leq 4(R+\sigma)$ as well, and hence $\|\mu - \mu_j\| \leq 4(R+\sigma) + R \leq 5(R+\sigma)$. We can therefore output the $\frac{4}{\alpha}$ elements $\hat{\mu}_j$ for $j \in L$ to obtain the desired result. □

2.5 Bibliographic Remarks

The concept of resilience was first systematically introduced in Steinhardt et al. (2018), with a preliminary version of Proposition 2.5 appearing in Section 8 of Charikar et al. (2017). Sections 2.1 and 2.2 closely follow material from Steinhardt et al. (2018). The results on finite norms (Section 2.4.1) do not to our knowledge appear in the literature, although Proposition B.1 of Diakonikolas et al. (2018b) contains similar ideas to Corollary 2.16, especially the introduction of the $v_{jj'}$ variables to reduce to a finite norm. Kothari and Steinhardt (2018) also uses resilience to prune down a list of candidate means; it employs a more complicated argument than Corollary 2.16, but has the advantage of achieving good results in the regime where $\epsilon \to 0$.

2.6 Exercises

Resilience in Matrix Norms

1. [1+] Given a matrix $X \in \mathbb{R}^{d\times d}$, let $\|X\|_2$ denotes its operator norm: $\|X\|_2 = \max_{\|v\|_2 \leq 1} \|Xv\|_2$. Let p be a distribution over X such that each coordinate X_{ij} is drawn independently from a Gaussian distribution $\mathcal{N}(0,1)$. Show that p is $(\mathcal{O}(\sqrt{\epsilon}), \epsilon)$-resilient in the operator norm.

Sparsity-inducing Norms

Define the norm $\|x\|_{\mathcal{S}_k} = \max\{\langle x, v\rangle \mid \|v\|_2 \leq 1, \|v\|_0 \leq k\}$. Here $\|\cdot\|_0$ denotes the ℓ_0-norm (number of non-zero entries). We call $\|\cdot\|_{\mathcal{S}_k}$ a *sparsity-inducing norm*.

2. [1] Show that if x and y both have at most k non-zero entries, then $\|x - y\|_{\mathcal{S}_{2k}} = \|x - y\|_2$.

3. [1] Let $X \in \mathbb{R}^d$ be Gaussian with independent mean-zero entries of variance 1. Show that the distribution over X is $(\mathcal{O}(\epsilon\sqrt{\log(2/\epsilon)}), \epsilon)$-resilient under $\|\cdot\|_{\mathcal{S}_k}$.

4. [2+] Define a matrix $M \in \mathbb{R}^{2^k \times 2^k}$ such that M_{ij} is the fraction of digits in which the binary representations of $i-1$ and $j-1$ agree with each other. Let $p = \mathcal{N}(0, M)$ be a Gaussian distribution with mean 0 and covariance M. Show that p is $(\mathcal{O}(\sqrt{\epsilon}), \epsilon)$-resilient in $\|\cdot\|_{\mathcal{S}_k}$, but not $(c, 1/2)$-resilient in the ℓ_2-norm for any constant c that is independent of k.

Low-rank recovery

Given a matrix $X \in \mathbb{R}^{d\times n}$, we say that X is *ϵ-rank-resilient* if for all $T \subseteq [n]$ we have $\mathrm{col}(X_T) = \mathrm{col}(X)$ and $\|X_T^\dagger X\|_2 \leq 2$, where X_T is the sub-matrix with columns indexed by T.

5. [2+] Given a matrix $X \in \mathbb{R}^{d \times n}$, suppose that X contains a subset $S \subseteq [n]$ of at least $(1-\epsilon)n$ columns such that X_S is $\frac{\epsilon}{1-\epsilon}$-rank-resilient. Show that it is possible to output a rank-k projection matrix P such that $\|(I-P)X\|_2 \leq 2\sigma_{k+1}(X_S)$, where σ_{k+1} denotes the $k+1$st-largest singular value.

Resilience vs. bounded moments

6. [2] We saw in Example 2.7 that any distribution with bounded kth moments is $(\mathcal{O}(\epsilon^{1-1/k}), \epsilon)$-resilient in the ℓ_2-norm. Show that the converse is not true: for every even $k \geq 2$, there is a distribution that is $(\epsilon^{1-1/k}, \epsilon)$-resilient for all $\epsilon \in (0, 1/3)$, but whose kth moment is infinite.

7. [2] Show that there is a partial converse to Example 2.7: if p is $(\epsilon^{1-1/k}, \epsilon)$-resilient in the ℓ_2-norm for all $\epsilon \in (0, 1/3)$, then p has bounded lth moments for all $l \in [1, k)$.

Chapter 3

Finite-Sample Concentration and Resilient Cores

In this chapter we continue the study of resilience, touching upon more advanced topics. In Sections 3.1 through 3.3, we discuss finite-sample concentration properties of resilience. We first establish a general theorem showing that resilience of a distribution implies resilience of samples from that distribution with high probability (provided the number of samples n is sufficiently large). We apply this to our examples from Chapter 2, and then study the semi-random stochastic block model, showing that resilient sets in this model occur roughly at the *Kesten-Stigum threshold*.

Next, in Section 3.4, we will show that for strongly convex norms, every resilient set has a "core" that has bounded covariance. This exposes a surprising geometric structure in resilient sets, and implies that bounded first moments essentially imply bounded second moments in strongly convex norms, except for some small fraction of outliers.

This chapter is somewhat more technical than the other chapters. Readers can safely skip to the next chapter if they wish.

3.1 Finite-Sample Concentration

We start by presenting a meta-result establishing that resilience of a population distribution p implies resilience of a finite set of samples from that distribution. The number of samples necessary depends on two quantities:

- B, the $\frac{\epsilon}{2}$-quantile of the norm. More precisely, B is such that $\mathbb{P}_{x\sim p^*}[\|x-\mu\| \geq B] \leq \frac{\epsilon}{2}$.
- $\log M$, the log-covering number of the unit ball in the dual norm. More precisely, M is the size of the minimum set of vectors $v_1, \dots, v_M$ such that (i) $\|v_j\|_* \leq 1$ for all j and (ii) $\max_{j=1}^{M} \langle x, v_j \rangle \geq \frac{1}{2}\|x\| = \frac{1}{2}\sup_{\|v\|_* \leq 1} \langle x, v \rangle$ for all vectors $x \in \mathbb{R}^d$.

Both B and $\log M$ are measures of the effective dimension of a space. For instance, if $\|\cdot\|$ is the ℓ_2-norm then B is roughly $\sqrt{d}$ (plus some function of ϵ) and $\log M$ is $\Theta(d)$. For the ℓ_∞-norm, $\log M$ is also $\Theta(d)$, while for the ℓ_1-norm it is $\Theta(\log d)$. These results are established as exercises at the end of the chapter.

Our main result, Theorem 3.1, says that if a distribution p is (σ, ϵ)-resilient, then n i.i.d. samples from p will have a large $(\mathcal{O}(\sigma), \epsilon)$-resilient subset provided $n \gg \max(\frac{1}{\epsilon^2}, \frac{B}{\sigma}) \log M$.

Theorem 3.1. *Suppose that a distribution p is (σ, ϵ)-resilient with $\epsilon < \frac{1}{2}$. Then, given n samples $x_1, \ldots, x_n \sim p$, with probability $1 - \delta - \exp(-\epsilon n/6)$ there is a subset T of $(1-\epsilon)n$ of the x_i such that T is (σ', ϵ)-resilient around the true mean μ with $\sigma' = \mathcal{O}\Big(\sigma \cdot \Big(1 + \sqrt{\frac{\log(M/\delta)}{\epsilon^2 n}} + \frac{(B/\sigma)\log(M/\delta)}{n}\Big)\Big)$.*

Note that Theorem 3.1 only guarantees resilience on a $(1-\epsilon)n$-element subset of the x_i, rather than all of $x_1, \ldots, x_n$. From the perspective of robust estimation, this is sufficient, as we can simply regard the remaining ϵn points as part of the "bad" outlier points. This type of pruning strategy has proved useful in robust estimation as well as other areas such as matrix completion; see the bibliographic remarks at the end of this chapter for further discussion.

The sample complexity in Theorem 3.1 is suboptimal in many cases, requiring roughly $d^{1.5}$ samples when d samples would suffice, due to the $\frac{B}{\sigma} \log M$ term. See Section 6.2 of Steinhardt et al. (2018) for an alternate bound that yields better results in some settings.

Theorem 3.1 yields bounds for each of the examples from Section 2.2; we discuss these next, analyzing how many samples are needed to obtain a large resilient set T as in Theorem 3.1.

Example: Bounded Moments. Suppose as in Example 2.7 that p has bounded kth moments. Then p is $(2\sigma_k \epsilon^{1-1/k}, \epsilon)$-resilient for all $\epsilon \leq \frac{1}{2}$. Additionally,

$$\mathbb{P}[\|x - \mu\|_2 \geq \tau] \leq \mathbb{E}_{x\sim p}[\|x-\mu\|_2^k]/\tau^k \leq C_k^k \mathbb{E}_{x\sim p, v\sim\{\pm 1\}^d}[|\langle x - \mu, v\rangle|^k]/\tau^k, \tag{3.1}$$

where the final inequality is Khinchine's inequality (Haagerup, 1981), which approximates the norm of a vector by its expected inner product with a random sign vector; the constant C_k is $\mathcal{O}(\sqrt{k})$. Fixing v and taking the expectation over x, we have

$$\mathbb{E}_{x\sim p, v\sim\{\pm 1\}^d}[|\langle x-\mu, v\rangle|^k] \leq \sigma_k^k \mathbb{E}_{v\sim\{\pm 1\}^d}[\|v\|_2^k] = \sigma_k^k d^{k/2} \tag{3.2}$$

by the bounded kth moment assumption. Putting these together yields $\mathbb{P}[\|x-\mu\|_2 \geq \tau] \leq \mathcal{O}(\sigma_k \sqrt{kd}/\tau)^k$, from which we see that $B = \mathcal{O}(\sigma_k \sqrt{kd} \epsilon^{-1/k})$.

Next, for the ℓ_2-norm we have $\log M \leq d \log(6)$ by a standard covering argument (see Exercise 2). Therefore, the number of samples needed to achieve $(\mathcal{O}(\sigma_k \epsilon^{1-1/k}), \epsilon)$-resilience is at most $\mathcal{O}(\max(\frac{1}{\epsilon^2}, \frac{B}{\sigma}) \log M) = \mathcal{O}(\frac{d}{\epsilon^2} + \frac{k^{0.5} d^{1.5}}{\epsilon})$.

Example: Item Frequency Estimation. Next consider item frequency estimation (Example 2.8). Here we measure resilience in the ℓ_1-norm, so $\log M$ is the log-covering number in the ℓ_∞-norm, which is at most $m\log(2)$ (since the ℓ_∞-ball is the convex hull of the 2^m sign vectors in $\{\pm 1\}^m$). We also know that $\|x-\mu\|_1 \le 2$ almost surely, so we can take $B=2$ for any value of ϵ. Finally, we know that p is $(6\epsilon\sqrt{\log(1/\epsilon)/k}, \epsilon)$-resilient from Example 2.8. The number of samples needed to achieve $(\mathcal{O}(\epsilon\sqrt{\log(1/\epsilon)/k}), \epsilon)$-resilience is therefore $\mathcal{O}(\max(\frac{1}{\epsilon^2}, \frac{B}{\sigma})\log M) = \mathcal{O}(\frac{m}{\epsilon^2} + \frac{m\sqrt{k}}{\epsilon\sqrt{\log(1/\epsilon)}})$. We believe that the $\sqrt{k}$ factor in the second term is unnecessary, although for $\epsilon \le \sqrt{\log(k)/k}$ this is irrelevant as the first term dominates.

Product distributions (Example 2.9) will be analyzed separately in Section 3.3, where we present a tighter analysis that improves upon Theorem 3.1. We next turn to proving Theorem 3.1.

3.2 Finite-Sample Concentration: Proof of the Main Result

In this section we prove Theorem 3.1. There are two main ideas. The first is that we can reduce to considering 1-dimensional projections by Lemma 2.6, and then union bound over the projections. The second idea is that, for each projection, we can split the contribution of the samples into a "bulk" term that is always small, and a "tail" term that can be bounded via a concentration argument.

Below we first address some preliminaries allowing us to focus our analysis on bounding the 1-dimensional sums (equation 3.3), and then go into the bulk and tail part of the argument.

Preliminaries. Let p' be the distribution of samples from p conditioned on $\|x-\mu\| \le B$. Note that p' is $(\sigma, \frac{\epsilon}{2})$-resilient around μ since every event with probability $1-\epsilon/2$ in p' is an event of probability $(1-\epsilon/2)^2 \ge 1-\epsilon$ in p. Moreover, with probability $1-\exp(-\epsilon n/6)$, at least $(1-\epsilon)n$ of the samples from p will come from p' (by the Chernoff bound). Therefore, we can focus on establishing resilience of the $n' = (1-\epsilon)n$ samples from p'.

With a slight abuse of notation, let $x_1, \dots, x_{n'}$ be the samples from p'. Then to check resilience we need to bound $\|\frac{1}{|T|}\sum_{i\in T}(x_i-\mu)\|$ for all sets T of size at least $(1-\epsilon)n'$. We first use the covering $v_1, \dots, v_M$ to obtain

$$\left\|\frac{1}{|T|}\sum_{i\in T}(x_i-\mu)\right\| \le 2\max_{j=1}^{M}\frac{1}{|T|}\sum_{i\in T}\langle x_i-\mu, v_j\rangle. \tag{3.3}$$

We will analyze the sum over $\langle x_i-\mu, v_j\rangle$ for a fixed v_j and then union bound over the M possibilities.

Splitting into bulk and tail. For a fixed v_j, we will split the sum into two components: those with small magnitude (roughly σ/ϵ) and those with large magnitude (between σ/ϵ and B). We can bound the former "bulk" term directly, and using resilience we will be able to upper-bound the second moment of the latter "tail" term, after which we can use Bernstein's inequality. More formally,

let $\tau = \frac{1-\epsilon}{\epsilon/4}\sigma$ and define

$$y_i = \langle x_i - \mu, v_j\rangle \mathbb{I}[|\langle x_i - \mu, v_j\rangle| < \tau], \tag{3.4}$$

$$z_i = \langle x_i - \mu, v_j\rangle \mathbb{I}[|\langle x_i - \mu, v_j\rangle| \geq \tau]. \tag{3.5}$$

Clearly $y_i + z_i = \langle x_i - \mu, v_j\rangle$. Also, we have $|y_i| \leq \tau$ almost surely, and $|z_i| \leq B$ almost surely (because $x_i \sim p'$ and hence $\langle x_i - \mu, v_j\rangle \leq \|x_i - \mu\| \leq B$).

The threshold τ ensures that z_i is non-zero with probability at most $\epsilon/2$ under p. Indeed, by Lemma 2.6 we have that $\tau_{\epsilon/4}(v_j) \leq \frac{1-\epsilon/4}{\epsilon/4}\sigma \leq \tau$. Therefore, the probability that $\langle x_i - \mu, v_j\rangle \geq \tau$ is at most $\epsilon/4$. Similarly, the probability that $\langle x_i - \mu, v_j\rangle \leq -\tau$ is at most $\epsilon/4$. By the union bound, $\mathbb{P}_p[|\langle x_i - \mu, v_j\rangle| \geq \tau] \leq \epsilon/2$, as claimed. Figure 3.1 summarizes this picture of the y_i and z_i, and previews the remainder of the argument.

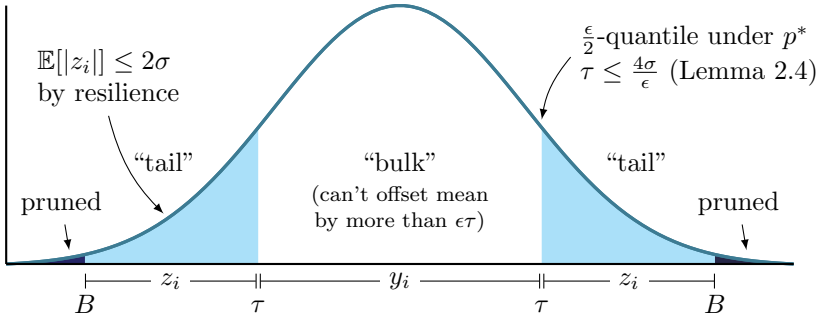

Figure 3.1: Decomposition into bulk (y_i) and tail (z_i). We will show that the bulk cannot change the mean by more than $\epsilon\tau$, while the tail is bounded in expectation by resilience. We will eventually bound the z_i with Bernstein's inequality.

Now, for any set T of size at least $(1-\epsilon)n'$, we have

$$\frac{1}{|T|}\sum_{i\in T}\langle x_i - \mu, v_j\rangle = \frac{1}{|T|}\sum_{i\in T} y_i + z_i \tag{3.6}$$

$$\leq \Big|\frac{1}{|T|}\sum_{i\in T} y_i\Big| + \frac{1}{|T|}\sum_{i\in T}|z_i| \tag{3.7}$$

$$\leq \Big|\frac{1}{|T|}\sum_{i=1}^{n'} y_i\Big| + \Big|\frac{1}{|T|}\sum_{i\notin T} y_i\Big| + \frac{1}{|T|}\sum_{i=1}^{n'}|z_i| \tag{3.8}$$

$$\leq \underbrace{\frac{1}{1-\epsilon}\Big|\frac{1}{n'}\sum_{i=1}^{n'} y_i\Big| + \frac{\epsilon}{1-\epsilon}\tau}_{\text{bulk}} + \underbrace{\frac{1}{(1-\epsilon)n'}\sum_{i=1}^{n'}|z_i|}_{\text{tail}}. \tag{3.9}$$

The last step uses the fact that $|y_i| \leq \tau$ for all i. It thus suffices to bound $|\frac{1}{n'}\sum_{i=1}^{n'} y_i|$ as well as $\frac{1}{n'}\sum_{i=1}^{n'}|z_i|$.

For the y_i term, note that $|\mathbb{E}_{p'}[y_i]| \le \sigma$ by $(\sigma, \epsilon/2)$-resilience of p' (since the event $|\langle x_i - \mu, v_j\rangle| < \tau$ occurs with probability at least $1 - \epsilon/2$ under p'). Moreover, $|y_i| \le \tau$ almost surely. Thus by Hoeffding's inequality, $|\frac{1}{n'}\sum_{i=1}^{n'} y_i| = \mathcal{O}(\sigma + \tau\sqrt{\log(2/\delta)/n'})$ with probability $1 - \delta$.

Bounding the tail. For the z_i term, note that $\mathbb{E}[|z_i|] = \mathbb{E}[\max(z_i, 0)] + \mathbb{E}[\max(-z_i, 0)]$. Let τ' be the ϵ-quantile of $\langle x_i - \mu, v_j\rangle$ under p, which is at most τ (since τ is at least the $(\epsilon/4)$-quantile). Then

$$\mathbb{E}_p[\max(z_i, 0)] = \mathbb{E}_p[\langle x_i - \mu, v_j\rangle \mathbb{I}[\langle x_i - \mu, v_j\rangle \ge \tau]] \tag{3.10}$$

$$\le \mathbb{E}_p[\langle x_i - \mu, v_j\rangle \mathbb{I}[\langle x_i - \mu, v_j\rangle \ge \tau']] \tag{3.11}$$

$$\overset{(i)}{\le} \epsilon \cdot \frac{1-\epsilon}{\epsilon}\sigma = (1-\epsilon)\sigma, \tag{3.12}$$

where (i) is by Lemma 2.6. Then $\mathbb{E}_{p'}[\max(z_i, 0)] \le \frac{1}{1-\epsilon}\mathbb{E}_p[\max(z_i, 0)] \le \sigma$, and hence $\mathbb{E}_{p'}[|z_i|] \le 2\sigma$ (as $\mathbb{E}[\max(-z_i, 0)] \le \sigma$ by the same argument as above).

Since $|z_i| \le B$, we then have $\mathbb{E}[|z_i|^2] \le 2B\sigma$. Therefore, by Bernstein's inequality, with probability $1 - \delta$ we have

$$\frac{1}{n'}\sum_{i=1}^{n'} |z_i| \le \mathcal{O}\Big(\sigma + \sqrt{\frac{\sigma B \log(2/\delta)}{n'}} + \frac{B\log(2/\delta)}{n'}\Big) = \mathcal{O}\Big(\sigma + \frac{B\log(2/\delta)}{n'}\Big). \tag{3.13}$$

Taking a union bound over the v_j for both y and z, and plugging back into (3.9), we get that $\frac{1}{|T|}\sum_{i\in T}\langle x_i - \mu, v_j\rangle \le \mathcal{O}\big(\sigma + \frac{\sigma}{\epsilon}\sqrt{\frac{\log(2M/\delta)}{n}} + \frac{B\log(2M/\delta)}{n}\big)$ for all T and v_j with probability $1 - \delta$. Plugging back into (3.3), we get that $\|\frac{1}{|T|}\sum_{i\in T}(x_i - \mu)\| \le \mathcal{O}\big(\sigma + \frac{\sigma}{\epsilon}\sqrt{\frac{\log(2M/\delta)}{n}} + \frac{B\log(2M/\delta)}{n}\big)$. Thus the points $x_1, \ldots, x_{n'}$ are resilient around μ with the claimed parameters.

3.3 Stochastic Block Model and Kesten-Stigum Threshold

Our final information-theoretic result concerns the semi-random stochastic block model from Charikar et al. (2017), which is a variant of a model proposed in Feige and Kilian (2001). Applying Theorem 3.1 directly yields an overly loose sample complexity bound, so we will show how to directly bound the sample complexity via a similar union bound argument.

In the semi-random stochastic block model, we consider a graph G on n vertices, with an unknown set S of αn "good" vertices. For simplicity we assume the graph is a directed graph. For $i, j \in S$, i is connected to j with probability $\frac{a}{n}$, and for $i \in S$, $j \notin S$, i is connected to j with probability $\frac{b}{n}$, where $b < a$. For $i \notin S$ the edges are allowed to be arbitrary. This is illustrated in Figure 3.2.

If we let $A \in \{0,1\}^{n\times n}$ be the adjacency matrix of G, then the rows in S (i.e., the good rows) are independent samples from a distribution $p = SBM(a, b, \alpha)$ on $\{0,1\}^n$, which is a product of $\mathrm{Ber}(\frac{a}{n})$ and $\mathrm{Ber}(\frac{b}{n})$ distributions as in Example 2.9. The mean μ of this distribution satisfies $\mu_j = \frac{a}{n}$ for

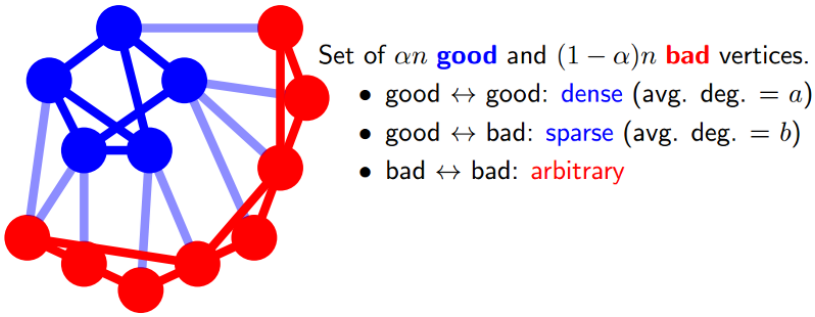

Figure 3.2: Illustration of the semi-random stochastic block model.

$j \in S$, and $\mu_j = \frac{b}{n}$ for $j \notin S$. As shown in Example 2.9, this distribution is resilient in a trimmed ℓ_1-norm. We will now establish finite-sample concentration for this distribution, and show how this can be used to recover S. Let $x_1, \dots, x_{\alpha n}$ denote the αn samples from p corresponding to the elements of S.

Lemma 3.2. *Suppose that $x_1, \dots, x_{\alpha n}$ are drawn from $SBM(a, b, \alpha)$. Take the norm $\|x\|_{(\alpha)} = \max_{|J|=\alpha n} \|x_J\|_1$, which is the maximum ℓ_1-norm over any αn coordinates of x. Then, with probability $1 - \exp(-\Omega(\alpha n))$, the x_i are $(\sigma, \alpha/2)$-resilient under $\|\cdot\|_{(\alpha)}$ with parameter*

$$\sigma = \mathcal{O}\big(\alpha\sqrt{a\log(2/\alpha)} + \log(2/\alpha)\big). \tag{3.14}$$

Proof. Note that we can express $\|x\|_{(\alpha)}$ as $\max_{v\in\mathcal{V}}\langle x, v\rangle$, where $\mathcal{V}$ is the set of αn-sparse $\{0, +1, -1\}$ vectors; in particular, $|\mathcal{V}| = \binom{n}{\alpha n}2^{\alpha n}$. By Lemma 2.4 and the definition of resilience, σ is equal to

$$\frac{\alpha/2}{1-\alpha/2} \max_{T\subseteq\{1,\dots,\alpha n\}, |T|=\frac{1}{2}\alpha^2 n} \max_{v\in\mathcal{V}} \Big\langle \frac{1}{|T|}\sum_{i\in T} x_i - \mu, v\Big\rangle. \tag{3.15}$$

We will union bound over all T and v. For a fixed T and v, the inner expression is equal to $\frac{2}{\alpha^2 n}\sum_{i\in T}\sum_{j=1}^n v_j(x_{ij} - \mu_j)$. Note that all of the $v_j(x_{ij} - \mu_j)$ are independent, zero-mean random variables with variance at most $\frac{a}{n}v_j^2$ and are bounded in $[-1, 1]$. By Bernstein's inequality, we have

$$\mathbb{P}\left[\sum_{i\in T}\sum_{j=1}^n v_j(x_{ij} - \mu_j) \ge t\right] \le \exp\Big(-\frac{1}{2}\frac{t^2}{\sum_{i,j}\frac{a}{n}v_j^2 + \frac{1}{3}t}\Big) \tag{3.16}$$

$$= \exp\Big(-\frac{t^2}{\alpha^3 an + \frac{2}{3}t}\Big), \tag{3.17}$$

Now, if we want our overall union bound to hold with probability $1 - \delta$, we need to set the term in (3.17) to be at most $\delta/\left[\binom{\alpha n}{\alpha^2 n/2}\binom{n}{\alpha n}2^{\alpha n}\right]$, so $\frac{t^2}{\alpha^3 an + \frac{2}{3}t} = \log(1/\delta) + \mathcal{O}\big(\alpha n\log(2/\alpha)\big) = \mathcal{O}\big(\alpha n\log(2\alpha)\big)$ (since we will take $\delta = \exp(-\Theta(\alpha n))$). Hence we can take $t = \mathcal{O}\Big(\sqrt{\alpha^4 an^2\log(2/\alpha)} + \alpha n\log(2/\alpha)\Big)$.

Dividing through by $\frac{1}{2}\alpha^2 n$ and multiplying by $\frac{\alpha/2}{1-\alpha/2}$ to match (3.15), we get

$$\sigma = \mathcal{O}\big(\alpha\sqrt{a\log(2/\alpha)} + \log(2/\alpha)\big), \tag{3.18}$$

as was to be shown. □

As a corollary, we obtain a result on robust recovery of the set S:

Corollary 3.3. *Under the semi-random stochastic block model with parameters α, a, and b, it is information-theoretically possible to obtain sets $\hat{S}_1, \dots, \hat{S}_{2/\alpha}$ satisfying*

$$\frac{1}{\alpha n}|\hat{S}_j \triangle S| = \mathcal{O}\Big(\sqrt{\frac{a\log(2/\alpha)}{\alpha^2(a-b)^2}}\Big). \tag{3.19}$$

for some j with probability $1 - \exp(-\Omega(\alpha n))$.

Proof. By Lemma 3.2 and Proposition 2.5, with probability $1 - \exp(-\Omega(\alpha n))$ we can recover $\hat{\mu}_1, \dots, \hat{\mu}_{2/\alpha}$ such that

$$\|\hat{\mu}_j - \mu\|_{(\alpha)} = \mathcal{O}\Big(\sqrt{a\log(2/\alpha)} + \frac{\log(2/\alpha)}{\alpha}\Big). \tag{3.20}$$

for some j. Note that $\mu_i = \frac{a}{n}$ if $i \in S$ and $\frac{b}{n}$ if $i \notin S$, where $b < a$. We will accordingly define $\hat{S}_j$ to be the set of coordinates i such that $(\hat{\mu}_j)_i \geq \frac{a+b}{2n}$. We then have that $\|\hat{\mu}_j - \mu\|_{(\alpha)} = \Omega(\frac{a-b}{n}\min(|\hat{S}_j \triangle S|, \alpha n))$, and hence $\frac{1}{\alpha n}|\hat{S}_j \triangle S| = \mathcal{O}(\frac{1}{\alpha(a-b)}\|\hat{\mu}_j - \mu\|_{(\alpha)})$ whenever the right-hand-side is at most 1. Using (3.20), we have that

$$\frac{1}{\alpha n}|\hat{S}_j \triangle S| = \mathcal{O}\bigg(\frac{1}{\alpha(a-b)}\Big(\sqrt{a\log(2/\alpha)} + \frac{\log(2/\alpha)}{\alpha}\Big)\bigg) \tag{3.21}$$

$$= \mathcal{O}\bigg(\sqrt{\frac{a\log(2/\alpha)}{\alpha^2(a-b)^2}} + \frac{\log(2/\alpha)}{\alpha^2(a-b)}\bigg) \leq \mathcal{O}\bigg(\sqrt{\frac{a\log(2/\alpha)}{\alpha^2(a-b)^2}} + \frac{a\log(2/\alpha)}{\alpha^2(a-b)^2}\bigg). \tag{3.22}$$

The last inequality multiplies the second term by $\frac{a}{a-b}$, which is at least 1. The first term in (3.22) dominates whenever the bound is meaningful, which yields the desired result. □

Interpretation. We get non-trivial recovery guarantees as long as $\frac{(a-b)^2}{a} \gg \frac{\log(2/\alpha)}{\alpha^2}$. This is close to the famous *Kesten-Stigum threshold* $\frac{(a-b)^2}{a} \gg \frac{1}{\alpha^2}$, which is the conjectured threshold for computationally efficient recovery in the classical stochastic block model (see Decelle et al. (2011) for the conjecture, and Mossel et al. (2013); Massoulié (2014) for a proof in the two-block case). The above upper bound coincides with the Kesten-Stigum threshold up to a $\log(2/\alpha)$ factor. This coincidence is somewhat surprising, and we conjecture that the upper bound is tight up to log factors; some evidence for this is given in Steinhardt (2017), which provides a nearly matching information-theoretic lower bound when $a = 1$, $b = \frac{1}{2}$.

3.4 Resilient Cores

In this section we show that for strongly convex norms, every resilient set contains a large subset with bounded variance. Recall Lemma 2.11, which states that $(\sigma, \frac{1}{2})$-resilience in a norm $\|\cdot\|$ is equivalent to having bounded first moments in the dual norm:

Lemma. Suppose that S is $(\sigma, \frac{1}{2})$-resilient in a norm $\|\cdot\|$, and let $\|\cdot\|_*$ be the dual norm. Then S has 1st moments bounded by 2σ: $\frac{1}{|S|}\sum_{i\in S}|\langle x_i - \mu, v\rangle| \le 2\sigma\|v\|_*$ for all $v \in \mathbb{R}^d$.

Conversely, if S has 1st moments bounded by σ, it is $(2\sigma, \frac{1}{2})$-resilient.

The straightforward proof is given in Section B.4. Thus resilience is closely tied to the 1st moment of p. Suppose now that the norm $\|\cdot\|$ is γ-*strongly convex* with respect to itself, in the sense that

$$\frac{1}{2}(\|x+y\|^2 + \|x-y\|^2) \ge \|x\|^2 + \gamma\|y\|^2. \tag{3.23}$$

In this case, whenever a set S has bounded 1st moments, it has a large "core" with bounded 2nd moments:

Proposition 3.4. *Let S be any set with 1st moments bounded by σ. Then if the norm $\|\cdot\|$ is γ-strongly convex, there exists a core S_0 of size at least $\frac{1}{2}|S|$ with variance bounded by $\frac{32\sigma^2}{\gamma}$. That is, $\frac{1}{|S_0|}\sum_{i\in S_0}|\langle x_i - \mu, v\rangle|^2 \le \frac{32\sigma^2}{\gamma}\|v\|_*^2$ for all $v \in \mathbb{R}^d$.*

While simple to state, the proof of Proposition 3.4 relies on non-trivial facts such as minimax duality and Khintchine's inequality (Khintchine, 1923). Moreover, the assumptions seem necessary: such a core does not exist when $\|\cdot\|$ is the ℓ_p-norm with $p > 2$ (which is a non-strongly-convex norm), or with bounded 3rd moments for $p = 2$ (see Exercise 3).

Proof of Proposition 3.4. Without loss of generality take $\mu = 0$ and suppose that $S = [n]$. We can pose the problem of finding a resilient core as an integer program:

$$\min_{c\in\{0,1\}^n, \|c\|_1 \ge \frac{n}{2}} \max_{\|v\|_* \le 1} \frac{1}{n}\sum_{i=1}^n c_i |\langle x_i, v\rangle|^2. \tag{3.24}$$

Here the variable c_i indicates whether the point i lies in the core S_0. By taking a continuous relaxation and applying a standard duality argument, we obtain the following (see Section C.1 for a proof):

Lemma 3.5. *Suppose that for all m and all vectors $v_1, \dots, v_m$ satisfying $\sum_{j=1}^m \|v_j\|_*^2 \le 1$, we have*

$$\frac{1}{n}\sum_{i=1}^n \sqrt{\sum_{j=1}^m |\langle x_i, v_j\rangle|^2} \le B. \tag{3.25}$$

Then the value of (3.24) is at most $8B^2$.

We can therefore focus on bounding (3.25). Let $s_1, \ldots, s_m \in \{-1, +1\}$ be i.i.d. random sign variables. We have

$$\frac{1}{n}\sum_{i=1}^{n}\sqrt{\sum_{j=1}^{m}|\langle x_i, v_j\rangle|^2} \overset{(i)}{\leq} \mathbb{E}_{s_{1:m}}\left[\frac{\sqrt{2}}{n}\sum_{i=1}^{n}\left|\sum_{j=1}^{m} s_j\langle x_i, v_j\rangle\right|\right] \tag{3.26}$$

$$= \mathbb{E}_{s_{1:m}}\left[\frac{\sqrt{2}}{n}\sum_{i=1}^{n}\left|\left\langle x_i, \sum_{j=1}^{m} s_j v_j\right\rangle\right|\right] \tag{3.27}$$

$$\overset{(ii)}{\leq} \mathbb{E}_{s_{1:m}}\left[\sqrt{2}\sigma\left\|\sum_{j=1}^{m} s_j v_j\right\|_*\right] \tag{3.28}$$

$$\leq \sqrt{2}\sigma\mathbb{E}_{s_{1:m}}\left[\left\|\sum_{j=1}^{m} s_j v_j\right\|_*^2\right]^{\frac{1}{2}}. \tag{3.29}$$

Here (i) is Khintchine's inequality (Haagerup, 1981) and (ii) is the assumed first moment bound. It remains to bound (3.29). The key is the following inequality asserting that the dual norm $\|\cdot\|_*$ is strongly smooth whenever $\|\cdot\|$ is strongly convex (c.f. Lemma 17 of Shalev-Shwartz (2007)):

Lemma 3.6. *If $\|\cdot\|$ is γ-strongly convex, then $\|\cdot\|_*$ is $(1/\gamma)$-strongly smooth: $\frac{1}{2}(\|v+w\|_*^2+\|v-w\|_*^2) \leq \|v\|_*^2 + (1/\gamma)\|w\|_*^2$.*

Applying Lemma 3.6 inductively to $\mathbb{E}_{s_{1:m}}\left[\left\|\sum_{j=1}^{m} s_j v_j\right\|_*^2\right]$, we obtain

$$\mathbb{E}_{s_{1:m}}\left[\left\|\sum_{j=1}^{m} s_j v_j\right\|_*^2\right] \leq \frac{1}{\gamma}\sum_{j=1}^{m}\|v_j\|_*^2 \leq \frac{1}{\gamma}, \tag{3.30}$$

where the final inequality uses the condition $\sum_j \|v_j\|_*^2 \leq 1$ from Lemma 3.5. Combining with (3.29), we have the bound $B \leq \sigma\sqrt{2/\gamma}$, which yields the desired result. □

3.5 Bibliographic Remarks

The results in this chapter mostly follow Steinhardt et al. (2018). The idea of pruning points (as in Theorem 3.1 and Proposition 3.4) to obtain better guarantees was also exploited in Charikar et al. (2017) to yield improved bounds for a graph partitioning problem. Steinhardt et al. (2018) also contains an extension of Proposition 3.4 allowing one to obtain sets of size $(1-\epsilon)|S|$ (rather than $\frac{1}{2}|S|$) under stronger assumptions. Beyond robust estimation, there has been recent interest in showing how to prune samples to achieve faster rates in random matrix settings (Guédon and Vershynin, 2014; Le et al., 2015; Rebrova and Tikhomirov, 2015; Rebrova and Vershynin, 2016). Some of these techniques are also related to the BSS sparsification procedure (Batson et al., 2012), which is a spectral approximation technique in the graph sparsification literature.

In the theory of Banach spaces, strong convexity and smoothness of a norm are referred to as *bounded cotype* and *bounded type* of the norm, respectively (see for instance Chapter 9 of Ledoux and Talagrand (1991)). It would be interesting to explore this connection further.

3.6 Exercises

1. [1+] Show that for the ℓ_1-norm we have $M \leq d$, while for the ℓ_∞-norm we have $M \leq 2^d$.

2. In this exercise we will show that $M \leq 6^d$ for the ℓ_2-norm. In other words, there are unit vectors $v_1, \ldots, v_M$ with $M \leq 6^d$ such that $\max_{j=1}^M \langle x, v_j \rangle \geq \frac{1}{2}\|x\|_2$.

   (a) [2] Let P be the maximum number of unit vectors $x_1, \ldots, x_P$ such that $\|x_i - x_j\|_2 \geq 1/2$ for all $i \neq j$. Show that $P \leq 6^d$. *(Hint: draw a sphere of radius* $1/4$ *around each point* x_i *and compare volumes.)*

   (b) [2] Let Q be the minimum number of unit vectors $y_1, \ldots, y_Q$ such that $\min_{j=1}^Q \|y_i - x\|_2 \leq 1/2$ whenever $\|x\|_2 = 1$. Show that $Q \leq P$.

   (c) [1+] Show that $\{y_1, \ldots, y_Q\}$ constitutes a valid covering. *(Hint: use the fact that* $\langle x, v \rangle = \frac{1}{2}(\|x\|_2^2 + \|v\|_2^2 - \|x - v\|_2^2)$*.)*

3. Let $S = \{e_1, \ldots, e_n\}$ where the e_i are the standard basis in $\mathbb{R}^n$.

   (a) [1+] Show that S is $(2n^{1/p-1}, \frac{1}{2})$-resilient in the ℓ_p-norm for all $p \in [1, \infty]$.

   (b) [1+] Let $\|\cdot\|_q$ be the dual to the ℓ_p-norm. Show that any subset $T \subseteq S$ with $|T| \geq n/2$ has $\max_{\|v\|_q \leq 1} \frac{1}{|S|} \sum_{i \in S} |\langle e_i, v \rangle|^k \geq (n/2)^{\max(-1, k(\frac{1}{p} - 1))}$.

   (c) [1+] Show that this can only be independent of n when $k \leq \frac{p}{p-1}$. What does this say about possible generalizations of Proposition 3.4?

4. [2+] Suppose that, for the stochastic block models in Section 3.3, we use the ℓ_1-norm instead of the trimmed ℓ_1-norm $\|\cdot\|_{(\alpha)}$. Show that concentration does not hold in the ℓ_1-norm—the samples are only $(\sigma, \alpha/2)$-resilient with $\sigma = \Omega(\sqrt{a\alpha})$, and hence applying Proposition 2.5 yields a weaker bound than Corollary 3.3.

Chapter 4

Robust Mean Estimation via Moments and Eigenvectors

We now turn our attention to computationally efficient robust estimation. Recall again the general setting: we are given n data points $x_1, \ldots, x_n$; $(1-\epsilon)n$ of the points are drawn from a distribution p^*, while the remaining are arbitrary outliers. The goal is to estimate the mean μ of p^*.

In Chapter 2 we saw that the property of resilience is information-theoretically sufficient to estimate μ. One consequence is that if a distribution p^* has covariance bounded by σ^2, then it is possible to estimate μ with error $\mathcal{O}(\sigma\sqrt{\epsilon})$. We will now see how to obtain this same result efficiently. The algorithm is a higher-dimensional analog of Algorithm 2, and admits a similar analysis. In Section 4.3, we will extend the algorithm to a general family of stochastic optimization problems.

4.1 ℓ_2 mean estimation via eigenvectors

Recall that in Chapter 1 we established Proposition 1.6, which showed that Algorithm 2 robustly estimates the mean in 1 dimension when the good data have bounded variance. In this section we will establish the multi-dimensional generalization of Proposition 1.6, under the assumption that the data have bounded *covariance*.

Proposition 4.1. *Suppose that $x_1, \ldots, x_n \in \mathbb{R}^d$ contain a subset S of size $(1-\epsilon)n$ that has bounded covariance: $\frac{1}{|S|}\sum_{i\in S}(x_i-\mu)(x_i-\mu)^\top \preceq \sigma^2 I$, where μ is the mean of S. Then if $\epsilon \leq \frac{1}{12}$, there is an efficient algorithm (Algorithm 4) whose output $\hat{\mu}$ satisfies $\|\hat{\mu}-\mu\|_2 = \mathcal{O}(\sigma\sqrt{\epsilon})$.*

While for our information-theoretic results we required resilience, here we require the stronger bounded covariance assumption. Note that bounded covariance implies $(\mathcal{O}(\sigma\sqrt{\epsilon}), \epsilon)$-resilience by Lemma 2.6. Intuitively, bounded covariance gives us an efficiently-checkable sufficient condition for

resilience. In some cases, resilience is also sufficient to imply bounded covariance—see Proposition 3.4 on strongly convex norms.

Algorithm 4 is given below, and is identical to Algorithm 2 except that we project onto the maximum eigenvector v of the covariance when computing $\hat{\sigma}_c$ and τ_i.

Algorithm 4 `FilterL2`

1: Input: $x_1, \dots, x_n \in \mathbb{R}^d$.
2: Initialize weights $c_1, \dots, c_n = 1$.
3: Compute the empirical mean $\hat{\mu}_c$ of the data, $\hat{\mu}_c \stackrel{\text{def}}{=} (\sum_{i=1}^n c_i x_i)/(\sum_{i=1}^n c_i)$.
4: Compute the empirical covariance $\hat{\Sigma}_c \stackrel{\text{def}}{=} \sum_{i=1}^n c_i (x_i - \hat{\mu}_c)(x_i - \hat{\mu}_c)^\top / \sum_{i=1}^n c_i$.
5: Let v be the maximum eigenvector of $\hat{\Sigma}_c$, and let $\hat{\sigma}_c^2 = v^\top \hat{\Sigma}_c v$.
6: If $\hat{\sigma}_c^2 \le 16\sigma^2$, output $\hat{\mu}_c$.
7: Otherwise, let $\tau_i = \langle x_i - \hat{\mu}_c, v\rangle^2$, and update $c_i \leftarrow c_i \cdot (1 - \tau_i/\tau_{\max})$, where $\tau_{\max} = \max_i \tau_i$.
8: Go back to line 3.

Proof of Proposition 4.1. The proof directly mirrors Proposition 1.6. As with Algorithm 2, the intuition is that whenever $\hat{\sigma}_c^2$ is much larger than the variance σ^2 of the good data, the bad points must on average be far away from the empirical mean, which implies that τ_i is large on average. We can formalize this with an analog of Lemma 1.4. Recall the invariant ($\mathcal{I}$) from Lemmas 1.4 and 1.5:

$$\sum_{i \in S} (1 - c_i) \le \frac{1-\epsilon}{2} \sum_{i \notin S} (1 - c_i) \tag{$\mathcal{I}$}$$

This invariant ensures that most of the mass of the c_i remains on the good points S. Assuming that ($\mathcal{I}$) holds, we can show that τ_i is small on S and large overall, and also relate $\|\mu - \hat{\mu}_c\|_2$ to $\hat{\sigma}_c$. This is done in Lemma 4.2, which is a direct analog of the 1-dimensional Lemma 1.4:

Lemma 4.2. *Suppose that the invariant ($\mathcal{I}$) holds. Then* $\|\mu - \hat{\mu}_c\|_2 \le \sigma\sqrt{\frac{\epsilon}{2-\epsilon}} + \hat{\sigma}_c\sqrt{\frac{\epsilon}{1-\epsilon}}$.

Suppose further that $\hat{\sigma}_c^2 \ge 16\sigma^2$ *and* $\epsilon \le \frac{1}{12}$. *Then we have*

$$\sum_{i \in S} c_i \tau_i \le \frac{1-\epsilon}{3} \hat{\sigma}_c^2 n, \text{ while } \sum_{i \notin S} c_i \tau_i \ge \frac{2}{3} \hat{\sigma}_c^2 n. \tag{4.1}$$

Proof. Note that Lemma 1.4 provides the identical statement in 1 dimension. Indeed, Lemma 4.2 follows by applying Lemma 1.4 along each vector u. The variances along u are $u^\top \Sigma u$ and $u^\top \hat{\Sigma}_c u$, which are at most σ^2 and $\hat{\sigma}_c^2$ respectively, so by Lemma 1.4 we have $\langle \mu - \hat{\mu}_c, u\rangle \le \sigma\sqrt{\frac{\epsilon}{2-\epsilon}} + \hat{\sigma}_c\sqrt{\frac{\epsilon}{1-\epsilon}}$. Since this holds for all unit vectors u, we obtain the corresponding bound on $\|\mu - \hat{\mu}_c\|_2$, which yields the first part of Lemma 4.2.

For the second part, note that $\tau_i = \langle x_i - \hat{\mu}_c, v\rangle^2$. This is equivalent to the definition of τ_i in Lemma 1.4, with the vectors x_i replaced by the scalars $\tilde{x}_i = \langle x_i - \hat{\mu}_c, v\rangle$. The variance of the $\tilde{x}_i$

across S is $v^\top \Sigma v \le \sigma^2$, while the variance with respect to the c_i is $v^\top \hat{\Sigma}_c v = \hat{\sigma}_c^2$. The condition of Lemma 1.4 therefore holds and we obtain the desired conclusion (4.1). □

The conclusion (4.1) allows us to separate the good points from the bad points. In particular, by Lemma 1.5, the update $c_i \leftarrow c_i \cdot (1 - \tau_i/\tau_{\max})$ on line 7 preserves the invariant ($\mathcal{I}$), so by induction ($\mathcal{I}$) holds when we output $\hat{\mu}_c$. By Lemma 4.2, we then have $\|\hat{\mu}_c - \mu\|_2 \le \mathcal{O}(\sigma\sqrt{\epsilon} + \hat{\sigma}_c\sqrt{\epsilon}) = \mathcal{O}(\sigma\sqrt{\epsilon})$. This completes the proof of Proposition 4.1. □

Remark 4.3. In Section 1.4, we presented a 1-dimensional algorithm that works in the list-decodable setting when $\alpha \le \frac{1}{2}$ (i.e., $\epsilon \ge \frac{1}{2}$). While that algorithm does not easily generalize to higher dimensions, it is possible to design a version of `FilterL2` that works when $\alpha \le \frac{1}{2}$. At a high level, in addition to downweighting based on the τ_i, one needs to check if the τ_i can be split into distinct subpopulations that are more tightly clustered than the original points. This idea is explored for Gaussian distributions in Diakonikolas et al. (2018b).

4.2 Moment estimation yields robust mean estimation

There was nothing special about the ℓ_2-norm in the previous subsection. Indeed, the proof of Lemma 4.2 followed by applying Lemma 1.4 projected along every unit vector u. If instead of the ℓ_2-norm we wished to estimate μ in some norm $\|\cdot\|$, we would instead want to project along all unit vectors u in the dual norm $\|\cdot\|_*$. The following algorithm generalizes `FilterL2`, and also allows for only approximately solving the corresponding eigenvector problem:

Algorithm 5 `FilterNorm`

1: Initialize weights $c_1, \dots, c_n = 1$.
2: Compute the empirical mean $\hat{\mu}_c$ of the data, $\hat{\mu}_c \stackrel{\text{def}}{=} (\sum_{i=1}^n c_i x_i)/(\sum_{i=1}^n c_i)$.
3: Compute the empirical covariance $\hat{\Sigma}_c \stackrel{\text{def}}{=} \sum_{i=1}^n c_i (x_i - \hat{\mu}_c)(x_i - \hat{\mu}_c)^\top / \sum_{i=1}^n c_i$.
4: Let v be any vector satisfying $\|v\|_* \le 1$ and $v^\top \hat{\Sigma}_c v \ge \frac{1}{\kappa} \max_{\|u\|_* \le 1} u^\top \hat{\Sigma}_c u$.
5: If $v^\top \hat{\Sigma}_c v \le 16\sigma^2$, output $\hat{\mu}_c$.
6: Otherwise, let $\tau_i = \langle x_i - \hat{\mu}_c, v\rangle^2$, and update $c_i \leftarrow c_i \cdot (1 - \tau_i/\tau_{\max})$, where $\tau_{\max} = \max_i \tau_i$.
7: Go back to line 2.

Algorithm 5 enjoys the following bound analogous to Proposition 4.1:

Proposition 4.4. . *Let $\Sigma = \frac{1}{|S|}\sum_{i\in S}(x_i - \mu)(x_i - \mu)^\top$ be the variance of the good data, and suppose that $u^\top \Sigma u \le \sigma^2$ whenever $\|u\|_* \le 1$. Furthermore suppose that $\epsilon \le \frac{1}{12}$. Then Algorithm 5 outputs an estimate $\hat{\mu}_c$ satisfying $\|\hat{\mu}_c - \mu\| \le \mathcal{O}(\sigma\sqrt{\kappa\epsilon})$.*

The proof is essentially identical to Proposition 4.1 and is given as an exercise.

In fact, one can generalize Algorithm 5 even further, although we will not go into detail here. First, rather than approximating the eigenvector problem with a unit vector v, it suffices to find

any *matrix* M that is (1) in the convex hull of $\{uu^\top \mid \|u\|_* \le 1\}$ and (2) satisfies $\langle M, \hat{\Sigma}_c \rangle \ge \frac{1}{\kappa} \max_{\|u\|_* \le 1} u^\top \hat{\Sigma}_c u$. This is important as in many norms the eigenvector problem admits efficient semidefinite approximations that yield such a matrix M. See Li (2017) and Steinhardt et al. (2018) for examples of this idea.

Second, one can apply analogs of Algorithm 5 for higher moments. In this case, $\hat{\Sigma}_c$ is replaced by the moment tensor $T_c = \sum_{i=1}^n c_i (x_i - \hat{\mu}_c)^{\otimes k}$, for some even $k \ge 2$. We can correspondingly let $\tau_i = \langle x_i - \hat{\mu}_c, v \rangle^k$. The main issue is that even approximating the eigenvector problem for higher moments is believed to be hard. We will return to this later in Chapter 5, where we will see some assumptions under which the eigenvector problem for higher moments can be efficiently approximated.

4.3 Generalization to robust stochastic optimization

So far we have focused almost exclusively on mean estimation. While this may seem limiting, in fact many problems in machine learning reduce to mean estimation. Here we show that mean estimation is sufficient for solving *stochastic optimization* problems.

In stochastic optimization (also sometimes called M-estimation), we observe functions f_i drawn from a distribution p^*, where $\mathbb{E}_{f \sim p^*}[f(w)] = \bar{f}(w)$ for some target function $\bar{f}$. For instance, in linear regression we might have $f_i(w) = (y_i - w \cdot x_i)^2$.

To model stochastic optimization with outliers, we assume that we observe n functions $f_1, \dots, f_n : \mathcal{H} \to \mathbb{R}$ (where $\mathcal{H}$ is the input domain). A $(1-\epsilon)$-fraction of the f_i are "good", while the remaining are arbitrary outliers. As before, let S denote the indices of the good functions.

The key idea is that we can use Algorithm 4 to perform robust mean estimation on the *gradients* of f_i, which will allow us to find an approximate stationary point of $\bar{f}$. The main twist is that rather than re-initializing the weights c_i to 1 every time we invoke Algorithm 4, we will want to keep a single set of weights c_i that are updated persistently throughout all runs of the algorithm. This procedure is summarized below:

Algorithm 6 `RobustStochasticOpt`

1: Input: functions $f_1, \dots, f_n$.
2: Initialize weights $c_1, \dots, c_n = 1$.
3: Find any γ-approximate stationary point, i.e. any point $\hat{w}$ such that

$$\Big\| \sum_{i=1}^n c_i \nabla f_i(\hat{w}) / \sum_{i=1}^n c_i \Big\|_2 \le \gamma. \tag{4.2}$$

4: Run `FilterL2` (Algorithm 4) initialized with weights c_i and with points $x_i = \nabla f_i(\hat{w})$.
5: If the c_i are updated by `FilterL2`, go back to line 3.
6: Otherwise, output $\hat{w}$.

When Algorithm 6 terminates, it outputs a point $\hat{w}$ such that $\big\| \frac{1}{|S|} \sum_{i \in S} \nabla f_i(\hat{w}) \big\|_2 \le \gamma + \mathcal{O}(\sigma\sqrt{\epsilon})$,

where σ is a bound on the covariance of the gradients $x_i = \nabla f_i(\hat{w})$. More formally, we have:

Proposition 4.5. *Suppose that the good points f_i have gradients with bounded covariance at all points w:*

$$\frac{1}{|S|}\sum_{i\in S}(\nabla f_i(w) - \nabla \bar{f}(w))(\nabla f_i(w) - \nabla \bar{f}(w))^\top \preceq \sigma^2 I \tag{4.3}$$

for all $w \in \mathcal{H}$, where $\bar{f} = \frac{1}{|S|}\sum_{i\in S} f_i$. Then the output $\hat{w}$ of Algorithm 6 satisfies $\|\nabla \bar{f}(\hat{w})\|_2 \le \gamma + \mathcal{O}(\sigma\sqrt{\epsilon})$.

To interpret the assumption (4.3), suppose first that $f_i(w) = \frac{1}{2}\|w - x_i\|_2^2$, which corresponds to mean estimation. Then $\nabla f_i(w) = w - x_i$, and $\nabla f_i(w) - \nabla \bar{f}(w) = \mu - x_i$. The assumption (4.3) then becomes $\frac{1}{|S|}\sum_{i\in S}(x_i - \mu)(x_i - \mu)^\top \preceq \sigma^2 I$, which is the same as the assumption in Proposition 4.1.

Product distributions. As another example, suppose that x_i is drawn from a product distribution on $\{0,1\}^d$, where the jth coordinate is 1 with probability p_j. Let $f_i(w) = \sum_{j=1}^d x_{ij}\log(w_j) + (1 - x_{ij})\log(1 - w_j)$. In this case $\bar{f}(w) = \sum_{j=1}^d p_j \log(w_j) + (1-p_j)\log(1-w_j)$, and $w_j^* = p_j$, so that $\bar{f}(w) - \bar{f}(w^*)$ is the KL divergence between p and w.

The jth coordinate of $\nabla f_i(w) - \nabla \bar{f}(w)$ is $(x_{ij} - p_j)(1/w_j + 1/(1-w_j))$. In particular, the matrix in (4.3) can be written as $D(w)\Sigma D(w)$, where $\Sigma = \frac{1}{|S|}\sum_{i\in S}(x_i - \mu)(x_i - \mu)^\top$ and $D(w)$ is the diagonal matrix with entries $1/w_j + 1/(1-w_j)$. Suppose that p is *balanced*, meaning that $p_j \in [1/4, 3/4]$, and that we restrict $w_j \in [1/4, 3/4]$ as well. Then $\|D(w)\|_{\text{op}} \le 16/3$, while the matrix Σ has maximum eigenvalue converging to $\max_{j=1}^d p_j(1-p_j) \le \frac{1}{4}$ for large enough n. Thus $\sigma^2 = \mathcal{O}(1)$ in this setting.

Proof of Proposition 4.5. An examination of the proof of Proposition 4.1 shows that it holds whenever the c_i in Algorithm 4 are initialized in a way that satisfies ($\mathcal{I}$). In particular, it holds for the c_i in Algorithm 6 (since the initial value $c_i = 1$ satisfies ($\mathcal{I}$), and each run of Algorithm 4 preserves the invariant).

It follows that when Algorithm 6 terminates, we have $\|\hat{\mu}_c - \mu\|_2 \le \mathcal{O}(\sigma\sqrt{\epsilon})$, where $\hat{\mu}_c = \sum_{i=1}^n c_i \nabla f_i(\hat{w}) / \sum_{i=1}^n c_i$ and $\mu = \frac{1}{|S|}\sum_{i\in S}\nabla f_i(\hat{w}) = \nabla \bar{f}(\hat{w})$. By the triangle inequality, $\|\nabla \bar{f}(\hat{w})\|_2 = \|\mu\|_2 \le \|\hat{\mu}_c\|_2 + \|\mu - \hat{\mu}_c\|_2 \le \gamma + \mathcal{O}(\sigma\sqrt{\epsilon})$, as claimed. □

Proposition 4.5 allows us to extend results on mean estimation to results for a broad family of optimization problems. For instance, if the f_i are convex, then we can make γ arbitrarily small using e.g. gradient descent or another appropriate convex optimization algorithm. We can thus assume that $\|\nabla \bar{f}(\hat{w})\|_2 = \mathcal{O}(\sigma\sqrt{\epsilon})$. We then obtain the following corollary:

Corollary 4.6. *Under the assumptions of Proposition 4.5, suppose that the f_i are convex and that $\gamma = \mathcal{O}(\sigma\sqrt{\epsilon})$ in Algorithm 6. Then if the diameter of $\mathcal{H}$ is at most r, Algorithm 6 outputs a $\hat{w}$ such that $\bar{f}(\hat{w}) - \bar{f}(w^*) = \mathcal{O}(\sigma r\sqrt{\epsilon})$.*

If, in addition, $\bar{f}$ is β-strongly convex, then Algorithm 6 outputs a $\hat{w}$ such that $\bar{f}(\hat{w}) - \bar{f}(w^*) = \mathcal{O}(\sigma^2\epsilon/\beta)$ and $\|\hat{w} - w^*\|_2 = \mathcal{O}(\sigma\sqrt{\epsilon}/\beta)$.

Proof. Let w^* be the global minimizer of $\bar{f}$. We have $\bar{f}(\hat{w}) - \bar{f}(w^*) \leq \langle \nabla \bar{f}(\hat{w}), \hat{w} - w^* \rangle \leq \|\nabla \bar{f}(\hat{w})\|_2 \|\hat{w} - w^*\|_2$, where the first step is by convexity and the second is Cauchy-Schwarz. Since $\|\hat{w} - w^*\|_2 \leq r$ by assumption, we then have that $\bar{f}(\hat{w}) - \bar{f}(w^*) \leq \mathcal{O}(\sigma r \sqrt{\epsilon})$.

Suppose further that $\bar{f}$ is β-strongly convex, meaning that $\bar{f}(w') - \bar{f}(w) \geq \langle \nabla \bar{f}(w), w' - w \rangle + \frac{\beta}{2}\|w' - w\|_2^2$. Applying this at $w' = \hat{w}$, $w = w^*$, we obtain $\frac{\beta}{2}\|\hat{w} - w^*\|_2^2 \leq \bar{f}(\hat{w}) - \bar{f}(w^*) \leq \|\nabla \bar{f}(\hat{w})\|_2 \cdot \|\hat{w} - w^*\|_2$, where the right-hand inequality is from the argument above. We thus obtain $\|\hat{w} - w^*\|_2 \leq \frac{2}{\beta}\|\nabla \bar{f}(\hat{w})\|_2 = \mathcal{O}(\sigma\sqrt{\epsilon}/\beta)$. Plugging back into the bound on $\bar{f}(\hat{w}) - \bar{f}(w^*)$, we obtain $\bar{f}(\hat{w}) - \bar{f}(w^*) \leq \mathcal{O}(\sigma^2 \epsilon / \beta)$. □

As a final remark, in some cases the uniform assumption (4.3) is too crude. For instance, in linear regression the gradients are larger (and have larger variance) when w is far away from the optimum w^*. In such cases, better results may be obtained by replacing σ^2 in (4.3) with a functional form that depends on w, such as $(\sigma_0 + \sigma_1\|w - w^*\|_2)^2$. We do not explore this here, but see Appendix B of Diakonikolas et al. (2018a) for an analysis of this case.

Remark 4.7. While Proposition 4.5 applies to many stochastic optimization problems, the quality of the bound depends on the parameter σ. In particular, the bounded covariance assumption means that we essentially measure the gradients in the ℓ_2-norm (although Proposition 4.4 could be applied to measure the gradients in other norms as well). Moreover, in some cases the second moments are too crude a measure—we may have bounds on much higher moments, in which case we might hope for better dependence on ϵ; or we may not even have bounded second moments. A particular challenge case is robust classification, where points far away from the decision boundary, as well as the existence of multiple classes, leads to large variance in the gradients of the loss but intuitively should not affect the result.

4.4 Bibliographic Remarks

Variants of the filtering idea appear in a number of works, going back to at least Klivans et al. (2009), where an SVD-based filter is used for robust classification. Klivans et al. (2009) uses a "hard" filtering step that loses log factors in the bounds. The filtering idea first appears in a form similar to Algorithm 4 in Diakonikolas et al. (2016); that work applies to Gaussian distributions, but Diakonikolas et al. (2017a) show that the same filtering idea also works under the second moment assumptions considered above. Concurrently with Diakonikolas et al. (2016), Lai et al. (2016) developed a SVD-based filter that operates via a somewhat different divide-and-conquer strategy.

The generalized Algorithm 5 is well-known to experts but we are not aware of it appearing explicitly in published form. Steinhardt et al. (2018) provide a similar generalization of a different duality-based algorithm (see Chapter 5 for more discussion of such algorithms). Li (2017) uses a filtering algorithm in a sparsity-inducing norm to obtain results for robust sparse mean estimation.

Under stonger assumptions, it is possible to improve the $\mathcal{O}(\sqrt{\epsilon})$ dependence in Proposition 4.1. Diakonikolas et al. (2016) show how to do this for Gaussian distributions, while Kothari and Steinhardt (2018) and Hopkins and Li (2018) do this assuming the higher moments of the distribution have bounded sum-of-squares norm.

Generalizations from mean estimation to stochastic optimization appear in Prasad et al. (2018) and Diakonikolas et al. (2018a). An earlier partial generalization (to generalized linear models and some other settings) appears in Du et al. (2017).

4.5 Exercises

1. Prove Proposition 4.4.

Sparsity-inducing norms Suppose that $p^* = \mathcal{N}(\mu, \sigma^2 I)$ is a d-dimensional Gaussian distribution with $\|\mu\|_0 \le k$, i.e. the mean is k-sparse. The next several exercises will show how to robustly estimate μ. Let $\|\cdot\|_{\mathcal{S}_k}$ denote the sparsity-inducing norm from Chapter 2, Exercise 2.

2. Let $\mathcal{M}_k = \{M \in \mathbb{R}^{d\times d} \mid M \succeq 0, \operatorname{tr}(M) \le 1, \sum_{i,j} |M_{ij}| \le k\}$.
   (a) [1+] Show that, for any matrix S, we have $\sup_{M\in\mathcal{M}_k} \langle S, M\rangle \le \sup_{\|v\|_2\le 1, \|v\|_0 \le k} v^\top S v$.
   (b) [1] Show that $\sup_{M\in\mathcal{M}_k} \langle M, \sigma^2 I\rangle = \sigma^2$.

3. [2+] Let $\mathcal{M}'$ be the set of symmetric k^2-sparse $d \times d$ matrices M with $\|M\|_F \le 1$. Show that $\mathcal{M}_k \subseteq 4\operatorname{conv}(\mathcal{M}')$, where conv denotes convex hull.

4. Let $x_1, \dots, x_n \sim \mathcal{N}(0, I)$, and let $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$.
   (a) [2+] Show that for any fixed $M \in \mathcal{M}'$ we have $|\langle \hat{\Sigma} - I, M\rangle| \le \mathcal{O}\left(\sqrt{\frac{\log(2/\delta)}{n}}\right)$ with probability $1-\delta$. *(Hint: Use the Hanson-Wright inequality.)*
   (b) [2] Show that with probability $1-\delta$, we have $|\langle \hat{\Sigma} - I, M\rangle| \le \mathcal{O}\left(\sqrt{\frac{k^2\log(d)+\log(2/\delta)}{n}}\right)$ for all $M \in \mathcal{M}'$. *(Hint: Use the result of Chapter 3, Exercise 2).*

5. [2] Suppose that $x_1, \dots, x_n \sim \mathcal{N}(\mu, \sigma^2 I)$ and let $\hat{\Sigma}$ be the empirical covariance matrix of the x_i. If $n = \Omega(k^2\log(d) + \log(2/\delta))$, Show that $\sup_{M\in\mathcal{M}_k} \langle \hat{\Sigma}, M\rangle \le \mathcal{O}(\sigma^2)$ with probability $1-\delta$.

6. [2] Suppose that $(1-\epsilon)n$ of the x_i are drawn from $\mathcal{N}(\mu, \sigma^2 I)$ and the remaining ϵn points are arbitrary outliers, where $\epsilon \le \frac{1}{12}$ and $n = \Omega(k^2\log(d) + \log(2/\delta))$. Design an algorithm outputting an estimate $\hat{\mu}$ such that $\|\hat{\mu} - \mu\|_2 = \mathcal{O}(\sigma\sqrt{\epsilon})$. *(Hint: Modify Algorithm 5.)*

   Remark 4.8. Li (2017) shows how to obtain a better error of $\mathcal{O}(\sigma\epsilon)$ when $n \ge \Omega(\frac{k^2\log(d)+\log(2/\delta)}{\epsilon^2})$.

Chapter 5

Robust Estimation via Duality

Our final chapter presents an alternate approach to robust estimation based on duality. This approach generalizes more cleanly than the eigenvector-based algorithm from Chapter 5, and allows one to directly handle stochastic optimization problems rather than going through the intermediate step of gradient estimation. While for the most part similar results can be established through the moment approach from Chapter 4 and the duality approach presented here, the duality approach works somewhat better when α is small while the moment approach yields better bounds when $\alpha \to 1$. However, both approaches are under active development and so this picture will likely change in the future.

The core elements of the duality approach already occur in Algorithm 4. As before, the goal will be to find some quantity τ_i such that (1) if $\sum_i c_i \tau_i$ is small, then we can output a good estimate of the parameters we wish to recover; and (2) if $\sum_i c_i \tau_i$ is large, then it must be much larger on the bad data than the good data, thus allowing us to filter via Lemma 1.5.

Rather than constructing τ_i as a secondary step, we will identify a family of optimization problems for which the τ_i fall out naturally, as certain dual potential functions. This family, given in terms of saddle point (min-max) problems, is introduced in Section 5.1; the key property is that for a fixed value of the dual variables, the optimization decomposes additively across data points.

We will first see (Section 5.2) that for non-negative cost functions, it is possible to recover a $\mathcal{O}(1/\alpha)$-approximation to the optimal cost, where α is the fraction of good points. In particular, if the fraction $\epsilon = 1 - \alpha$ of outliers is at most $\frac{1}{2}$, then we recover a constant factor approximation to the optimum. This will allow us to recover the same $\mathcal{O}(\sigma\sqrt{\epsilon})$ bound as in Chapter 4, as well as a bound of $\mathcal{O}(\sigma/\alpha^{1.5})$ that holds even when the fraction α of good points is less than $\frac{1}{2}$. This latter bound is new to this section (but see Remark 4.3 on extending the moment approach to the small-α regime).

Next, we will see how to obtain stronger bounds that hold even when the cost functions are potentially negative, as long as the functions satisfy a certain coupling inequality in terms of non-negative *regularization* functions. This improves the mean estimation bound to $\mathcal{O}(1/\sqrt{\alpha})$

(Section 5.3.2); we can also do even better under a stronger *sum-of-squares* assumption (Section 5.3.3).

5.1 A Family of Saddle Point Problems

In this section we identify the family of optimization problems for which our results will hold. Given a primal domain $\mathcal{H}$ and a dual domain $\mathbf{\Lambda}$, suppose we are given functions $f_1, \dots, f_n : \mathcal{H} \times \mathbf{\Lambda} \to \mathbb{R}$. Our goal will be to solve optimization problems of the following form:

$$\min_{w_i \in \mathcal{H}} \max_{\lambda \in \mathbf{\Lambda}} \sum_{i=1}^{n} f_i(w_i, \lambda). \tag{5.1}$$

Here we abuse notation and use $w_i \in \mathcal{H}$ to mean $w_1, \dots, w_n \in \mathcal{H}$. In Section 5.2 we will treat a subset S of the f_i as good points and the remaining f_i are outliers, and try to minimize the sum only over the good points. For now, however, we examine the basic structure of (5.1).

We will typically assume that each f_i is convex in w_i and concave in λ, in which case (under mild assumptions) Sion's minimax theorem implies that the order of min and max can be switched:

$$\min_{w_i \in \mathcal{H}} \max_{\lambda \in \mathbf{\Lambda}} \sum_{i=1}^{n} f_i(w_i, \lambda) = \max_{\lambda \in \mathbf{\Lambda}} \min_{w_i \in \mathcal{H}} \sum_{i=1}^{n} f_i(w_i, \lambda) = \max_{\lambda \in \mathbf{\Lambda}} \sum_{i=1}^{n} \underbrace{\min_{w_i \in \mathcal{H}} f_i(w_i, \lambda)}_{\stackrel{\text{def}}{=} \tau_i}. \tag{5.2}$$

For a fixed λ, the terms $\min_{w_i \in \mathcal{H}} f_i(w_i, \lambda)$ will play the role of τ_i. We will see this in more detail below. Before that, we give examples of problems taking the form (5.1).

Example 5.1 (Matrix reconstruction)**.** Suppose we are given n data points $x_1, \dots, x_n \in \mathbb{R}^d$, which we stack into a matrix $X \in \mathbb{R}^{d \times n}$. Consider the reconstruction problem

$$\text{minimize } \|X - XW\|_2^2 \text{ subject to } 0 \le W_{ij} \le \frac{1}{\alpha n}, \ \sum_j W_{ij} = 1 \, \forall i. \tag{5.3}$$

This asks to reconstruct X in terms of convex combinations of its columns, such that each weight in the convex combination is at most $\frac{1}{\alpha n}$. Let $\mathcal{H} = \{w \in \mathbb{R}^n \mid 0 \le w_j \le \frac{1}{\alpha n}, \sum_j w_j = 1\}$. Using the fact that $\|Z\|_2^2 = \max\{\operatorname{tr}(Z^\top Y Z) \mid Y \succeq 0, \operatorname{tr}(Y) \le 1\}$, we can re-write (5.3) as

$$\min_{w_i \in \mathcal{H}} \max_{Y \succeq 0, \operatorname{tr}(Y) \le 1} \sum_{i=1}^{n} (x_i - X w_i)^\top Y (x_i - X w_i), \tag{5.4}$$

which has the form of (5.1) with $f_i(w_i, Y) = (x_i - X w_i)^\top Y (x_i - X w_i)$.

Example 5.2 (Low-rank Approximation)**.** Suppose again that we are given data points $x_1, \dots, x_n$ stacked into a matrix $X \in \mathbb{R}^{d \times n}$. We can regularize by the nuclear norm $\|\cdot\|_\star$ to approximate X by

a low-rank matrix W:

$$\min_W \|X - W\|_2^2 + \gamma \|W\|_\star, \tag{5.5}$$

for some constant γ. We can use the same dual form for the operator norm as before, as well as the relation $\|W\|_\star = \max_{\|Z\|_2 \le 1} \operatorname{tr}(Z^\top W)$. This yields the equivalent optimization

$$\min_{w_i \in \mathbb{R}^d} \max_{\substack{Y \succeq 0, \operatorname{tr}(Y) \le 1 \\ \|Z\|_2 \le 1}} \sum_{i=1}^n (x_i - w_i)^\top Y (x_i - w_i) + \gamma \langle w_i, z_i \rangle. \tag{5.6}$$

This has the form (5.1) with $\mathcal{H} = \mathbb{R}^d$, $\mathbf{\Lambda} = \{(Y, Z) \mid Y \succeq 0, \operatorname{tr}(Y) \le 1, \|Z\|_2 \le 1\}$, and $f_i(w_i, Y, Z) = (x_i - w_i)^\top Y (x_i - w_i) + \gamma \langle w_i, z_i \rangle$.

Example 5.3 (Stochastic optimization)**.** Suppose that we are given convex functions $g_i : \mathcal{H} \to \mathbb{R}$ and wish to minimize $\sum_i g_i(w)$. This does not have the form of (5.1) (because we are only minimizing over a single w) but we can approximate the objective by minimizing $\sum_i g_i(w_i) + R_i(w_i; \lambda)$, where R_i is a regularizer encouraging the w_i to have similar values. For instance, similarly to the previous example we can pose the optimization

$$\min_{w_i \in \mathcal{H}} \max_{\|Z\|_2 \le 1} \sum_{i=1}^n g_i(w_i) + \gamma \langle w_i, z_i \rangle. \tag{5.7}$$

This regularizes the nuclear norm of the matrix $W = [w_i]_{i=1}^n$, which encourages W to have low rank. In the extreme case where W has rank 1, the w_i are all scalar multiples of each other. In later sections we will show formally that an analog of (5.7) approximates the objective $\sum_{i=1}^n g_i(w)$.

We will see more examples later in this chapter. We next study the min-max problem (5.1) in the presence of outliers.

5.2 Robustly Approximating Saddle Point Problems

We will now see how to solve the optimization problem (5.1) even in the presence of some number of outlier functions. More formally, suppose that we are given functions $f_1, \dots, f_n : \mathcal{H} \times \mathbf{\Lambda} \to \mathbb{R}$, and that there is an unknown subset $S \subseteq [n]$ of good functions. Our goal is to find $w_1, \dots, w_n \in \mathcal{H}$ so as to minimize

$$\max_{\lambda \in \mathbf{\Lambda}} \sum_{i \in S} f_i(w_i, \lambda). \tag{5.8}$$

If we knew S, then solving (5.8) would be a purely algorithmic problem, but because S is unknown we also need a strategy for dealing with the outliers. Our first main result is that, under certain assumptions, we can obtain an $\mathcal{O}(1/\alpha)$-approximation to (5.8):

Theorem 5.4. *Suppose that each f_i is a continuous non-negative function that is convex in w and concave in λ, and that $\mathcal{H}$ and $\mathbf{\Lambda}$ are convex and compact. For an unknown good set S, let V be the minimum value of (5.8). Then there is a procedure (Algorithm 7) that outputs parameters $\hat{w}_1, \dots, \hat{w}_n$ such that*

$$\max_{\lambda \in \mathbf{\Lambda}} \sum_{i \in S'} f_i(\hat{w}_i, \lambda) \leq \mathcal{O}(V/\alpha) \tag{5.9}$$

for some $S' \subseteq S$ satisfying $|S'| \geq \frac{\alpha(1+\alpha)}{2} n$.

Note that Theorem 5.4 only guarantees that f_i is small across some large subset S' of S, rather than S itself. This is because when we remote outliers we might end up removing some good points along with the bad points. Previously, this was not an issue because resilience ensured that any large subset would have a similar mean to S. Now, we need to ensure that minimizing f_i over a large subset S' is sufficient for a given problem of interest.

When $\alpha = 1 - \epsilon$ the guarantee yields $|S'| \geq (1 - 3\epsilon/2)n$, meaning that f_i is small on almost all of the good points S. For any α, Theorem 5.4 implies that f_i is small on at least half of S.

The compactness assumption on $\mathcal{H}$ and $\mathbf{\Lambda}$ is needed to rule out degenerate cases, but can often be effectively ignored (for instance, by taking the intersection of $\mathcal{H}$ and $\mathbf{\Lambda}$ with a sufficiently large ball). We will mostly ignore this issue in the sequel.

As background, we recall Sion's minimax theorem (presented as a special case for clarity):

Theorem 5.5 (Sion (1958))**.** *Suppose that $F : \mathcal{H} \times \mathbf{\Lambda} \to \mathbb{R}$ is a continuous function that is convex in $\mathcal{H}$ and concave in $\mathbf{\Lambda}$. Also suppose that $\mathcal{H}$ and $\mathbf{\Lambda}$ are convex and compact. Then*

$$\min_{w \in \mathcal{H}} \max_{\lambda \in \mathbf{\Lambda}} F(w, \lambda) = \max_{\lambda \in \mathbf{\Lambda}} \min_{w \in \mathcal{H}} F(w, \lambda). \tag{5.10}$$

Moreover, there is a saddle point $(w^\star, \lambda^\star)$ *such that* $\max_{\lambda \in \mathbf{\Lambda}} F(w^\star, \lambda) = \min_{w \in \mathcal{H}} F(w, \lambda^\star)$.

We will apply Theorem 5.5 to the function $F(w_{1:n}, \lambda) = \sum_{i=1}^n c_i f_i(w_i, \lambda)$, which will allow us to construct appropriate scalars $\tau_i = f_i(w_i^\star, \lambda^\star)$ in Algorithm 7.

Algorithm 7 `DualFilter`

1: Input: functions $f_1, \dots, f_n$.
2: Initialize weights $c_1, \dots, c_n = 1$.
3: Let $(w_{1:n}^\star, \lambda^\star)$ be the solutions to the saddle point problem given by

$$F(w_{1:n}, \lambda) = \sum_{i=1}^n c_i f_i(w_i, \lambda). \tag{5.11}$$

4: Let $\tau_i^\star = f_i(w_i^\star, \lambda^\star)$.
5: If $\sum_{i=1}^n c_i \tau_i^\star \leq 5V/\alpha$, output $w_{1:n}^\star$.
6: Otherwise, update $c_i \leftarrow c_i \cdot (1 - \tau_i^\star / \tau_{\max})$, where $\tau_{\max} = \max_i \tau_i$. Go back to line 3.

The key observation for analyzing Algorithm 7 is that $\sum_{i\in S} c_i\tau_i^\star \le V$. To see this, note that

$$\sum_{i\in S} c_i\tau_i^\star \le \sum_{i\in S}\tau_i^\star \tag{5.12}$$

$$\overset{(i)}{=} \sum_{i\in S}\min_{w_i\in\mathcal{H}} f_i(w_i,\lambda^\star) \tag{5.13}$$

$$\le \max_{\lambda\in\mathbf{\Lambda}}\sum_{i\in S}\min_{w_i\in\mathcal{H}} f_i(w_i,\lambda) \tag{5.14}$$

$$\overset{(ii)}{=} \min_{w_i\in\mathcal{H}}\max_{\lambda\in\mathbf{\Lambda}}\sum_{i\in S} f_i(w_i,\lambda) = V. \tag{5.15}$$

Here the two key steps are (i), which exploits the form (5.1) (in particular, the fact that the w_i are optimized independently for a fixed value of λ); and (ii), which is Sion's minimax theorem.

Since by assumption $\sum_{i=1}^n c_i\tau_i^\star \ge 5V/\alpha$ whenever we update the c_i, we thus have that $\sum_{i\in S} c_i\tau_i^\star \le \frac{\alpha}{5}\sum_{i=1}^n c_i\tau_i^\star$. Re-arranging, we obtain $\sum_{i\in S} c_i\tau_i^\star \le \frac{\alpha}{5-\alpha}\sum_{i\notin S} c_i\tau_i^\star \le \frac{\alpha}{4}\sum_{i\notin S} c_i\tau_i^\star$. This is sufficient for the $\tau_i^\star$ to yield an effective filter similarly to Lemma 1.5. We state this result below:

Lemma 5.6. *Suppose that τ_i is any quantity such that $\sum_{i\in S} c_i\tau_i \le \frac{\alpha}{4}\sum_{i\notin S} c_i\tau_i$, where $\alpha = |S|/n$. Then, the update $c_i \leftarrow c_i(1-\tau_i/\tau_{\max})$ preserves the invariant ($\mathcal{I}$) defined by*

$$\sum_{i\in S}(1-c_i) \le \frac{\alpha}{4}\sum_{i\notin S}(1-c_i). \tag{$\mathcal{I}$}$$

In other words, if ($\mathcal{I}$) holds before the update, it will continue to hold after it.

Note that Lemma 5.6 is just Lemma 1.5 but with different constants. We are now ready to prove Theorem 5.4.

Proof of Theorem 5.4. Since the invariant ($\mathcal{I}$) holds at the beginning of Algorithm 7, by Lemma 5.6 it also holds when we output $w^\star_{1:n}$. Therefore, we have $\sum_{i\in S}(1-c_i) \le \frac{\alpha}{4}\sum_{i\notin S}(1-c_i) \le \frac{\alpha(1-\alpha)}{4}n$. Therefore, $c_i \le \frac{1}{2}$ for at most $\frac{\alpha(1-\alpha)}{2}n$ elements of S. Let S' be the set of at least $\frac{\alpha(1+\alpha)}{2}n$ remaining elements in S for which $c_i \ge \frac{1}{2}$. Then we have

$$\max_{\lambda\in\mathbf{\Lambda}}\sum_{i\in S'} f_i(w_i^\star,\lambda) \overset{(i)}{\le} 2\max_{\lambda\in\mathbf{\Lambda}}\sum_{i\in S'} c_i f_i(w_i^\star,\lambda) \tag{5.16}$$

$$\overset{(ii)}{\le} 2\max_{\lambda\in\mathbf{\Lambda}}\sum_{i=1}^n c_i f_i(w_i^\star,\lambda) \tag{5.17}$$

$$= 2\sum_{i=1}^n c_i f_i(w_i^\star,\lambda^\star) \le 10V/\alpha. \tag{5.18}$$

Here (i) is because $c_i \ge \frac{1}{2}$ for $i\in S'$, while (ii) is by the non-negativity of the f_i. This yields the desired result. □

5.2.1 Applications of Theorem 5.4

We next apply Theorem 5.4 to Examples 5.1 and 5.2. Recall that in both examples we are given a matrix $X = [x_1 \ \cdots \ x_n] \in \mathbb{R}^{d\times n}$ and wish to approximately reconstruct some subset of its columns. To aid us in this, we will assume that for some subset S of the columns, the covariance of the x_i is bounded: $\frac{1}{|S|}\sum_{i\in S}(x_i-\mu)(x_i-\mu)^\top \preceq \sigma^2 I$.

In addition to the matrix reconstruction results, we will obtain obtain results for robust mean estimation as a corollary of analyzing Example 5.2. (We can obtain such results from Example 5.1 as well, but the argument is more complicated so we omit it.)

5.2.2 Matrix Reconstruction (Example 5.1)

Recall that in $X = [x_1 \ \cdots \ x_n] \in \mathbb{R}^{d\times n}$, we wished to minimize $\|X - XW\|_2^2$ and accordingly defined $f_i(w_i, Y) = (x_i - Xw_i)^\top Y(x_i - Xw_i)$. Here $w_i \in \mathcal{H}$ is constrained to satisfy $\sum_j w_{ij} = 1, 0 \le w_{ij} \le \frac{1}{\alpha n}$; the matrix $Y \in \mathbf{\Lambda}$ must satisfy $Y \succeq 0$, $\mathrm{tr}(Y) = 1$. Note that $\max_{Y\in\mathbf{\Lambda}} \sum_{i=1}^n f_i(w_i, Y) = \|X - XW\|_2^2$.

Suppose there is a set S of αn points satisfying $\frac{1}{|S|}\sum_{i\in S}(x_i-\mu)(x_i-\mu)^\top \preceq \sigma^2 I$, where $\mu = \frac{1}{|S|}x_i$ is the mean of S. Then by taking $\tilde{w}_{ij} = \frac{\mathbb{I}[j\in S]}{|S|}$, we obtain $X\tilde{w}_i = \mu$, and hence

$$\max_{\mathrm{tr}(Y)\le 1, Y\succeq 0} \sum_{i\in S} f_i(\tilde{w}_i, Y) = \max_{\mathrm{tr}(Y)\le 1, Y\succeq 0} \sum_{i\in S} (x_i-\mu)^\top Y(x_i-\mu) = \|[x_i-\mu]_{i\in S}\|_2^2 \le \alpha n\sigma^2. \quad (5.19)$$

Therefore, $V \le \alpha n\sigma^2$, and Theorem 5.4 thus yields a set $S' \subseteq S$ and $W^\star = [w_1^\star \ \cdots \ w_n^\star]$ such that $\|[x_i - Xw_i^\star]_{i\in S'}\|_2^2 \le \mathcal{O}(n\sigma^2)$.

5.2.3 Low-Rank Approximation (Example 5.2)

As before, suppose we are given a matrix $X = [x_1 \ \cdots \ x_n] \in \mathbb{R}^{d\times n}$ with a subset S of good columns of covariance at most σ^2. In Example 5.2 we wished to minimize $\|X - W\|_2^2 + \gamma\|W\|_*^2$, where $\|\cdot\|_*$ denotes the nuclear norm. We accordingly define $f_i(w_i, Y, Z) = (x_i - w_i)^\top Y(x_i - w_i) + \gamma\langle w_i, z_i\rangle$ with $\mathbf{\Lambda} = \{(Y, Z) \mid Y \succeq 0, \mathrm{tr}(Y) \le 1, \|Z\|_2 \le 1\}$. Note that for $i \in S$, the f_i depend only on good data points x_i, while for $i \notin S$ the f_i are influenced by outliers. We will show:

Corollary 5.7. *Let $W^\star$ be the output of Algorithm 7, and define $r = \max_{w\in\mathcal{H}} \|w\|_2$. For appropriately chosen functions f_i, there is a set S' with $|S'| \ge \frac{\alpha(1+\alpha)}{2}|S|$ such that*

$$\|X_{S'} - W^\star_{S'}\|_2^2 \le \mathcal{O}(n\sigma^2) \text{ and } \|W^\star_{S'}\|_* \le \mathcal{O}(r\sqrt{n/\alpha}), \quad (5.20)$$

Proof. An obstacle to applying Theorem 5.4 is that the $\langle w_i, z_i\rangle$ term could be negative. However, we can avoid this by instead defining

$$f_i(w_i, Y, Z) = (x_i - w_i)^\top Y(x_i - w_i) + \gamma \max(\langle w_i, z_i\rangle, 0). \quad (5.21)$$

Since we can always set $z_i = 0$, this does not affect the value of the maximum, but ensures that the f_i are all non-negative. Moreover, setting $w_i = \mu$, we obtain

$$\max_{(Y,Z)\in\mathbf{\Lambda}} \sum_{i\in S} f_i(\mu, Y, Z) = \max_{\mathrm{tr}(Y)\le 1, Y\succeq 0, \|Z\|_2\le 1} \sum_{i\in S} (x_i-\mu)^\top Y(x_i-\mu) + \gamma \max(\langle \mu, z_i\rangle, 0) \tag{5.22}$$

$$= \|[x_i-\mu]_{i\in S}\|_2^2 + \gamma \|[\mu]_{i\in S}\|_* \le \alpha n\sigma^2 + \gamma\sqrt{\alpha n}\|\mu\|_2. \tag{5.23}$$

Letting $r = \max_{w\in\mathcal{H}} \|w\|_2$, Theorem 5.4 yields a matrix $W^\star$ and set S' for which $\|X_{S'} - W^\star_{S'}\|_2^2 + \gamma\|W^\star_{S'}\|_* \le \mathcal{O}(n\sigma^2 + \gamma r\sqrt{n/\alpha})$. We now apply a standard trick of setting γ to balance the two terms; in this case, $\gamma = \frac{n\sigma^2}{r\sqrt{n/\alpha}}$. Since both terms are non-negative, we then obtain $\|X_{S'} - W^\star_{S'}\|_2^2 \le \mathcal{O}(n\sigma^2)$ and $\|W^\star_{S'}\|_* \le \mathcal{O}(r\sqrt{n/\alpha})$, as claimed. □

Consequence for mean estimation. We can use the above result to obtain estimates of the mean μ with error $\mathcal{O}(\sigma/\alpha^{1.5})$. A key fact about the nuclear norm (which is why it is a good proxy for matrix rank) is that for any matrix $W^\star$ there is a rank-k matrix $\hat{W}$ such that $\|W^\star - \hat{W}\|_2 \le \|W^\star\|_*/k$ (for instance, we can take $\hat{W}$ to be the top k components of the singular value decomposition). Applying this here, we can obtain a $\hat{W}$ with rank k such that $\|X_{S'} - \hat{W}_{S'}\|_2^2 \le \mathcal{O}(n\sigma^2 + nr^2/(\alpha k^2))$.

Let $U = \mu\mathbb{1}^\top$ be a matrix whose columns are all μ. Since by assumption $\|X_{S'} - U_{S'}\|_2^2 \le \|X_S - U_S\|_2^2 \le \alpha n\sigma^2$, we then have

$$\|U_{S'} - \hat{W}_{S'}\|_F^2 \le \mathrm{rank}(U-\hat{W})\cdot\|U_{S'} - \hat{W}_{S'}\|_2^2 \tag{5.24}$$

$$\le (k+1)\cdot\mathcal{O}(n\sigma^2 + nr^2/(\alpha k^2)) = \mathcal{O}(nk\sigma^2 + nr^2/(\alpha k)). \tag{5.25}$$

By taking $k = \Theta(1/\alpha^2)$, we obtain $\|U_{S'} - \hat{W}_{S'}\|_F^2 \le \mathcal{O}(n\sigma^2/\alpha^2) + \frac{1}{10}\alpha r^2$. This means that on average across S', the squared distance between $\hat{w}_i$ and μ is at most $\mathcal{O}(\sigma^2/\alpha^3) + r^2/5$. This is better than the naïve bound of r^2 by a constant factor, and in fact by iterating this result we can eventually reach a squared distance of $\mathcal{O}(\sigma^2/\alpha^3)$, which gives us an approximation to the mean with ℓ_2-distance $\mathcal{O}(\sigma/\alpha^{1.5})$. We defer the details of this iterative algorithm to Section 5.3.2, where we also obtain a better bound of $\tilde{\mathcal{O}}(\sigma/\alpha^{0.5})$ via a more sophisticated algorithm.

5.3 Better Approximations via Dual Coupling Inequalities

In the previous section, we saw how to obtain a $\mathcal{O}(1/\alpha)$-approximation to the optimum if each of the f_i are non-negative. This result has two weaknesses. First, it only applies to non-negative f_i. Second, if the optimal value V is large compared to typical devitations from the optimum, then a

$\mathcal{O}(1/\alpha)$-approximation may be meaningless. For instance, consider the function

$$\frac{1}{n}\sum_{i=1}^{n}\|x_i - w\|_2^2, \tag{5.26}$$

where the x_i are sampled from a normal distribution $\mathcal{N}(\mu, I)$. Then if we set $w = \mu$ for all i, the value of (5.26) is roughly d. On the other hand, if we set $w = \mu + \Delta$ for some Δ, the value will be roughly $d + \|\Delta\|_2^2$. Thus while the optimum of (5.26) is d, we would need bounds of $d + \mathcal{O}(1)$ to obtain good control over $\|\Delta\|_2$.

To solve this, in this section we present an improvement of Algorithm 7 that can handle f_i taking negative values and that can give tight bounds even when the optimal cost V is large.

Our approach is to couple the potentially negative functions f_i with non-negative *regularization* functions R_i. Specifically, given functions $f_1, \ldots, f_n : \mathcal{H} \to \mathbb{R}$, we seek non-negative functions $R_1, \ldots, R_n : \mathcal{H} \times \mathbf{\Lambda} \to \mathbb{R}$ satisfying the following property:

Definition 5.8 (Dual coupling property)**.** For a set S and functions f_i, the functions R_i are said to possess the *dual coupling property* relative to target parameters $\overline{w}_i$ if

$$\sum_{i\in S} c_i(f_i(\overline{w}_i) - f_i(w_i)) \le \beta\left(\max_{\lambda\in\mathbf{\Lambda}}\sum_{i=1}^{n} c_i R_i(w_i, \lambda)\right)^{1/s} + \gamma \tag{5.27}$$

for all $w_i \in \mathcal{H}$, $\lambda \in \mathbf{\Lambda}$, and $c_i \in [0, 1]$, and some $s > 1$.

In addition to the parameters s, β, and γ, define $\zeta = \max_{\lambda\in\mathbf{\Lambda}} \sum_{i\in S} R_i(\overline{w}_i, \lambda)$. We then say that the R_i possess the dual coupling property with parameters $(s, \beta, \gamma, \zeta)$. Typically we will take $s = 2$, but later corollaries will consider larger values of s as well.

As an example of the self-bounding property, suppose that $f_i(w_i) = \|w_i - x_i\|_2^2$. Then $f_i(\overline{w}_i) - f_i(w_i) = \|\overline{w}_i - x_i\|_2^2 - \|w_i - x_i\|_2^2 = 2\langle w_i - \overline{w}_i, x_i - \overline{w}_i\rangle - \|w_i - \overline{w}_i\|_2^2$. It then suffices to find regularizers R_i that upper bound $2\sum_{i\in S} c_i\langle w_i - \overline{w}_i, x_i - \overline{w}_i\rangle$. A general recipe for doing so is the following: First, upper-bound the sum by a sum of non-negative functions involving the x_i. For instance, using Cauchy-Schwarz we can upper bound the sum in terms of $\sum_{i\in S} c_i\langle w_i - \overline{w}_i, x_i - \overline{w}_i\rangle^2$. Second, replace the x_i term by a variable λ_i, and use a-priori knowledge about the x_i to constrain the space Λ of feasible $\lambda_{1:n}$. In this case, assuming that $\sum_{i\in S}(x_i - \overline{w}_i)(x_i - \overline{w}_i)^\top \preceq \sigma^2 I$, we can replace $\langle w_i - \overline{w}_i, x_i - \overline{w}_i\rangle^2$ with $(w_i - \overline{w}_i)^\top Z_i(w_i - \overline{w}_i)$, where $Z_{1:n} \in \Lambda$ are constrained to satisfy $\sum_i Z_i \preceq \sigma^2 I$. A complication is that we do not know $\overline{w}_i$, but since it is a constant we can fold it into the γ term, and define $R_i(w_i, Z_{1:n}) = w_i^\top Z_i w_i$. We carry out this analysis in detail in Proposition 5.10 below.

Note that the f_i are no longer allowed to depend on the dual variable λ. This is a drawback of the dual coupling approach (though we hold hope that future techniques may circumvent it).

The main result in this section says that we can recover parameters $\hat{w}_i$ for which the $f_i(\hat{w}_i)$ are close in value to $f_i(\overline{w}_i)$, and moreover the R_i are small.

Theorem 5.9. *Let $\mathcal{H}$ and $\mathbf{\Lambda}$ be convex compact sets. Suppose that $f_1, \dots, f_n : \mathcal{H} \to \mathbb{R}$ are convex in $\mathcal{H}$ and that $R_1, \dots, R_n : \mathcal{H} \times \mathbf{\Lambda} \to \mathbb{R}$ are non-negative, convex in $\mathcal{H}$, and concave in $\mathbf{\Lambda}$. For target parameters $\overline{w}_i$, suppose the R_i possess the dual coupling property with parameters $(s, \beta, \gamma, \zeta)$. Then there is an algorithm (Algorithm 8) outputting parameters $\hat{w}_1, \dots, \hat{w}_n$ such that*

$$\sum_{i \in S} c_i (f_i(\hat{w}_i) - f_i(\overline{w}_i)) \le \mathcal{O}(\gamma + \beta \cdot (\zeta/\alpha)^{1/s}) \text{ and } \max_{\lambda \in \mathbf{\Lambda}} \sum_{i \in S} c_i R_i(\hat{w}_i, \lambda) \le \mathcal{O}(\zeta/\alpha), \tag{5.28}$$

where the $c_i \in [0,1]$ satisfy $\sum_{i \in S}(1 - c_i) \le \frac{\alpha(1-\alpha)}{4} n$.

Note that the bound (5.28) on $f_i(\hat{w}_i) - f_i(\overline{w}_i)$ depends only on the parameters $(s, \beta, \gamma, \zeta)$ and not on the magnitude of the f_i. This will allow us to obtain good bounds even when the f_i themselves are large and potentially negative.

Algorithm 8 is given below:

Algorithm 8 `RegularizedDualFilter`

1: Input: functions $f_1, \dots, f_n$, regularizers $R_1, \dots, R_n$.
2: Initialize weights $c_1, \dots, c_n = 1$.
3: Let $(w_{1:n}^\star, \lambda^\star)$ be the solutions to the saddle point problem given by

$$F(w_{1:n}, \lambda) = \sum_{i=1}^n c_i (f_i(w_i) + \kappa R_i(w_i, \lambda)), \text{ where } \kappa = \gamma/\zeta + \beta/(\alpha^{\frac{1}{s}} \zeta^{\frac{s-1}{s}}). \tag{5.29}$$

4: Let $\tau_i^\star = R_i(w_i^\star, \lambda^\star)$.
5: If $\sum_{i=1}^n c_i \tau_i^\star \le \frac{5}{\alpha}\big(\frac{\beta}{\kappa}(\sum_{i=1}^n c_i \tau_i^\star)^{1/s} + \frac{\gamma}{\kappa} + \zeta\big)$, output $w_{1:n}^\star$.
6: Otherwise, update $c_i \leftarrow c_i \cdot (1 - \tau_i^\star / \tau_{\max})$, where $\tau_{\max} = \max_i \tau_i$. Go back to line 3.

Proof of Theorem 5.9. To analyze Algorithm 8, we first need to bound the $\tau_i^\star$. We have

$$\begin{aligned}
\sum_{i \in S} c_i \tau_i^\star &= \sum_{i \in S} c_i R_i(w_i^\star, \lambda^\star) && (5.30)\\
&\overset{(i)}{\le} \sum_{i \in S} c_i \big[\frac{1}{\kappa}(f_i(\overline{w}_i) - f_i(w_i^\star)) + R_i(\overline{w}_i, \lambda^\star)\big] && (5.31)\\
&\overset{(ii)}{\le} \frac{\beta}{\kappa}\left(\sum_{i=1}^n c_i R_i(w_i^\star, \lambda^\star)\right)^{1/s} + \frac{\gamma}{\kappa} + \sum_{i \in S} c_i R_i(\overline{w}_i, \lambda^\star) && (5.32)\\
&\overset{(iii)}{\le} \frac{\beta}{\kappa}\left(\sum_{i=1}^n c_i R_i(w_i^\star, \lambda^\star)\right)^{1/s} + \frac{\gamma}{\kappa} + \zeta, && (5.33)
\end{aligned}$$

where (i) is by the optimality of $w_i^\star$ relative to $\overline{w}_i$ for the functions $f_i + \kappa R_i$, (ii) is by the dual coupling property and the fact that $\lambda^\star$ maximizes $\sum_{i=1}^n c_i R_i(w_i^\star, \lambda)$, and (iii) is by the definition of ζ. We thus have that $\sum_{i \in S} c_i \tau_i^\star \le \frac{\beta}{\kappa}(\sum_{i=1}^n c_i \tau_i^\star)^{1/s} + \frac{\gamma}{\kappa} + \zeta \le \frac{\alpha}{5} \sum_{i=1}^n c_i \tau_i^\star$, and hence we can apply

Lemma 5.6 as before to obtain that $\sum_{i\in S}(1-c_i) \le \frac{\alpha}{4}\sum_{i\notin S}(1-c_i)$ throughout the execution of Algorithm 8.

Now, when Algorithm 8 terminates, we have $z \le \frac{5}{\alpha}\left(\frac{\beta}{\kappa}z^{1/s} + \frac{\gamma}{\kappa} + \zeta\right)$, where $z = \sum_{i=1}^n c_i R_i(w_i^\star, \lambda^\star)$. Inverting the inequality yields $z = \mathcal{O}((\frac{\beta}{\alpha\kappa})^{\frac{s}{s-1}} + \frac{\gamma}{\alpha\kappa} + \frac{\zeta}{\alpha})$. By the non-negativity of the R_i, we have

$$\max_{\lambda\in\mathbf{\Lambda}} \sum_{i\in S} c_i R_i(w_i^\star, \lambda) \le \max_{\lambda\in\mathbf{\Lambda}} \sum_{i=1}^n c_i R_i(w_i^\star, \lambda) = \sum_{i=1}^n c_i R_i(w_i^\star, \lambda^\star) = z \tag{5.34}$$

as well. It remains to bound $\sum_{i\in S} c_i(f_i(w_i^\star) - f_i(\overline{w}_i))$. By the optimality of $w_i^\star$ at $\lambda^\star$, we have

$$\sum_{i\in S} c_i(f_i(w_i^\star) - f_i(\overline{w}_i)) \le \kappa \sum_{i\in S} c_i(R_i(\overline{w}_i, \lambda^\star) - R_i(w_i^\star, \lambda^\star)) \tag{5.35}$$

$$\overset{(i)}{\le} \kappa \sum_{i\in S} R_i(\overline{w}_i, \lambda^\star) \overset{(ii)}{\le} \kappa\zeta, \tag{5.36}$$

where (i) is the non-negativity of R_i and (ii) is the definition of ζ. Plugging in $\kappa = \frac{\gamma}{\zeta} + \frac{\beta}{\alpha}(\frac{\alpha}{\zeta})^{\frac{s-1}{s}}$ to (5.34) and (5.36) yields the desired result. □

5.3.1 Application: Robust Stochastic Optimization

We next discuss applications of Theorem 5.9. Suppose, as in Section 4.3, that the functions f_i satisfy

$$\frac{1}{|S|}\sum_{i\in S}(\nabla f_i(w) - \nabla\bar{f}(w))(\nabla f_i(w) - \nabla\bar{f}(w))^\top \preceq \sigma^2 I \text{ for all } w \in \mathcal{H}. \tag{5.37}$$

We will construct a regularizer that has the dual coupling property for the f_i. Let $\mathbf{\Lambda} = \{(Z_1, \ldots, Z_n) \mid Z_i \succeq 0, \sum_{i=1}^n Z_i \preceq I\}$ and $R_i(w_i, \lambda) = w_i^\top Z_i w_i$. Then the R_i satisfy dual coupling with respect to the f_i:

Proposition 5.10. *Suppose that the f_i are convex and satisfy the bound (5.37). Also let $\overline{w} = \arg\min_{w\in\mathcal{H}} \sum_{i\in S} f_i(w)$ and $r = \max_{w\in\mathcal{H}} \|w\|_2$. Then*

$$\sum_{i\in S} c_i(f_i(\overline{w}) - f_i(w_i)) \le \alpha n\sigma\left(\max_{Z_{1:n}\in\mathbf{\Lambda}} \sum_{i=1}^n c_i w_i^\top Z_i w_i\right)^{1/2} + 3\alpha n\sigma r. \tag{5.38}$$

Moreover, $\sum_{i\in S} \overline{w}^\top Z_i \overline{w} \le r^2$ for all $Z_{1:n} \in \mathbf{\Lambda}$. In particular, the functions $R_i(w_i, Z_{1:n}) = w_i^\top Z_i w_i$ possess the dual coupling property with parameters $s = 2$, $\beta = \alpha n\sigma$, $\gamma = 3\alpha n\sigma r$, and $\zeta = r^2$.

Proof. Let $\tilde{w} = \sum_{i\in S} c_i w_i / \sum_{i\in S} c_i$. Using (5.37) and the optimality of $\overline{w}$, we can show that

$$\sum_{i\in S} c_i(f_i(\overline{w}) - f_i(\tilde{w})) \le \alpha n\sigma\|\overline{w} - \tilde{w}\|_2. \tag{5.39}$$

See Section D.1 for details. Next, we have

$$\sum_{i\in S} c_i(f_i(\tilde{w}) - f_i(w_i)) \overset{(i)}{\leq} \sum_{i\in S} c_i\langle \nabla f_i(\tilde{w}), \tilde{w} - w_i\rangle \tag{5.40}$$

$$\overset{(ii)}{=} \sum_{i\in S} c_i\langle \nabla f_i(\tilde{w}) - \nabla \bar{f}(\tilde{w}), \tilde{w} - w_i\rangle \tag{5.41}$$

$$\overset{(iii)}{\leq} \Big(\sum_{i\in S} c_i\Big)^{1/2} \Big(\underbrace{\sum_{i=1}^{n} c_i(\tilde{w} - w_i)^\top (\nabla f_i(\tilde{w}) - \nabla \bar{f}(\tilde{w}))(\nabla f_i(\tilde{w}) - \nabla \bar{f}(\tilde{w}))^\top (\tilde{w} - w_i)}_{A}\Big)^{1/2}. \tag{5.42}$$

Here (i) is by convexity of the f_i, (ii) uses the fact that $\sum_{i\in S} c_i(\tilde{w} - w_i) = 0$, and (iii) is Cauchy-Schwarz. Now, we know that $\sum_{i\in S}(\nabla f_i(\tilde{w}) - \nabla \bar{f}(\tilde{w}))(\nabla f_i(\tilde{w}) - \nabla \bar{f}(\tilde{w}))^\top \preceq \alpha n\sigma^2 I$ by (5.37). Therefore, the term $A^{1/2}$ in (5.42) is bounded by

$$A^{1/2} \leq (\alpha n\sigma^2)^{1/2} \max_{Z_{1:n}\in\mathbf{\Lambda}} \Big(\sum_{i\in S} c_i(\tilde{w} - w_i)^\top Z_i(\tilde{w} - w_i)\Big)^{1/2} \tag{5.43}$$

$$\leq (\alpha n\sigma^2)^{1/2} \max_{Z_{1:n}\in\mathbf{\Lambda}} \Big(\Big(\sum_{i\in S} c_i w_i^\top Z_i w_i\Big)^{1/2} + \Big(\sum_{i\in S} c_i \tilde{w}^\top Z_i \tilde{w}\Big)^{1/2}\Big) \tag{5.44}$$

$$\leq (\alpha n\sigma^2)^{1/2}\Big(\Big(\max_{Z_{1:n}\in\mathbf{\Lambda}} \sum_{i=1}^{n} c_i w_i^\top Z_i w_i\Big)^{1/2} + \|\tilde{w}\|_2\Big). \tag{5.45}$$

Combining with (5.39) and (5.42) and using $\sum_{i\in S} c_i \leq \alpha n$, , we obtain $\sum_{i\in S} c_i(f_i(\overline{w}) - f_i(w_i)) \leq \alpha n\sigma((\sum_{i=1}^n c_i w_i^\top Z_i w_i)^{1/2} + \|\tilde{w}\|_2 + \|\overline{w} - \tilde{w}\|_2)$. The functions $R_i(w_i, Z_{1:n}) = w_i^\top Z_i w_i$ thus possess the dual coupling property for target parameters $\overline{w}_i = \overline{w}$ with $s = 2$, $\beta = \alpha n\sigma$, $\gamma = \alpha n\sigma(\|\overline{w} - \tilde{w}\|_2 + \|\tilde{w}\|_2) \leq 3\alpha n\sigma r$. Moreover, we have $\zeta = \max_{Z_{1:n}\in\mathbf{\Lambda}} \sum_{i\in S} \overline{w}^\top Z_i \overline{w} \leq \|\overline{w}\|_2^2 \leq r^2$, as claimed. □

Since Proposition 5.10 establishes dual coupling of R_i with f_i, we can apply Theorem 5.9 to obtain:

Corollary 5.11. *If the f_i are convex and satisfy (5.37), and $R_i(w_i, Z_{1:n}) = w_i^\top Z_i w_i$, then Algorithm 8 outputs parameters $\hat{w}_1, \ldots, \hat{w}_n$ satisfying*

$$\sum_{i\in S} c_i(f_i(\hat{w}_i) - f_i(\overline{w}))/\sum_{i\in S} c_i \leq \mathcal{O}(\sigma r/\sqrt{\alpha}), \tag{5.46}$$

$$\max_{Z_{1:n}\in\mathbf{\Lambda}} \sum_{i\in S} c_i \hat{w}_i^\top Z_i \hat{w}_i \leq \mathcal{O}(r^2/\alpha). \tag{5.47}$$

Moreover, if $\tilde{w} = \sum_{i\in S} c_i\hat{w}_i/\sum_{i\in S} c_i$, then $\sum_{i\in S} c_i(f_i(\tilde{w}) - f_i(\overline{w}))/\sum_{i\in S} c_i \leq \mathcal{O}(\sigma r/\sqrt{\alpha})$ as well.

Proof. We apply Theorem 5.9, which states that $\sum_{i\in S} c_i(f_i(\hat{w}_i) - f_i(\overline{w})) \leq \mathcal{O}(\gamma + \beta\cdot(\zeta/\alpha)^{1/s} = 3\alpha n\sigma r + \alpha n\sigma(r^2/\alpha)^{1/2}) = \mathcal{O}(\alpha n\cdot\sigma r/\sqrt{\alpha})$. Moreover, $\sum_{i\in S} c_i \geq 3\alpha n/4$. Combining these yields the

first inequality. The second inequality also follows from Theorem 5.9 upon noting that $\zeta/\alpha = r^2/\alpha$. Finally, applying (5.42) and (5.45) implies that $\sum_{i\in S} c_i(f_i(\tilde{w}) - f_i(\hat{w}_i)) \le \mathcal{O}(\alpha n \cdot \sigma r/\sqrt{\alpha})$, which gives the final inequality. □

5.3.2 Consequence for Mean Estimation

As a particular case of Corollary 5.11, suppose we take $f_i(w_i) = \|w_i - x_i\|_2^2$. Then the minimizer $\overline{w}$ of $\sum_{i\in S} f_i(w)$ is $\mu = \frac{1}{|S|}\sum_{i\in S} x_i$. In addition, $\nabla f_i(w) - \nabla \bar{f}(w) = x_i - \mu$. Therefore, σ^2 is the maximum eigenvalue of the matrix $\frac{2}{|S|}\sum_{i\in S}(x_i-\mu)(x_i-\mu)^\top$. We then obtain that $\frac{1}{C}\sum_{i\in S} c_i(\|\hat{w}_i - x_i\|_2^2 - \|\mu - x_i\|_2^2) = \mathcal{O}(\sigma r/\sqrt{\alpha})$, where $C = \sum_{i\in S} c_i$. But we can also write

$$\frac{1}{C}\sum_{i\in S} c_i\|\hat{w}_i - \mu\|_2^2 = \frac{1}{C}\sum_{i\in S} c_i(\|\hat{w}_i - x_i\|_2^2 - \|\mu - x_i\|_2^2 + 2\langle \hat{w}_i - x_i, \mu - x_i\rangle) \tag{5.48}$$

$$\le \mathcal{O}(\sigma r/\sqrt{\alpha}) + \frac{2}{C}\sum_{i\in S} c_i\langle \hat{w}_i - x_i, \mu - x_i\rangle \tag{5.49}$$

$$\le \mathcal{O}(\sigma r/\sqrt{\alpha}) + 2\sqrt{\frac{1}{C}\sum_{i\in S}(\hat{w}_i - x_i)^\top(\mu - x_i)(\mu - x_i)^\top(\hat{w}_i - x_i)} \tag{5.50}$$

$$= \mathcal{O}(\sigma r/\sqrt{\alpha}). \tag{5.51}$$

This shows that the $\hat{w}_i$ are close to μ for most elements of S, with a typical distance of $\mathcal{O}(\sqrt{\sigma r/\sqrt{\alpha}})$. Unfortunately, this distance depends on the radius r of the space. However, we can remove this dependence by iteratively re-centering around our current estimate $\hat{\mu}$ of μ and re-running Algorithm 8. We split into two cases based on whether α is large or small.

Case 1: $\alpha \ge 0.55$. In this case, let $\hat{\mu} = \sum_{i=1}^n c_i\hat{w}_i / \sum_{i=1}^n c_i$. Then using the bound (5.51), we can show (see Section D.2) that $\|\hat{\mu} - \mu\|_2 \le \mathcal{O}(\sqrt{\sigma r}) + 0.96r$. As long as $r \gg \sigma$, we can re-center around $\hat{\mu}$ and re-run the algorithm in a space of smaller radius $r' = \mathcal{O}(\sqrt{\sigma r}) + 0.96r < r$. Running this until convergence yields:

Proposition 5.12. *If the x_i have covariance at most σ^2 around their mean μ, and $\epsilon \le 0.45$, then there is a procedure that outputs an estimate $\hat{\mu}$ of μ satisfying $\|\hat{\mu} - \mu\|_2 = \mathcal{O}(\sigma)$.*

The $\mathcal{O}(\sigma)$ bound can be further improved to $\mathcal{O}(\sigma\sqrt{\epsilon})$ by searching for a set that has bounded covariance around $\hat{\mu}$. We leave this as an exercise (Exercise 2).

Case 2: general α. By (5.51), there must be some $\hat{w}_i$ such that $\|\hat{w}_i - \mu\|_2 = \mathcal{O}(\sqrt{\sigma r/\sqrt{\alpha}})$. By Corollary 2.16, we can output a list L of at most $\frac{4}{\alpha}$ of the $\hat{w}_i$ such that $\|\hat{w}_i - \mu\|_2 = \mathcal{O}(\sqrt{\sigma r/\sqrt{\alpha}} + \sigma\sqrt{\log(2/\alpha)/\alpha})$ for some $i \in L$.

We can then re-run Algorithm 8 centered around each of the $\hat{w}_i$ (for $i \in L$), with a smaller radius $r' = \mathcal{O}(\sqrt{\sigma r/\sqrt{\alpha}} + \sigma\sqrt{\log(2/\alpha)/\alpha})$. We will now have up to $4n/\alpha$ candidate parameters (n for each of

the $4/\alpha$ runs of the algorithm), such that at least one is within distance $\mathcal{O}(\sqrt{\sigma r'/\sqrt{\alpha}}+\sigma\sqrt{\log(2/\alpha)/\alpha})$ of μ. We can again use Corollary 2.16 to narrow down to a list of at most $\frac{4}{\alpha}$ of the $\hat{w}_i$. Iterating this until convergence, we will eventually end up with at most $\frac{4}{\alpha}$ candidates such that one is within $\mathcal{O}(\sigma\sqrt{\log(2/\alpha)/\alpha})$ of μ. This yields:

Proposition 5.13. *If the x_i have covariance at most σ^2 around their mean μ, then there is a procedure that outputs estimates $\hat{\mu}_1, \ldots, \hat{\mu}_m$ of μ with $m \le \frac{4}{\alpha}$ and $\min_{j=1}^m \|\hat{\mu}_j - \mu\|_2 = \mathcal{O}\Big(\sigma\sqrt{\frac{\log(2/\alpha)}{\alpha}}\Big)$.*

5.3.3 Better Bounds via Sum-of-Squares Relaxations

As a final application of Theorem 5.9, we show how to obtain better estimates of the mean assuming control over higher moments of the x_i. To do this, we will need to make use of a tool called *sum-of-squares relaxations* to obtain efficient algorithms. Roughly, we will show that if the $2t$th moments of the x_i have bounded sum-of-squares norm (which is a relaxation of the spectral norm) then we can estimate the mean with error $\tilde{\mathcal{O}}(\sigma/\alpha^{1/2t})$. We first go over preliminaries around sum-of-squares (SOS) relaxations, then present our assumptions and algorithm.

Sum-of-squares preliminaries. Let $\mathbb{R}[v]$ denote the space of real-valued polynomials in v, and $P_{2t} \subseteq \mathbb{R}[x]$ denote the polynomials of degree at most $2t$. We will call a linear functional $E : P_{2t} \to \mathbb{R}$ a *degree-$2t$ pseudodistribution over the unit sphere* if it satisfies the following properties:

$$E[p(v)^2] \ge 0 \text{ for all polynomials } p(v) \text{ of degree at most } t, \tag{5.52}$$

$$E[1] = 1, \tag{5.53}$$

$$E[(\|v\|_2^2 - 1)p(x)] = 0 \text{ for all polynomials } p(v) \text{ of degree at most } 2t-2. \tag{5.54}$$

The constraints (5.52) and (5.53) ask that E behaves similarly to the expectation under a probablity distribution: the square of any polynomial should have non-negative "expected value" under E, and the constant function should have a value of 1. The constraint (5.54) asks E to act as if its probability mass is supported on the unit sphere (as then $\|x\|_2^2 - 1 = 0$). In particular, any probability distribution over the unit sphere will have an expectation operator E that satisfies (5.52-5.54), but there are potentially many other E that satisfy these constraints as well. We let $\mathcal{D}_{2t}$ denote the set of linear functionals satisfying (5.52) and (5.53), and let $\mathcal{S}_{2t} \subseteq \mathcal{D}_{2t}$ denote the set of linear functionals that further satisfy (5.54). It is a standard result that one can optimize over $\mathcal{S}_{2t}$ and $\mathcal{D}_{2t}$ in polynomial time (see e.g. Barak and Steurer (2016)).

Sum-of-squares norm. Given an order-$2t$ tensor T, we can define a polynomial $p_T(v) = \langle T, v^{\otimes 2t}\rangle$. We define the SOS-norm as $\|T\|_{\text{sos}-2t} = \max\{E[p_T(v)] \mid E \in \mathcal{S}_{2t}\}$; in other words, $\|T\|_{\text{sos}-2t}$ is the maximum "expectation" of $\langle T, v^{\otimes 2t}\rangle$ over all pseudodistributions on the unit sphere. Note that

$\|T\|_{\text{sos}-2t}$ is a relaxation of the injective norm of T, i.e.

$$\|T\|_{\text{sos}-2t} \le \max_{\|v\|_2 \le 1} \langle T, v^{\otimes 2t} \rangle. \tag{5.55}$$

It turns out that for $t = 1$, (5.55) holds with equality, i.e. $\|T\|_{\text{sos}-2} = \lambda_{\max}(T)$ when T is a matrix.

Assumption: bounded SOS-norm. Given data $x_1, \dots, x_n \in \mathbb{R}^d$, we will assume that there is a set S of αn good points with mean μ such that the $2t$-th moment tensor has bounded SOS-norm:

$$\|M_{2t}\|_{\text{sos}-2t} \le \sigma^{2t}, \text{ where } M_{2t} \stackrel{\text{def}}{=} \frac{1}{|S|} \sum_{i \in S} (x_i - \mu)^{\otimes 2t}. \tag{5.56}$$

This is a generalization of the bounded-covariance assumption from before. Since the SOS-norm is larger than the injective norm, this is stronger than assuming that M_{2t} has bounded injective norm (or equivalently, that the x_i have bounded $2t$-th moment). However, there are many distributions p for which samples from p will satisfy (5.56). For instance, Kothari and Steinhardt (2018) show that this holds whenever p satisfies the *Poincaré inequality*, which includes Gaussian distributions, log-concave distributions, and any Lipschitz function of a log-concave distribution.

Algorithm. We start with some notation. Recall from before that given a tensor T, we can define the polynomial $p_T(v) = \langle T, v^{\otimes 2t} \rangle$. However, we can also define the pseudodistribution E_T such that $E_T[p_{T'}] = \langle T, T' \rangle$. This gives us two dual viewpoints on a tensor T: as a polynomial, and as a pseudodistribution. Finally, for a vector u we define the polynomial $p_u(w) \stackrel{\text{def}}{=} p_{u^{\otimes 2t}}(w) = \langle u, w \rangle^{2t}$.

Assuming that (5.56) holds, we will apply Algorithm 8 with $f_i(w_i) = \|w_i - x_i\|_2^2$ and $R_i(w_i, Z_{1:n}) = \langle Z_i, v_i^{\otimes 2t} \rangle$, where now the $Z_i \in \mathbb{R}^{\otimes 2t}$ are order-$2t$ tensors. The space $\mathbf{\Lambda}$ of admissible Z_i is defined as

$$\mathbf{\Lambda} = \Big\{ (Z_1, \dots, Z_n) \mid E_{Z_i} \in \mathcal{D}_{2t}, \ \|\sum_{i=1}^n Z_i\|_{\text{sos}-2t} \le 1 \Big\}. \tag{5.57}$$

In other words, the Z_i must correspond to valid pseudodistributions (not necessarily over the unit sphere), and their sum must have bounded SOS-norm. Note that when $t = 1$, f_i and R_i are the same as in Section 5.3.2. As with $\mathcal{D}_{2t}$, we can optimize over $\mathbf{\Lambda}$ in polynomial time.

Applying Algorithm 8 with the f_i, R_i, and $\mathbf{\Lambda}$ given above, we obtain:

Proposition 5.14. *Suppose that $x_1, \dots, x_n$ contain a set S of size αn with mean μ for which $M_{2t} = \frac{1}{|S|} \sum_{i \in S} (x_i - \mu)^{\otimes 2t}$ satisfies $\|M_{2t}\|_{\text{sos}-2t} \le \sigma^{2t}$. Then there is an efficient algorithm outputting parameters $\hat{\mu}_1, \dots, \hat{\mu}_m$ with $m \le 4/\alpha$ and $\min_{j=1}^m \|\hat{\mu}_j - \mu\|_2 = \mathcal{O}\Big(\sigma \Big(\frac{\log(2\alpha)}{\alpha}\Big)^{1/2t}\Big)$.*

Note that in comparison to Proposition 5.13, the $\Big(\frac{\log(2/\alpha)}{\alpha}\Big)^{1/2}$ factor is replaced by a $\Big(\frac{\log(2/\alpha)}{\alpha}\Big)^{1/2t}$ factor. The bounded SOS-norm assumption thus enabled mean estimation with better dependence

on the fraction α of good points.

Proof of Proposition 5.14. We will apply Theorem 5.9 to the f_i and R_i given above, with $\overline{w}_i = \mu$. A technicality is that we need the R_i to be convex in w, i.e. we need $\langle Z_i, (cv + (1-c)w)^{\otimes 2t}\rangle \leq \langle Z_i, cv^{\otimes 2t} + (1-c)w^{\otimes 2t}\rangle$. An equivalent condition is $E_{Z_i}[cp_v + (1-c)p_w - p_{cv+(1-c)w}] \geq 0$. But the polynomial $cp_v + (1-c)p_w - p_{cv+(1-c)w}$ can be written as a sum of squares of other polynomials (Frenkel and Horváth, 2014), whence the desired inequality follows from $E_{Z_i} \in \mathcal{D}_{2t}$.

We now turn to the dual coupling property. We can bound ζ by noting that for the target parameters $\overline{w}_i = \mu$, we have

$$\sum_{i\in S} R_i(\overline{w}_i, Z_{1:n}) = \sum_{i\in S}\langle Z_i, \mu^{\otimes 2t}\rangle \tag{5.58}$$

$$= \|\mu\|_2^{2t}\langle \sum_{i\in S} Z_i, (\mu/\|\mu\|_2)^{\otimes 2t}\rangle \tag{5.59}$$

$$\overset{(i)}{\leq} \|\mu\|_2^{\otimes 2t} \leq r^{2t}, \tag{5.60}$$

where (i) is because $\|\sum_i Z_i\|_{\text{sos}-2t} \leq 1$ and the operator $p \mapsto p(\mu/\|\mu\|_2)$ is a valid pseudodistribution over the unit sphere. Thus we can take $\zeta = r^{2t}$.

Next, we relate f_i to R_i for the dual coupling property. We have

$$\sum_{i\in S} c_i(f_i(\overline{w}_i) - f_i(w_i)) = \sum_{i\in S} c_i(\|\mu - x_i\|_2^2 - \|w_i - x_i\|_2^2) \tag{5.61}$$

$$\leq 2\sum_{i\in S} c_i\langle \mu - w_i, \mu - x_i\rangle \tag{5.62}$$

$$\overset{(ii)}{\leq} 2\alpha n\sigma\|\mu\|_2 - 2\sum_{i\in S} c_i\langle w_i, \mu - x_i\rangle \tag{5.63}$$

$$\overset{(iii)}{\leq} 2\alpha n\sigma\|\mu\|_2 + 2\Big(\sum_{i\in S} c_i\Big)^{\frac{2t-1}{2t}}\Big(\sum_{i\in S} c_i\langle w_i, \mu - x_i\rangle^{2t}\Big)^{\frac{1}{2t}} \tag{5.64}$$

$$= 2\alpha n\sigma\|\mu\|_2 + 2\Big(\sum_{i\in S} c_i\Big)^{\frac{2t-1}{2t}}\Big(\alpha n\sigma^{2t}\sum_{i\in S} c_i\langle \hat{Z}_i, w_i^{\otimes 2t}\rangle\Big)^{\frac{1}{2t}} \tag{5.65}$$

$$\leq 2\alpha n\sigma\|\mu\|_2 + 2\alpha n\sigma\Big(\sum_{i\in S} c_i R_i(w_i, \hat{Z}_{1:n})\Big)^{\frac{1}{2t}}, \tag{5.66}$$

where $\hat{Z}_i = (\mu - x_i)^{\otimes 2t}/\alpha n\sigma^{2t}$. Here (ii) is because bounded SOS-norm implies bounded covariance (which in turn bounds $\|\sum_i c_i(\mu - x_i)\|_2$), and (iii) is Hölder's inequality.

We need to ensure that the $\hat{Z}_{1:n} \in \mathbf{\Lambda}$. We have $\sum_{i\in S}\hat{Z}_i = \frac{1}{\alpha n\sigma^{2t}}\sum_{i\in S}(x_i - \mu)^{\otimes 2t} = M_{2t}/\sigma^{2t}$, and $\|M_{2t}/\sigma^{2t}\|_{\text{sos}-2t} = \|M_{2t}\|_{\text{sos}-2t}/\sigma^{2t} \leq 1$. Therefore, we indeed have $\hat{Z}_{1:n} \in \mathbf{\Lambda}$.

The dual coupling property thus holds with $s = 2t$, $\beta = 2\alpha n\sigma$, $\gamma = 2\alpha n\sigma r$, and $\zeta = r^{2t}$. Applying

Theorem 5.9, the output of Algorithm 8 satisfies

$$\sum_{i \in S} c_i(\|\hat{w}_i - x_i\|_2^2 - \|\mu - x_i\|_2^2) \leq \mathcal{O}(\alpha n \sigma r / \alpha^{1/2t}). \tag{5.67}$$

As before, we can use this to conclude that

$$\frac{1}{C}\sum_{i \in S} c_i \|\hat{w}_i - \mu\|_2^2 \leq \mathcal{O}(r\sigma/\alpha^{1/2t}), \text{ where } C = \sum_{i \in S} c_i. \tag{5.68}$$

Using the same recursive procedure as in Section 5.3.2, we then eventually end up with parameters $\hat{\mu}_1, \ldots, \hat{\mu}_m$ with $m \leq 4/\alpha$ and $\min_{j=1}^m \|\hat{\mu}_j - \mu\|_2 = \mathcal{O}\Big(\sigma\Big(\frac{\log(2/\alpha)}{\alpha}\Big)^{1/2t}\Big)$. □

5.4 Bibliographic Remarks

The duality-based approach to robust estimation was first introduced in Charikar et al. (2017). That work used the dual coupling approach specialized to the setting in Section 5.3.1, although it employed a more complicated algorithm and argument than the one given here. The saddle point formulation from Section 5.2 actually appeared later, in in Steinhardt et al. (2018). That work uses a matrix reconstruction objective along the lines of Example 5.1 to obtain bounds for mean estimation. The saddle point formulation was also used in Kothari and Steinhardt (2018) to obtain mean estimation results using sum-of-squares algorithms. The sum-of-squares results presented here are in general stronger than those in Kothari and Steinhardt (2018), as they make use of the stronger dual coupling bounds given by Theorem 5.9.

Multiple previous papers (Charikar et al., 2017; Kothari and Steinhardt, 2018) make use of an iterative re-clustering algorithm to obtain stronger bounds. This general re-clustering approach has been streamlined here through the use of Algorithm 3 for finite norms and the resulting Corollary 2.16. Previous re-clustering techniques sometimes used substantially more complex analysis, e.g. the approach in Charikar et al. (2017) makes use of metric embeddings. On the other hand, Corollary 2.16 does not provide an immediate way to obtain better results as the fraction ϵ of outliers approaches 0, which is a disadvantage relative to the approach in Kothari and Steinhardt (2018). Similarly to Corollary 2.16, the approach in Kothari and Steinhardt (2018) also works by finding resilient sets.

5.5 Exercises

1. Suppose that we are given a matrix $X \in \mathbb{R}^{d \times n}$ such that $\|X_S - \bar{W}_S\|_2^2 \leq \mathcal{O}(|S|\sigma^2)$ for some subset $S \subseteq [n]$ of size αn. Here $\bar{W}_S$ is an unknown matrix with the guarantee that the nuclear norm $\|\bar{W}_S\|_*$ is at most ρ.

(a) [2] The matrix Hölder's inequality states that $\langle A, B\rangle \le \|A\|_2 \|B\|_*$. Use this to obtain a dual coupling inequality with exponent $s = 1$, for the function $f_i(w_i) = \|x_i - w_i\|_2^2$. *(Hint: use the regularizer* $R_i(w_i, Z) = \max(\langle w_i, z_i\rangle, 0)$ *with the constraint* $\|Z\|_2 \le 1$*.)*

(b) [1+] Theorem 5.9 requires $s > 1$. Show, nevertheless, that an analog of Theorem 5.9 holds even when $s = 1$, provided that we set the regularization constant κ to be large enough.

(c) [1+] Use (a) and (b) to obtain an algorithm that recovers a $\hat{W}$ for which $\|\hat{w}_i - \overline{w}_i\|_2^2$ is small on average on a large subset S' of S.

(d) [2] Using the regularizer $R_i(w_i, Z_{1:n}) = w_i^\top Z_i w_i$ from Section 5.3.1, can you obtain a dual coupling inequality with $s = 2$?

2. Suppose that we are in the typical mean estimation scenario where we are given $x_1, \dots, x_n \in \mathbb{R}^d$, and an unknown subset of $(1-\epsilon)n$ of the points have covariance at most σ^2 around their mean μ. Suppose further that we are given a list of candidate means $\hat{\mu}_1, \dots, \hat{\mu}_m$ such that $\|\hat{\mu}_j - \mu\|_2 \le \sigma$ for some j. In this exercise we will show that we can then obtain an estimate $\hat{\mu}$ such that $\|\hat{\mu} - \mu\|_2 \le \mathcal{O}(\sigma\sqrt{\epsilon})$.

(a) [1+] For any vector v, consider the optimization problem

$$\min_{0 \le c_i \le 1, \sum_i c_i \ge (1-\epsilon)n} F(c; v), \text{ where } F(c; v) = \lambda_{\max}\Big(\sum_{i=1}^n c_i (x_i - v)(x_i - v)^\top\Big). \tag{5.69}$$

Show that $F(c; v)$ is convex in c.

(b) [2−] Suppose that for any v, $F(c; v) \le \mathcal{O}(\sigma^2)$. Let $S = \{i \mid c_i \ge \frac{1}{2}\}$. Show that S is $(\mathcal{O}(\sigma\sqrt{\epsilon}), \epsilon)$-resilient and that $|S| \ge (1 - 2\epsilon)n$.

(c) [1+] Using the results of (a) and (b), show that given the list $\hat{\mu}_1, \dots, \hat{\mu}_m$, we can obtain a single estimate $\hat{\mu}$ with $\|\hat{\mu} - \mu\|_2 \le \mathcal{O}(\sigma\sqrt{\epsilon})$.

3. Consider a distribution learning problem: we are given data points $x_1, \dots, x_n \in \{1, \dots, m\}$, where an unknown subset S of αn of the points have empirical probability distribution π. Our goal is to recover a distribution w that approximates π. We represent w and π as functions from $\{1, \dots, m\}$ to $[0, 1]$ that sum to 1.

(a) [2+] Consider the objective function $f_i(w_i) = -\log w(x_i)$. Show that $\sum_{i \in S} c_i (f_i(\pi) - f_i(w_i)) \le \alpha n[\log(\sum_{j=1}^m \max_{i=1}^n c_i w_i(j)) + 1/e]$.

(b) [2] Derive an analog of Theorem 5.9 showing that we can recover a probability distribution $\hat{w}$ such that the KL divergence from π to $\hat{w}$ is $\mathcal{O}(\log(2/\alpha))$.

(c) [1+] How meaningful is this reuslt?

Chapter 6

Discussion

We have now seen how to perform robust estimation in high dimensions, both information-theoretically (Chapters 2 and 3) and computationally (Chapters 4 and 5). From the information-theoretic perspective, a key takeaway is to look at populations rather than individual points, which allows us to avoid the typical $\sqrt{d}$ error of traditional estimators. This leads to the concept of *resilience*, which gave tight characterizations of information-theoretic recoverability in many cases. Algorithmically, we saw two techniques, based on moment estimation and duality, that enabled robust mean estimation, robust stochastic optimization, and, through the list-decodable model, robust clustering.

While these techniques give an approach to many problems, the field is still new—there are many unexplored questions. Both the moment estimation and duality results rely on tractable relaxations of resilience. Unfortunately, the spectral norm of the covariance (or of higher moments) is the only such relaxation that is currently known. This relaxation is well-suited to mean estimation, but there are surely other yet-undiscovered relaxations more suitable for other problems. One particular challenge problem is the semi-random stochastic block model presented in Section 3.3. The best efficient algorithms perform far worse than the information-theoretic threshold (Charikar et al., 2017; Steinhardt, 2017; McKenzie et al., 2018). It seems likely that no algorithm based on second moment estimation can circumvent this, so any improvement on the known bounds will likely lead to new algorithmic techniques. Another challenge problem is robust classification, where we are given a $(1-\epsilon)$-fraction of (x, y) pairs that can be linearly separated by a vector w^*, and the remaining ϵ-fraction of points are arbitrary outliers. Without further assumptions, robust classification is computationally hard (Guruswami and Raghavendra, 2009; Feldman et al., 2009). However, the known conditions that enable robust classification hew closely to the bounded covariance assumption (Klivans et al., 2009; Awasthi et al., 2014; Diakonikolas et al., 2017b; 2018a), which is not a natural fit for classification because points far from the decision boundary increase the covariance but intuitively should not increase the difficulty of classification.

For mean estimation, we can hope for better dependence on ϵ as $\epsilon \to 0$. The second moment

algorithms presented here achieve an $\mathcal{O}(\sqrt{\epsilon})$ dependence, but with better control over the tails one might hope for e.g. an $\tilde{\mathcal{O}}(\epsilon)$ dependence. This distinction likely matters in practice. Unfortunately, known algorithms for surpassing $\sqrt{\epsilon}$ either rely strongly on Gaussian assumptions (Diakonikolas et al., 2016), or else involve expensive sum-of-squares algorithms (Kothari and Steinhardt, 2018; Hopkins and Li, 2018). Another shortcoming in the $\epsilon \to 0$ regime is the performance of the duality approach. Algorithm 8 in particular often does not directly produce estimates with vanishing error as $\epsilon \to 0$, and instead must rely on separate post-processing to yield good results.

Finally, while resilience helps to illuminate robust mean estimation, it does not currently apply to more complex problems such as stochastic optimization (although see Steinhardt et al. (2018) for an extension of resilience that handles low-rank approximation). Can we similarly characterize information-theoretic recoverability for these problems? Moreover, how accurate is resilience as a characterization for mean estimation? Does it always give sharp estimates, or are there problems where the upper bound given by resilience differs substantially from the true information-theoretic threshold? We hope that readers of this manuscript are inspired to tackle these problems, and to propose new questions of their own.

Appendix A

Proofs for Chapter 1

A.1 Proof of Lemma 1.1

Let $\mathbb{I}[E]$ denote the indicator that E occurs. Then we have

$$|\mathbb{E}_{X\sim p}[X \mid E] - \mu| = |\mathbb{E}_{X\sim p}[(X-\mu)\mathbb{I}[E]]|/\mathbb{P}[E] \tag{A.1}$$

$$\leq \sqrt{\mathbb{E}_{X\sim p}[(X-\mu)^2\mathbb{I}[E]] \cdot \mathbb{E}_{X\sim p}[\mathbb{I}[E]]}/\mathbb{P}[E] \tag{A.2}$$

$$\leq \sqrt{\sigma^2 \cdot \mathbb{P}[E]}/\mathbb{P}[E] = \sigma/\sqrt{\mathbb{P}[E]}. \tag{A.3}$$

In particular, if we let E_0 be the event that $X \geq \mu + \sigma/\sqrt{\delta}$, we get that $\sigma/\sqrt{\delta} \leq \sigma/\sqrt{\mathbb{P}[E_0]}$, and hence $\mathbb{P}[E_0] \leq \delta$, which proves the first part of the lemma.

For the second part, if $\mathbb{P}[E] \leq \frac{1}{2}$ then (A.3) already implies the desired result since $\sigma/\sqrt{\delta} \leq \sigma\sqrt{2(1-\delta)/\delta}$ when $\delta \leq \frac{1}{2}$. If $\mathbb{P}[E] \geq \frac{1}{2}$, then consider the same argument applied to $\neg E$ (the event that E does not occur). We get

$$|\mathbb{E}_{X\sim p}[X \mid E] - \mu| = \frac{1-\mathbb{P}[E]}{\mathbb{P}[E]}|\mathbb{E}_{X\sim p}[X \mid \neg E] - \mu| \tag{A.4}$$

$$\leq \frac{1-\mathbb{P}[E]}{\mathbb{P}[E]} \cdot \sigma/\sqrt{1-\mathbb{P}[E]}. \tag{A.5}$$

Again the result follows since $\sigma\sqrt{1-\delta}/\delta \leq \sigma\sqrt{2(1-\delta)/\delta}$ when $\delta \geq \frac{1}{2}$.

A.2 Proof of Lemma 1.4

First note that if ($\mathcal{I}$) holds, then most of the mass of the c_i lies within S. More precisely, $\sum_{i\in S} c_i / \sum_{i=1}^n c_i \geq 1-\epsilon$. This is because ($\mathcal{I}$) ensures that the mass removed from $[n]\backslash S$ is greater

than the mass removed from S, which ensures that the ratio is always at least its initial value of $1-\epsilon$.

Now, let $\tilde{\mu}_c = \sum_{i\in S} c_i x_i / \sum_{i\in S} c_i$ be the weighted mean of the good points. Then $|\hat{\mu}_c - \tilde{\mu}_c| \leq \hat{\sigma}\sqrt{\frac{2\epsilon}{1-\epsilon}}$ by Lemma 1.1, since $\hat{\mu}_c$ is the weighted mean of all the points and $\tilde{\mu}_c$ is the mean of the at least $1-\epsilon$ fraction (under c) of remaining good points. In addition, $|\mu - \tilde{\mu}_c| \leq \sigma\sqrt{\frac{2\epsilon}{2-\epsilon}}$, since $\sum_{i\in S} c_i/|S| \geq 1 - \frac{\epsilon}{2}$ (here we think of $\tilde{\mu}_c$ as the mean of an event occuring with probability $\sum_{i\in S} c_i/|S|$ under the uniform distribution on $|S|$).

This establishes the first part of the lemma. For the second part, first note that

$$\sum_{i\in S} c_i\tau_i \overset{(1.1)}{\leq} (1-\epsilon)n\cdot[\sigma^2 + (\mu-\hat{\mu}_c)^2] \tag{A.6}$$

$$\overset{(i)}{\leq} (1-\epsilon)n\cdot[\sigma^2 + (\sigma\sqrt{2\epsilon/(2-\epsilon)} + \hat{\sigma}_c\sqrt{2\epsilon/(1-\epsilon)})^2] \tag{A.7}$$

$$\overset{(ii)}{\leq} (1-\epsilon)\hat{\sigma}_c^2 n\cdot[\tfrac{1}{16} + 2\epsilon\cdot(\tfrac{1}{4\sqrt{2-\epsilon}} + \tfrac{1}{\sqrt{1-\epsilon}})^2] \tag{A.8}$$

$$\leq 0.32(1-\epsilon)\hat{\sigma}_c^2 n \text{ (for } \epsilon \leq 1/12). \tag{A.9}$$

Here (i) is the first part of the lemma, and (ii) uses the assumption that $\sigma^2 \leq \hat{\sigma}_c^2$. The final step is simple calculation. We also have

$$\sum_{i=1}^{n} c_i\tau_i = \hat{\sigma}_c^2\cdot(\sum_{i=1}^{n} c_i\tau_i) \tag{A.10}$$

$$\overset{(iii)}{\geq} \hat{\sigma}_c^2\cdot(1-3\epsilon/2)n \geq \frac{2}{3}\hat{\sigma}_c^2 n. \tag{A.11}$$

Here (iii) uses ($\mathcal{I}$) to conclude that at most an $\frac{\epsilon}{2}$-fraction of the mass is removed from S (since it is less than half the mass removed from $[n]\backslash S$ and hence the total fraction of mass remaining is at least $1-3\epsilon/2$. This completes the lemma.

Appendix B

Proofs for Chapter 2

B.1 Proof of Lemma 2.4

Recall that a distribution p is (σ, ϵ)-resilient iff $\|\mathbb{E}[X - \mu \mid E]\| \leq \sigma$ for all events E with probability at least $1-\epsilon$. Note that this is equivalent to asking that $\|\mathbb{E}[X-\mu \mid E]\| \leq \sigma$ for all E with probability *exactly* $1-\epsilon$—if $\mathbb{P}[E] > 1-\epsilon$, we can remove the points X' from E for which $\langle v, X' \rangle$ is smallest (where v is the dual unit vector to the point $\mathbb{E}[X-\mu \mid E]$) and thus increase $\mathbb{E}[X-\mu \mid E]$. Now note that for events with probability $1-\epsilon$,

$$\|\mathbb{E}[X-\mu \mid E]\| = \left\| -\frac{1-\mathbb{P}[E]}{\mathbb{P}[E]}\mathbb{E}[X-\mu \mid \neg E]\right\| \tag{B.1}$$

$$= \frac{\epsilon}{1-\epsilon}\|\mathbb{E}[X-\mu \mid \neg E]\|, \tag{B.2}$$

and that $\neg E$ is an event with probability ϵ. Therefore, $\|\mathbb{E}[X-\mu \mid E]\| \leq \sigma$ for all E with probability $1-\epsilon$ if and only if $\|\mathbb{E}[X-\mu \mid \neg E]\| \leq \frac{1-\epsilon}{\epsilon}\sigma$ for all $\neg E$ with probability ϵ, as claimed.

B.2 Proof of Lemma 2.6

By Lemma 2.4, it suffices to show that $(1-\epsilon, \frac{1-\epsilon}{\epsilon}\sigma)$-resilience is equivalent to (2.3). Suppose that E is an event with probability ϵ, and let v be such that $\|v\|_* = 1$ and $\langle \mathbb{E}[X-\mu \mid E], v\rangle = \|\mathbb{E}[X-\mu \mid E]\|$.

Then we have

$$\|\mathbb{E}[X-\mu \mid E]\| = \langle \mathbb{E}[X-\mu \mid E], v\rangle \tag{B.3}$$
$$= \langle \mathbb{E}[\langle X-\mu, v\rangle \mid E] \tag{B.4}$$
$$\overset{(i)}{\leq} \mathbb{E}[\langle X-\mu, v\rangle \mid \langle X-\mu, v\rangle \geq \tau_\epsilon(v)] \tag{B.5}$$
$$\overset{(2.3)}{\leq} \frac{1-\epsilon}{\epsilon}\sigma. \tag{B.6}$$

Here (i) is because $\langle X-\mu, v\rangle$ is at least as large for the ϵ-quantile as for any other event E of probability ϵ. This shows that (2.3) implies $(1-\epsilon, \frac{1-\epsilon}{\epsilon}\sigma)$-resilience. For the other direction, given any v let E_v denote the event that $\langle X-\mu, v\rangle \geq \tau_\epsilon(v)$. Then E_v has probability ϵ and hence

$$\mathbb{E}[\langle X-\mu, v\rangle \mid \langle X-\mu, v\rangle \geq \tau_\epsilon(v)] = \mathbb{E}[\langle X-\mu, v\rangle \mid E_v] \tag{B.7}$$
$$= \langle \mathbb{E}[X-\mu \mid E_v], v\rangle \tag{B.8}$$
$$\overset{(ii)}{\leq} \|\mathbb{E}[X-\mu \mid E_v]\| \tag{B.9}$$
$$\overset{(iii)}{\leq} \frac{1-\epsilon}{\epsilon}\sigma, \tag{B.10}$$

where (ii) is Hölder's inequality and (iii) invokes resilience. Therefore, resilience implies (2.3), so the two properties are equivalent, as claimed.

B.3 Proof of Proposition 2.10

We first note that $(e^{\lambda x} - \lambda x - 1) \leq (e^\lambda - \lambda - 1)x^2$ whenever $|x| \leq 1$ and $\lambda \geq 0$. We therefore have

$$\mathbb{E}[\exp(\lambda(X-\mu))] \leq \mathbb{E}[1 + \lambda(X-\mu) + (e^\lambda - \lambda - 1)(X-\mu)^2] \tag{B.11}$$
$$= 1 + (e^\lambda - \lambda - 1)\operatorname{Var}[X] \tag{B.12}$$
$$\leq \exp((e^\lambda - \lambda - 1)\operatorname{Var}[X]) \leq \exp((e^\lambda - \lambda - 1)S), \tag{B.13}$$

as was to be shown.

B.4 Proof of Lemma 2.11

Let E_+ be the event that $\langle x_i - \mu, v\rangle$ is positive, and E_- the event that it is non-negative. Then $\mathbb{P}[E_+] + \mathbb{P}[E_-] = 1$, so at least one of E_+ and E_- has probablity at least $\frac{1}{2}$. Without loss of generality

assume it is E_+. Then we have

$$\frac{1}{|S|}\sum_{i\in S}|\langle x_i-\mu,v\rangle| = \frac{2}{|S|}\sum_{i\in S}\max(\langle x_i-\mu,v\rangle,0) \tag{B.14}$$

$$= 2\mathbb{P}[E_+]\mathbb{E}[\langle x-\mu,v\rangle \mid E_+] \tag{B.15}$$

$$\leq 2\mathbb{P}[E_+]\|\mathbb{E}[x-\mu \mid E_+]\| \leq 2\sigma, \tag{B.16}$$

where the last step invokes resilience applies to E_+ together with $\mathbb{P}[E_+] \leq 1$. Conversely, if S has bounded 1st moments then

$$\mathbb{E}[\langle X-\mu,v\rangle \mid \langle X-\mu,v\rangle \geq \tau_{1/2}(v)] \leq \mathbb{E}[|\langle X-\mu,v\rangle|]/\mathbb{P}[\langle X-\mu,v\rangle \geq \tau_{1/2}(v)] \tag{B.17}$$

$$= 2\mathbb{E}[|\langle X-\mu,v\rangle|] \leq 2\sigma, \tag{B.18}$$

so S is $(2\sigma,\frac{1}{2})$-resilient by Lemma 2.6.

B.5 Proof of Lemma 2.12

First, if S is (σ,ϵ)-resilient around μ_0, then invoking resilience with $T=S$ yields $\|\mu-\mu_0\| \leq \sigma$. It then follows that

$$\|\frac{1}{|T|}\sum_{i\in T}(x_i-\mu)\| \overset{(i)}{\leq} \|\frac{1}{|T|}\sum_{i\in T}(x_i-\mu_0)\| + \|\mu-\mu_0\| \tag{B.19}$$

$$\overset{(ii)}{\leq} \sigma+\sigma = 2\sigma \tag{B.20}$$

whenever $|T| \geq (1-\epsilon)|S|$. Here (i) is the triangle inequality and (ii) invokes resilience around μ_0.

Conversely, if S is (σ,ϵ)-resilient around its actual mean μ, we have

$$\|\frac{1}{|T|}\sum_{i\in T}(x_i-\mu_0)\| \overset{(i)}{\leq} \|\frac{1}{|T|}\sum_{i\in T}(x_i-\mu)\| + \|\mu-\mu_0\| \tag{B.21}$$

$$\overset{(ii)}{\leq} \sigma + \|\mu-\mu_0\|, \tag{B.22}$$

where (i) is again the triangle inequality and (ii) again invokes resilience (this time around μ). This completes the proof.

B.6 Proof of Lemma 2.14

First, we have

$$\sum_{i\in S} c_i\tau_i \le \sum_{i\in S}\tau_i \tag{B.23}$$

$$= \sum_{i\in S}\max(\langle x_i-\mu, v_j\rangle - \sigma, 0) \tag{B.24}$$

$$\le \sum_{i\in S}\langle x_i-\mu, v_j\rangle \mathbb{I}[\langle x_i-\mu, v_j > \sigma] \tag{B.25}$$

$$\overset{(i)}{\le} \frac{\alpha}{32\log(4/3\alpha)}|S|\cdot\sigma = \frac{\alpha^2}{32\log(4/3\alpha)}n. \tag{B.26}$$

Here (i) invokes $(\sigma, \frac{\alpha}{32\log(4/3\alpha)})$-resilience of S, together with the fact that at most $\frac{\alpha}{32\log(4/3\alpha)}|S|$ elements of S can exceed σ. Next, let Q be the indices of the $\frac{\alpha}{8}$-fraction of largest values under p. Then we have

$$\sum_{i=1}^n c_i\tau_i = \sum_{i=1}^n c_i\max(\langle x_i-\mu, v_j\rangle - \sigma, 0) \tag{B.27}$$

$$\ge \sum_{i\in Q} c_i(\langle x_i-\mu, v_j\rangle - \sigma) \tag{B.28}$$

$$\overset{(ii)}{\ge} (\sum_{i\in Q} c_i)\cdot(2\sigma-\sigma) \tag{B.29}$$

$$= \frac{\alpha}{8}\sigma\cdot\sum_{i=1}^n c_i. \tag{B.30}$$

Here (ii) invokes the fact that the weighted average of $\langle x_i-\mu, v_j\rangle$ across Q is σ_+ by definition, which is in turn at least 2σ.

B.7 Proof of Lemma 2.15

Let c_i' denote the value after the update and c_i denote the value before the update. Then we have

$$\sum_{i=1}^n c_i' = \sum_{i=1}^n c_i - \frac{1}{\tau_{\max}}\sum_{i=1}^n c_i\tau_i \tag{B.31}$$

$$\overset{(2.18)}{\le} \sum_{i=1}^n c_i - \frac{1}{\tau_{\max}}\frac{\alpha\sigma}{8}\sum_{i=1}^n c_i \tag{B.32}$$

$$= (\sum_{i=1}^n c_i)(1-\frac{1}{\tau_{\max}}\frac{\alpha\sigma}{8}). \tag{B.33}$$

On the other hand, we have

$$\sum_{i\in S} c_i - c_i' = \frac{1}{\tau_{\max}} \sum_{i\in S} c_i \tau_i \tag{B.34}$$

$$\overset{(2.17)}{\leq} \frac{1}{\tau_{\max}} \frac{\alpha^2}{32\log(4/3\alpha)} \sigma \cdot n. \tag{B.35}$$

Combining these and assuming that $(\mathcal{R})$ holds for the c_i, we have

$$\sum_{i=1}^n c_i' \leq \big(\sum_{i=1}^n c_i\big)\Big(1 - \frac{4\log(4/3\alpha)}{\alpha n} \sum_{i\in S}(c_i - c_i')\Big) \tag{B.36}$$

$$\leq \big(\sum_{i=1}^n c_i\big) \exp\Big(-\frac{4\log(4/3\alpha)}{\alpha n} \sum_{i\in S}(c_i - c_i')\Big) \tag{B.37}$$

$$\overset{(\mathcal{R})}{\leq} n \exp\Big(-\frac{4\log(4/3\alpha)}{\alpha n} \sum_{i\in S}(1 - c_i)\Big) \cdot \exp\Big(-\frac{4\log(4/3\alpha)}{\alpha n} \sum_{i\in S}(c_i - c_i')\Big) \tag{B.38}$$

$$= n \exp\Big(-\frac{4\log(4/3\alpha)}{\alpha n} \sum_{i\in S}(1 - c_i')\Big), \tag{B.39}$$

which shows that $(\mathcal{R})$ is indeed preserved. Finally, suppose that $\sum_{i\in S}(1-c_i) = \frac{\alpha}{4}n + \delta$ and hence $\sum_{i\in S} c_i = \frac{3\alpha}{4}n - \delta$. Then $(\mathcal{R})$ implies that $\sum_{i=1}^n c_i \leq \frac{3\alpha}{4}n \cdot \exp(-4\log(4/3\alpha)\delta/\alpha n) \leq \frac{3\alpha}{4}n - 3\log(4/3\alpha)\delta < \sum_{i\in S} c_i$, which is a contradiction. This implies that $\sum_{i\in S}(1-c_i) \leq \frac{\alpha}{4}n$, as claimed.

Appendix C

Proofs for Chapter 3

C.1 Proof of Lemma 3.5

We start by taking a continuous relaxation of (3.24), asking for weights $c_i \in [0,1]$ rather than $\{0,1\}$:

$$\min_{c\in[0,1]^n, \|c\|_1 \geq \frac{3n}{4}} \max_{\|v\|_* \leq 1} \frac{1}{n}\sum_{i=1}^n c_i |\langle x_i, v\rangle|^2. \tag{C.1}$$

Note that we strengthened the inequality to $\|c\|_1 \geq \frac{3n}{4}$, whereas in (3.24) it was $\|c\|_1 \geq \frac{n}{2}$. Given any solution $c_{1:n}$ to (C.1), we can obtain a solution c' to (3.24) by letting $c'_i = \mathbb{I}[c_i \geq \frac{1}{2}]$. Then $c'_i \in \{0,1\}$ and $\|c'\|_1 \geq \frac{n}{2}$. Moreover, $c'_i \leq 2c_i$, so $\frac{1}{n}\sum_{i=1}^n c'_i |\langle x_i, v\rangle|^2 \leq \frac{2}{n}\sum_{i=1}^n c_i |\langle x_i, v\rangle|^2$ for all v. Therefore, the value of (3.24) is at most twice the value of (C.1).

Now, by the minimax theorem, we can swap the min and max in (C.1) in exchange for replacing the single vector v with a distribution over vectors v_j, thus obtaining that (C.1) is equal to

$$\lim_{m\to\infty} \max_{\substack{\alpha_1+\cdots+\alpha_m \leq 1 \\ \alpha \geq 0, \|v_j\|_* \leq 1}} \min_{\substack{c\in[0,1]^n \\ \|c\|_1 \geq \frac{3n}{4}}} \frac{1}{n}\sum_{i=1}^n c_i \sum_{j=1}^m \alpha_j |\langle x_i, v_j\rangle|^2. \tag{C.2}$$

By letting $v'_j = \alpha_j v_j$, the above is equivalent to optimizing over v_j satisfying $\sum_j \|v_j\|_*^2 \leq 1$:

$$\lim_{m\to\infty} \max_{\|v_1\|_*^2+\cdots+\|v_m\|_*^2 \leq 1} \min_{\substack{c\in[0,1]^n \\ \|c\|_1 \geq \frac{3n}{4}}} \frac{1}{n}\sum_{i=1}^n c_i \sum_{j=1}^m |\langle x_i, v_j\rangle|^2. \tag{C.3}$$

For any $v_1, \ldots, v_m$, we will find c such that the above sum is bounded. Indeed, define $B(v_{1:m})$ to be $\frac{1}{n}\sum_{i=1}^n \sqrt{\sum_{j=1}^m |\langle x_i, v_j\rangle|^2}$. Then take $c_i = \mathbb{I}[\sum_{j=1}^m |\langle x_i, v_j\rangle|^2 < 16B^2]$, which has $\|c\|_1 \geq \frac{3n}{4}$ by Markov's inequality, and for which $\sum_i c_i \sum_j |\langle x_i, v_j\rangle|^2 \leq 4B(v_{1:m})^2$.

Therefore, the value of (C.1) is bounded by $\max_{m, v_{1:m}} 4B(v_{1:m})^2$, and so the value of (3.24) is

bounded by $\max_{m,v_{1:m}} 8B(v_{1:m})^2$, which yields the desired result.

Appendix D

Proofs for Chapter 5

D.1 Proof of (5.39)

Note that

$$\sum_{i\in S} c_i(f_i(\overline{w}) - f_i(\tilde{w})) = \int_0^1 \sum_{i\in S} c_i \langle \nabla f_i(s\overline{w} + (1-s)\tilde{w}), \overline{w} - \tilde{w}\rangle ds \tag{D.1}$$

$$\overset{(i)}{\leq} \alpha n\sigma \|\overline{w} - \tilde{w}\|_2 + \int_0^1 \sum_{i\in S} c_i \langle \nabla \bar{f}(s\overline{w} + (1-s)\tilde{w}), \overline{w} - \tilde{w}\rangle ds \tag{D.2}$$

$$= \alpha n\sigma \|\overline{w} - \tilde{w}\|_2 + \bar{f}(\overline{w}) - \bar{f}(\tilde{w}) \tag{D.3}$$

$$\overset{(ii)}{\leq} \alpha n\sigma \|\overline{w} - \tilde{w}\|_2. \tag{D.4}$$

Here (i) uses the bound (5.37) to conclude that ∇f_i and $\nabla \bar{f}$ are close, while (ii) uses the optimality of $\overline{w}$ for $\bar{f}$.

D.2 Bounding $\|\hat{\mu} - \mu\|_2$

We have

$$\|\hat{\mu} - \mu\|_2 = \|\sum_{i=1}^{n} c_i(\hat{w}_i - \mu)/\sum_{i=1}^{n} c_i\|_2 \tag{D.5}$$

$$\leq \frac{\sum_{i\in S} c_i}{\sum_{i=1}^{n} c_i}\Big(\sum_{i\in S} c_i\|\hat{w}_i - \mu\|_2/\sum_{i\in S} c_i\Big) + \frac{\sum_{i\notin S} c_i}{\sum_{i=1}^{n} c_i}\Big(\sum_{i\in S} c_i\|\hat{w}_i - \mu\|_2/\sum_{i\notin S} c_i\Big) \tag{D.6}$$

$$\overset{(i)}{\leq} \frac{\sum_{i\in S} c_i}{\sum_{i=1}^{n} c_i} \cdot \mathcal{O}(\sqrt{\sigma r/\sqrt{\alpha}}) + \frac{\sum_{i\notin S} c_i}{\sum_{i=1}^{n} c_i} \cdot 2r \tag{D.7}$$

$$\overset{(ii)}{\leq} \mathcal{O}(\sqrt{\sigma r/\sqrt{\alpha}}) + \frac{2(1-\alpha)}{1-\alpha(1-\alpha)/4} r. \tag{D.8}$$

Here (i) uses the bound (5.51) together with $\|\hat{w}_i - \mu\|_2 \leq 2r$, and (ii) uses Lemma 5.6 to bound $\frac{\sum_{i\notin S} c_i}{\sum_{i=1}^{n} c_i}$. Suppose that $\alpha > 0.55$ (i.e., $\epsilon = 1 - \alpha < 0.45$). Then simple calculation shows that $\frac{2(1-\alpha)}{1-\alpha(1-\alpha)/4} < 0.96$. Therefore, we have $\|\hat{\mu} - \mu\|_2 \leq \mathcal{O}(\sqrt{\sigma r}) + 0.96r$, as claimed.

Bibliography

P. Awasthi, M. F. Balcan, and P. M. Long. The power of localization for efficiently learning linear separators with noise. In *Symposium on Theory of Computing (STOC)*, pages 449–458, 2014.

M. Balcan, A. Blum, and S. Vempala. A discriminative framework for clustering via similarity functions. In *Symposium on Theory of Computing (STOC)*, pages 671–680, 2008.

M. F. Balcan, H. Röglin, and S. Teng. Agnostic clustering. In *International Conference on Algorithmic Learning Theory*, pages 384–398, 2009.

B. Barak and D. Steurer. Proofs, beliefs, and algorithms through the lens of sum-of-squares. https://www.sumofsquares.org/public/index.html, 2016.

J. Batson, D. A. Spielman, and N. Srivastava. Twice-Ramanujan sparsifiers. *SIAM Journal on Computing*, 41(6):1704–1721, 2012.

G. E. Box. Non-normality and tests on variances. *Biometrika*, 40:318–335, 1953.

M. Charikar, J. Steinhardt, and G. Valiant. Learning from untrusted data. In *Symposium on Theory of Computing (STOC)*, 2017.

A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6), 2011.

I. Diakonikolas, G. Kamath, D. Kane, J. Li, A. Moitra, and A. Stewart. Robust estimators in high dimensions without the computational intractability. In *Foundations of Computer Science (FOCS)*, 2016.

I. Diakonikolas, G. Kamath, D. Kane, J. Li, A. Moitra, and A. Stewart. Being robust (in high dimensions) can be practical. *arXiv*, 2017a.

I. Diakonikolas, D. M. Kane, and A. Stewart. Learning geometric concepts with nasty noise. *arXiv*, 2017b.

I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, J. Steinhardt, and A. Stewart. Sever: A robust meta-algorithm for stochastic optimization. *arXiv preprint arXiv:1803.02815*, 2018a.

I. Diakonikolas, D. M. Kane, and A. Stewart. List-decodable robust mean estimation and learning mixtures of spherical Gaussians. In *Symposium on Theory of Computing (STOC)*, 2018b.

D. L. Donoho. Breakdown properties of multivariate location estimators. Ph.D. qualifying paper, 1982.

D. L. Donoho and M. Gasko. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Annals of Statistics*, 20(4):1803–1827, 1992.

S. S. Du, S. Balakrishnan, and A. Singh. Computationally efficient robust estimation of sparse functionals. *arXiv preprint arXiv:1702.07709*, 2017.

U. Feige and J. Kilian. Heuristics for semirandom graph problems. *Journal of Computer and System Sciences*, 63(4):639–671, 2001.

V. Feldman, P. Gopalan, S. Khot, and A. K. Ponnuswami. On agnostic learning of parities, monomials, and halfspaces. *SIAM Journal on Computing*, 39(2):606–645, 2009.

P. E. Frenkel and P. Horváth. Minkowski's inequality and sums of squares. *Central European Journal of Mathematics*, 12(3):510–516, 2014.

O. Guédon and R. Vershynin. Community detection in sparse networks via Grothendieck's inequality. *arXiv*, 2014.

V. Guruswami and P. Raghavendra. Hardness of learning halfspaces with noise. *SIAM Journal on Computing*, 39(2):742–765, 2009.

U. Haagerup. The best constants in the khintchine inequality. *Studia Mathematica*, 70(3):231–283, 1981.

F. R. Hampel. *Contributions to the theory of robust estimation*. PhD thesis, University of California at Berkeley, 1968.

F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley, 2011.

S. B. Hopkins and J. Li. Mixture models, robustness, and sum of squares proofs. In *Symposium on Theory of Computing (STOC)*, 2018.

P. J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35 (1):73–101, 1964.

P. J. Huber and E. M. Ronchetti. *Robust Statistics*. Wiley, 2009.

D. S. Johnson and F. P. Preparata. The densest hemisphere problem. *Theoretical Computer Science*, 6:93–107, 1978.

W. E. Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8:118–127, 2007.

M. Kearns and M. Li. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 22 (4):807–837, 1993.

A. Khintchine. Über dyadische brüche. *Mathematische Zeitschrift*, 18:109–116, 1923.

A. R. Klivans, P. M. Long, and R. A. Servedio. Learning halfspaces with malicious noise. *Journal of Machine Learning Research (JMLR)*, 10:2715–2740, 2009.

P. Kothari and J. Steinhardt. Better agnostic clustering via tensor norms. In *Symposium on Theory of Computing (STOC)*, 2018.

S. Kushagra, S. Samadi, and S. Ben-David. Finding meaningful cluster structure amidst background noise. In *International Conference on Algorithmic Learning Theory*, pages 339–354, 2016.

K. A. Lai, A. B. Rao, and S. Vempala. Agnostic estimation of mean and covariance. In *Foundations of Computer Science (FOCS)*, 2016.

C. M. Le, E. Levina, and R. Vershynin. Concentration and regularization of random graphs. *arXiv*, 2015.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer Berlin Heidelberg, 1991.

J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10), 2010.

J. Li. Robust sparse estimation tasks in high dimensions. *arXiv*, 2017.

M. Mahajan, P. Nimbhorkar, and K. Varadarajan. The planar k-means problem is NP-hard. *International Workshop on Algorithms and Computation*, pages 274–285, 2009.

R. A. Maronna. Robust estimation of multivariate location and scatter. *Annals of Statistics*, 4(1): 51–67, 1976.

L. Massoulié. Community detection thresholds and the weak Ramanujan property. In *Symposium on Theory of Computing (STOC)*, pages 694–703, 2014.

T. McKenzie, H. Mehta, and L. Trevisan. A new algorithm for the robust semi-random independent set problem. *arXiv*, 2018.

E. Mossel, J. Neeman, and A. Sly. A proof of the block model threshold conjecture. *arXiv*, 2013.

S. Newcomb. A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics*, 8(4):343–366, 1886.

A. Prasad, A. S. Suggala, S. Balakrishnan, and P. Ravikumar. Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*, 2018.

E. Rebrova and K. Tikhomirov. Coverings of random ellipsoids, and invertibility of matrices with iid heavy-tailed entries. *arXiv*, 2015.

E. Rebrova and R. Vershynin. Norms of random matrices: local and global problems. *arXiv*, 2016.

S. Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, The Hebrew University of Jerusalem, 2007.

M. Sion. On general minimax theorems. *Pacific journal of mathematics*, 8(1):171–176, 1958.

J. Steinhardt. Does robustness imply tractability? A lower bound for planted clique in the semi-random model. *arXiv*, 2017.

J. Steinhardt, G. Valiant, and M. Charikar. Avoiding imposters and delinquents: Adversarial crowdsourcing and peer prediction. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

J. Steinhardt, P. W. Koh, and P. Liang. Certified defenses for data poisoning attacks. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

J. Steinhardt, M. Charikar, and G. Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. In *Innovations in Theoretical Computer Science (ITCS)*, 2018.

J. W. Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, 2:448–485, 1960.

J. W. Tukey. Mathematics and picturing of data. In *ICM*, volume 6, pages 523–531, 1975.