Lecture 3: Robust mean estimation in high dimensions

October 2, 2019

1 Introduction

We have so far only considered robust mean estimation for univariate distributions. We now turn our attention to the much more complicated problem of robust mean estimation for multivariate distributions. We will consider the natural high dimensional analogs of the two settings we considered previously, and we'll see the information theoretic rates we can achieve, as well as some of the difficulties with coming up with efficient algorithms for these problems that achieve these rates. The two problems we will consider are:

Problem 1.1 (Robust mean estimation for Gaussians in high dimensions)**.** *Let $\mu \in \mathbb{R}^d, \sigma > 0$, and let $\varepsilon \in [0, 1/2)$. Let $S = \{X_1, \ldots, X_n\}$ be an ε-corrupted set of samples from $\mathcal{N}(\mu, \sigma^2 I)$. Given S, ε, σ, output $\widehat{\mu}$ minimizing $\|\widehat{\mu} - \mu\|_2$.*

One can also ask the question when the covariance is not spherical and not known. If the covariance is known but non-spherical, one can appropriately rotate the problem and solve the problem in the rotated space. As we shall see in a future class, when the distribution is truly Gaussian, we can also learn the covariance to high precision even in the presence of outliers, so that we can also appropriately rotate.

Problem 1.2 (Robust mean estimation with bounded second moments in high dimensions)**.** *Let $\mu \in \mathbb{R}^d, \sigma > 0$, and let $\varepsilon \in [0, 1/2)$. Let D be a distribution with mean μ and covariance $\Sigma \preceq \sigma^2 I$. Let $S = \{X_1, \ldots, X_n\}$ be an ε-corrupted set of samples from D. Given S, ε, σ, output $\widehat{\mu}$ minimizing $\|\widehat{\mu} - \mu\|_2$.*

As a remark, note that for both problems, our metric of recovery will be ℓ_2 norm. We shall see later that this is indeed the "correct" metric for closeness. Intuitively, this is because both problems have some underlying ℓ_2-structure: Gaussians are invariant under rotations (an ℓ_2-rotation), and spectral properties (including PSD ordering) is also an ℓ_2 property.

As these problems are strict generalizations of the 1D problem, it is not hard to see that $\Omega(\varepsilon)$ and $\Omega(\sqrt{\varepsilon})$ are still lower bounds for Problems 1.1 and 1.2, respectively. In this lecture, we will also show that there exist inefficient estimators which match these lower bounds. However, a number of naive estimators do not attain this bound, and have error which get worse as the dimension increases, as we shall see next.

2 Things that don't work

Here we'll give two natural estimators which fail to get the right rates. There is a veritable zoo of such estimators, but these will demonstrate the basic difficulties of designing good estimators for this problem.

Learning coordinate-wise For both problems, one natural attempt would be to simply reduce the problem to a series of d univariate learning problems. For instance, for Problem 1.1, every projection of a Gaussian is a Gaussian, and hence we can use the univariate estimator (i.e. the median) to learn each coordinate to error $O(\varepsilon)$. However, aggregating over all d directions yields an error $O(\varepsilon\sqrt{d})$, which is not optimal.

Truncated mean Another try is a truncated mean, as in the univariate case. Probably the most natural way to do this is to try to find a threshold T, throw away all points $X \in S$ so that $\|X - \mu\|_2 > T$, and take the mean of the remaining points. It is not immediately obvious how one can do something like this, but even if we could, this will be insufficient. The problem is that this threshold T will necessarily scale with the dimension:

Fact 2.1. *Let $X \sim \mathcal{N}(\mu, I)$. Then there is some universal constant $c > 0$ so that*

$$\Pr\left[\left|\|X - \mu\|_2^2 - d\right| > t\sqrt{d}\right] \leq 2\exp(-ct) \; .$$

Proof. Let Y_i be the i-th component of $X - \mu$, so that the Y_i's are all i.i.d. as $\mathcal{N}(0,1)$. Then $\|X - \mu\|_2^2 = \sum_{i=1}^d Y_i^2$. We know that $\mathbb{E}[Y_i^2] = 1$, and each Y_i is a sub-exponential random variable. The result then follows from sub-exponential concentration bounds. □

In particular, this implies that the threshold T must be at least some constant multiple of $\sqrt{d}$. But if that's the case, the outliers can all be $\sqrt{d}$ far from μ, and it is not hard to see that since there are an ε-fraction of them, this can cause error $\Omega(\varepsilon\sqrt{d})$ for Problem 1.1. In contrast, the estimators we will develop later in this lecture achieve error $O(\varepsilon)$ for this problem. When d is large and ε is reasonably sized, say 0.1, this additional factor can easily render the algorithm statistically useless.

The fact that norm-based statistics are so noisy in high dimensions is quite problematic, and causes many natural candidate algorithms to suffer a loss which scales as $\Omega(\sqrt{d})$, or worse. For instance, even though the median is robust in one dimension, some natural generalizations of it based on norms such as geometric median fail to achieve a good error in high dimensions. As we shall see, to avoid this problem, it will be very important to always try to only consider statistics of projections of our data onto low-dimensional spaces.

3 Tukey median

The Tukey median is a way to make this idea more concrete. Specifically, for any set of points $S \subset \mathbb{R}^d$ of size n, and any point $\eta \in \mathbb{R}^d$, define the *Tukey depth*, or simply *depth* [1], of this point with respect to S, denoted $\mathrm{depth}(S, \eta)$ to be

$$\mathrm{depth}(S, \eta) = \inf_{\|v\|_2 = 1} \frac{|\{X \in S : \langle X - \eta, v\rangle \geq 0\}|}{n} \; . \tag{1}$$

In other words, the depth of η with respect to S is the minimum over all half-planes that pass through η of the fraction of points on any side of that half-space. Intuitively, the larger the depth of η, the more central the point is with respect to the dataset S. The *Tukey median* of S, which we will denote $\mathrm{Tukey}(S)$, is simply defined to be the most central point, with respect to this notion of depth:

$$\mathrm{Tukey}(S) = \arg\max_{\eta} \mathrm{depth}(S, \eta) \; . \tag{2}$$

Observe that this is fundamentally a property of projections of the data. Thus this can hope to circumvent these dimensionality issues that faced the previous estimators we considered. Indeed, we will show:

Theorem 3.1. *Let $S \subset \mathbb{R}^d$ be an ε-corrupted set of samples of size n from $\mathcal{N}(\mu, I)$, where $\varepsilon < 1/6$. Then, there exists some universal constant $C > 0$ so that with probability $1 - \delta$, we have that*

$$\|\mathrm{Tukey}(S) - \mu\|_2 \leq \Phi^{-1}\left(\frac{1}{2} + 3\varepsilon + C\sqrt{\frac{d + \log 1/\delta}{n}}\right) \; .$$

In particular, if n is sufficiently large, and ε is relatively small, then the RHS is $O(\varepsilon)$. As it turns out, the rate of convergence here is also minimax optimal, so this algorithm obtains both the right asymptotic error, as well as the tight rate of convergence.

Proof sketch of Theorem 3.1. For now, let's ignore issues of concentration, and see why we should expect that when we get enough samples, we can get error $O(\varepsilon)$. Recall that we can always write our set S using its canonical decomposition as $S = S_{\text{good}} \cup S_{\text{bad}} \setminus S_r$, where $|S_{\text{good}}| = n$, $|S_{\text{bad}}| = \varepsilon n$, and $|S_r| = \varepsilon n$. Suppose for now that for every point $\eta \in \mathbb{R}^d$ and all unit vectors v, we had that

$$\frac{|\{X \in S_{\text{good}} : \langle X - \eta, v\rangle \geq 0\}|}{|S_{\text{good}}|} = \Pr_{X \sim \mathcal{N}(\mu, I)}[\langle X - \eta, v\rangle \geq 0] = \Phi(\langle \mu - \eta, v\rangle) \,. \tag{3}$$

Then the analysis of the Tukey median would be almost identical to the analysis of the univariate median. First, for any unit vector v, we have that

$$\begin{aligned}|\{X \in S : \langle X - \mu, v\rangle \geq 0\}| &\geq |\{X \in S_{\text{good}} : \langle X - \mu, v\rangle \geq 0\}| + |\{X \in S_{\text{bad}} : \langle X - \mu, v\rangle \geq 0\}| - \varepsilon n \\ &\geq \frac{1}{2}|S_{\text{good}}| - \varepsilon n = \left(\frac{1}{2} - \varepsilon\right) n \,,\end{aligned}$$

so in particular $\text{depth}(S, \mu) \geq 1/2 - \varepsilon$. On the other hand, consider any point $\eta \in \mathbb{R}^d$ that satisfies $\|\eta - \mu\|_2 > \Phi^{-1}(1/2 + 3\varepsilon)$. Let $v = \frac{\eta - \mu}{\|\eta - \mu\|_2}$. Then

$$\begin{aligned}|\{X \in S : \langle X - \eta, v\rangle \geq 0\}| &\leq |\{X \in S_{\text{good}} : \langle X - \eta, v\rangle \geq 0\}| + \varepsilon n \\ &= |S_{\text{good}}| \cdot \Phi(-\|\eta - \mu\|_2) + \varepsilon n \\ &= |S_{\text{good}}| \cdot \left(\frac{1}{2} - 3\varepsilon\right) + \varepsilon n < \left(\frac{1}{2} - \varepsilon\right) n \,.\end{aligned}$$

Therefore in particular, no point with distance from μ greater than $\Phi^{-1}(1/2 + 3\varepsilon)$ can have depth greater than μ. This in particular implies that the true Tukey median must have distance at most $\Phi^{-1}(1/2 + 3\varepsilon)$ from μ.

However, this analysis only works in an infinite sample regime. To get finite sample guarantees, we need that these empirical statistics concentrate, i.e. that (3) holds except up to some small error. One can indeed prove this via empirical process theory:

Fact 3.2 (Theorem 6 in [2])**.** *Let $Y_1, \ldots, Y_m \sim \mathcal{N}(\mu, I)$, and let $\delta > 0$. Then, with probability $1 - \delta$, we have*

$$\left|\frac{|\{i \in [m] : \langle Y_i - \eta, v\rangle \geq 0\}|}{m} - \Pr_{Y \sim \mathcal{N}(\mu, I)}[\langle Y - \eta, v\rangle \geq 0]\right| \lesssim \sqrt{\frac{d + \log 1/\delta}{n}} \,, \tag{4}$$

for all η and all unit vectors $v \in \mathbb{R}^d$ simultaneously.

Plugging this bound into the same calculation as above will yield the Theorem. □

4 Bounded cores

Tukey median establishes the right minimax rate for robust mean estimation for Gaussians, but it uses relatively strong properties of Gaussians to do so. Here we will turn our attention to a more recently introduced notion, namely *resilience*, which recovers the information-theoretic rates (or close to them) for robust mean estimation in a much more general setting.

Definition 4.1 (Resilience, see [3])**.** Let D be a distribution over $\mathbb{R}^d$ with mean μ. Let $\sigma, \varepsilon > 0$. We say that D is (σ, ε)-*resilient* if for all events E so that $\Pr_D[E] \geq 1 - \varepsilon$, then

$$\left\|\mathop{\mathbb{E}}_{D}[X|E] - \mu\right\|_2 \leq \sigma \,. \tag{5}$$

Given a dataset S of size n, we say that S is (σ, ε)-resilient if the uniform distribution over S is (σ, ε)-resilient.

In other words, a distribution is resilient if conditioned on any large probability event, the mean cannot change by much. One can also define a generalization of this to other norms (see [3]). This characterization is useful because it gives a very simple (inefficient) algorithm to recover the mean, given corruption:

Theorem 4.1. *Let $\sigma > 0$, and let $\varepsilon < 1/4$. Let S be an ε-corrupted dataset, and let $S = S_{\text{good}} \cup S_{\text{bad}} \setminus S_r$ be its canonical decomposition. Let μ_g be the empirical mean of S_{good}, and suppose that S_{good} is $(\sigma, 2\varepsilon)$-resilient. Then, there exists an (inefficient) algorithm which, given S, ε, σ, finds $\widehat{\mu}$ so that $\|\widehat{\mu} - \mu_g\|_2 \leq 2\sigma$.*

Proof. First, notice that $S_{\text{good}} \setminus S_r$ is $(2\sigma, \varepsilon/(1-\varepsilon))$-resilient (why?). Given this, the algorithm is simple: given S, find any subset T of size $(1-\varepsilon)n$ which is $(2\sigma, \varepsilon/(1-\varepsilon))$-resilient, and output the empirical mean $\widehat{\mu}$ of T. Such a set exists (as $S_{\text{good}} \setminus S_r$ is a valid solution). We now prove that this has the desired properties. Let $A = T \cap S_{\text{good}}$. Notice that $|A| \geq (1-2\varepsilon)n$. By the $(\sigma, 2\varepsilon)$-resilience of S_{good}, we know that if μ_A is the empirical mean of A, then $\|\mu_A - \mu_g\|_2 \leq \sigma$. On the other hand, by the resilience of T, we know that $\|\mu_A - \mu_g\|_2 \leq \sigma$. The result follows from a triangle inequality. □

This result says that given an ε-corrupted version of a resilient dataset, we can recover the empirical mean of the original data set up to some error. To go from this to learning the mean of the distribution itself, we would additionally need concentration inequalities which says that the empirical mean is close to the true mean, and that the empirical distribution is resilient with high probability.

5 Computational complexity

While these estimators are great from a purely statistical perspective, it is not clear how to implement them efficiently. For instance:

Theorem 5.1 ([4])**.** *Computing the Tukey median of an arbitrary set of points is NP-hard.*

In general, the above reference [4] demonstrates the worst-case hardness of computing a number of natural robust estimators that have good statistical properties. Similarly, it was shown recently:

Theorem 5.2 ([5])**.** *Under the SSE hypothesis, for any $c > 0$, and $\varepsilon > 0$ sufficiently small, certifying $(\varepsilon^{1/2+c}, \varepsilon)$-resilience is is NP-hard.*

Note that the latter theorem only holds for some parameter regimes of resilience. This is because, as we will discuss in the next lecture, there are related efficiently certifiable bounds which, in some regimes, allow us to match these bounds efficiently.

This interplay between statistical efficiency and robust statistics is very interesting and will be a recurring topic throughout this unit. As we shall see, many statistical-computational tradeoffs arise only in the presence of these worst-case corruptions.

References

[1] John W Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, volume 2, pages 523–531, 1975.

[2] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[3] Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[4] Thorsten Bernholt. Robust estimators are hard to compute. Technical report, Technical Report/Universität Dortmund, SFB 475 Komplexitätsreduktion in ..., 2006.

[5] Samuel B Hopkins and Jerry Li. How hard is robust mean estimation? *arXiv preprint arXiv:1903.07870*, 2019.