PRASAD RAGHAVENDRA

CS294: EFFICIENT ALGORITHMS AND COMPUTATIONAL COMPLEXITY IN STATISTICS

Preliminaries

Useful facts in probability

Lemma 1. *(Chebyshev's inequality) Suppose X is a real-valued random variable such that its mean is μ and variance σ^2 then for all $C > 0$,*

$$\Pr[|X - \mu| \geq C\sigma] \leq \frac{1}{C^2}$$

Lemma 2. *(Chernoff's bound) Suppose $X_1, \ldots, X_k$ are $\{0,1\}$-valued random variables then,*

$$\Pr\left[\left|\sum_{i=1}^{k} X_i - \mathbb{E}[\sum_{i=1}^{k} X_i]\right| \geq \delta \cdot k\right] \leq e^{-\delta^2 k/2}$$

Lemma 3. *(McDiarmid's inequality) A function $f : \mathcal{X}_1 \times \mathcal{X}_2 \times \cdot \mathcal{X}_n$ is said to satisfy the bounded difference property if for some $L > 0$, for all i, changing the value of the i^{th} coordinate changes the value of the function by at most L. Formally,*

$$\sup_{\substack{x_1 \in \mathcal{X}_1, \ldots, x_n \in \mathcal{X}_n \\ x_i' \in \mathcal{X}_i}} \left| f(x_1, x_2, \ldots, x_{i-1}, x_i, \ldots, x_n) - f(x_1, x_2, \ldots, x_{i-1}, x_i', \ldots, x_n) \right| \leq L$$

Then for $\epsilon > 0$,

$$\Pr\left[\left|f(x_1, \ldots, x_n) - \mathbb{E}[f(x_1, \ldots, x_n)]\right| \geq \epsilon\right] \leq 2\exp\left(-\frac{2\epsilon^2}{nL^2}\right)$$

Lecture 1

Heavy-tailed mean estimation on $\mathbb{R}$

Let $\mathcal{D}$ be a probability distribution over $\mathbb{R}$. Given samples from $\mathcal{D}$, the most natural estimator for the mean is the so-called "empirical mean" given by,

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where $x_1, \ldots, x_n$ are i.i.d samples from $\mathcal{D}$.

Suppose $\mathcal{D}$ is the Gaussian distribution with mean μ and variance σ^2. Then the empirical mean $\hat{\mu}$ is also a Gaussian with mean μ and variance $\frac{1}{n} \cdot \sigma^2$. In particular, the empirical mean satisfies the following guarantee for every $\delta > 0$

$$\Pr\left[|\hat{\mu} - \mu| \geq \sqrt{\frac{2\sigma^2}{n} \cdot \log(2/\delta)}\right] \leq \delta$$

Stating the guarantee in words, the estimate *mîu* has *accuracy* $r = \sqrt{\frac{2\sigma^2}{n} \cdot \log(2/\delta)}$ with probability $1 - \delta$. The parameter δ is the confidence of the estimate.

For Gaussian distributions, the empirical mean is known to achieve the optimal trade-off between accuracy and confidence. We will treate the the above trade-off between accuracy and confidence as the *gold standard*.

It is natural to ask whether this *gold-standard* can be achieved for more general families of distributions $\mathcal{D}$. Indeed if the moments of the distribution $\mathcal{D}$ are appropriately bounded, one can achieve the same guarantees up to constant factors. For example, if the distribution $\mathcal{D}$ is L-subGaussian in that it satisfies,

$$\mathbb{E}[e^{\lambda(X - \mathbb{E}[x])}] \leq e^{\sigma^2\lambda^2/L} \qquad \forall \lambda > 0$$

for a fixed $L > 0$ and all $\lambda > 0$, then we can conclude that with probability $1 - \delta$, the accuracy of empirical mean is $L \cdot \sqrt{\frac{2\sigma^2}{n} \cdot \log(2/\delta)}$.

In many applications, the input distribution $\mathcal{D}$ can be "heavy-tailed", in that its higher moments need not be nicely bounded. For the sake of concreteness, consider the following setting. The distribution $\mathcal{D}$ has bounded mean μ and variance σ^2, but one may assume no bounds on its higher moments. [1]

Many real world distributions satisfy power laws and are often heavy-tailed. Can we still achieve the gold-standard in accuracy vs confidence, for all distributions $\mathcal{D}$ with bounded mean and variance?

[1] For example, consider the family of distributions $\mathcal{D}_\epsilon$,

$$x = \begin{cases} 0 & \text{w. prob. } 1-\epsilon^2 \\ 1/\epsilon & \text{w. prob. } \epsilon^2 \end{cases}$$

The distributions have bounded mean and variance but its third or higher moments $\to \infty$ as $\epsilon \to 0$.

Median of Means Estimator

Input: Samples $x_1, \ldots, x_n \sim \mathcal{D}$ and a parameter $\delta > 0$.

1. Fix $k = \lceil 100 \log(1/\delta) \rceil$.
2. Partition the samples $\{x_1 \ldots, x_n\}$ into k buckets $B_1, B_2, \ldots, B_k$ each with n/k samples.
3. Set Z_i to be the mean of bucket B_i,
$$Z_i = \frac{1}{n/k} \sum_{x_j \in B_i} x_j$$
4. Output the median of the bucket means, $\hat{\mu} = Median(Z_1, \ldots, Z_k)$

Theorem 4. *The estimate $\hat{\mu}$ given by the median-of-means algorithm satisfies,*

$$\Pr\left[|\hat{\mu} - \mu| \geq 4\sigma \cdot \sqrt{\frac{k}{n}}\right] \leq e^{-\Omega(k)}$$

Proof. We will show the claim for the upper-tail of the distribution of $\hat{\mu}$, the other one follows by a similar argument. More precisely, let us show that,

$$\Pr\left[\hat{\mu} \geq \mu + 4\sigma \cdot \sqrt{\frac{k}{n}}\right] \leq e^{-\Omega(k)}$$

Observe that for each i, Z_i is average of n/k samples from $\mathcal{D}$,

$$\mathbb{E}[Z_i] = \mu$$

and

$$\mathrm{Var}[Z_i] = \mathbb{E}[(Z_i - \mu)^2] = \frac{1}{n/k} \cdot \mathrm{Var}[\mathcal{D}]$$

Hence by using Chebyshev's inequality on Z_i, we get that

$$\Pr\left[Z_i \geq \mu + 4\sigma \cdot \sqrt{\frac{k}{n}}\right] \leq \frac{1}{16}$$

Set $\Theta = \mu + 4\sigma \cdot \sqrt{\frac{k}{n}}$. For the median of $\{Z_1, \ldots, Z_k\}$ to be larger than a threshold Θ, we need at least half of $\{Z_1, \ldots, Z_k\}$ to be larger than Θ. Since each Z_i is independent, Hence,

$$\Pr\left[\hat{\mu} > \Theta\right] \leq \Pr\left[\sum_{i=1}^{k} \mathbb{1}[Z_i \geq \Theta] \geq \frac{k}{2}\right]$$

As each $\mathbb{1}[Z_i \geq \Theta]$ is an independent Bernoulli random variable with mean $< 1/16$, by Chernoff bound, we conclude that

$$\Pr\left[\hat{\mu} > \Theta\right] \leq e^{-\Omega(k)}$$

□

Median of means estimator is a very useful primitive for streaming and randomized algorithms. Suppose we had a randomized (or streaming) algorithm $\mathcal{A}$ to estimate a quantity μ. The idea is to execute the algorithm $\mathcal{A}$, n times independently, and use the median of means estimator on the outputs. In this case, the distribution $\mathcal{D}$ is the distribution of outputs of the randomized algorithm.

Lecture 2

Heavy-tailed mean estimation on $\mathbb{R}^d$

We now consider samples from a distribution $\mathcal{D}$ over $\mathbb{R}^d$ with finite first two moments:

$$\mu = \mathbb{E}_{X\sim\mathcal{D}}[X]$$
$$\mathrm{Cov}(X) = \mathrm{Id.},$$

but possibly unbounded higher moments. As before, we wish to estimate the mean μ: $\hat{\mu}$.

For normally distributed variables - $X \sim \mathcal{N}(\mu, \mathrm{Id.})$ - we obtain the multivariate version of the *gold standard* for the usual sample mean estimator:

$$\Pr\left[\|\hat{\mu} - \mu\| \geq r^*\right] \leq \delta$$
$$r^* = \mathcal{O}\left(\sqrt{\frac{d}{n}} + \sqrt{\frac{log(1/\delta)}{n}}\right)$$

At a higher level, our goals are to get familiar with the following aspects of the problem:

- *Problem setting:* heavy tails / robust estimation
- *Algorithmic tools:* Ellipsoid algorithm, Spectral methods
- *Statistical concepts:* the Rademacher complexity

Definition 1. A point x in $\mathbb{R}^d$ is a $(r, 0.1)$**-central point** for a set of points $Z_1 \ldots Z_K \in \mathbb{R}^d$ if for every direction $u \in \mathbb{R}^d$ at most a 0.1 fraction of the set are further than r away:

$$\mathrm{Far}_u(x) = \{Z_i : |(Z_i - x) \cdot u| > r\}$$
$$\|\mathrm{Far}_u(x)\| < 0.1K,\ \forall u \implies x \text{ is a } (r, 0.1)\text{-central point}$$

The Lugosi-Mendelson estimator

Input: Samples $x_1, \ldots, x_n \sim \mathcal{D}$ and a parameter $\delta > 0$.

1. Split $X_1 \ldots X_n$ into K buckets, where $K = \log(1/\delta)$
2. Compute the means of each bucket $Z_i \in \mathbb{R}^d$
3. Find the estimator $\hat{\mu}$: a $(r^*, 0.1)$-central point for $Z_i \ldots Z_K$

We now, without loss of generality, assume the true mean μ to be zero. We wish to show that $\hat{\mu}$ is close to this mean. First, we wish to show:

Theorem 5. *The true mean μ is a $(r, 0, 1)$-central point for a set $Z_1 \ldots Z_K$ with probability of at least $1 - \delta$:*

$$\Pr\left[\|Far_u(\mu)\| \geq 0.1K\right] < \delta, \quad \forall u$$

$$r = \sqrt{\frac{d}{n}} + \sqrt{\frac{log(1/\delta)}{n}}$$

In proving the above theorem, we will look at the *worst direction* of u: $\max_u \ \|\text{Far}_u\| \in [0, ..., K] = \phi(Z_1, ..., Z_K, u)$.

We will make use of two lemmas:

Lemma 6.

$$\mathbb{E}_{Z_1,\ldots,Z_K}\left[\max_u \ \|Far_u\|\right] = 0.001K$$

Lemma 7.

$$\Pr\left[max_u\|Far_u\| - \mathbb{E}\left[\max_u \ \|Far_u\|\right] \geq 0.001K\right] \leq e^{-\Omega(K)}$$

Let us prove Lemma 7 first.

Proof. $\max_u \|\text{Far}_u\|$ is obviously 1-Lipschitz since it's a simple count. We can, therefore, use McDiarmid's inequality (Lemma 3), which immediately gives us the desired form. □

Moving on to the first lemma:

Proof.

$$\mathbb{E}_{Z_1,\ldots,Z_K}\left[\max_u \ \|\text{Far}_u\|\right] = \mathbb{E}_{Z_1,\ldots,Z_K}\left[\max_u \sum_{i=1}^{K} \mathbb{1}\{Z_i \in \text{Far}_u\}\right]$$

$$\leq \mathbb{E}_{Z_1,\ldots,Z_K}\left[\max_u \sum_{i=1}^{K} \frac{|Z_i \cdot u|}{r}\right]$$

The following steps come up quite often in Statistics and are known as *symmetrization*:

$$= \frac{1}{r}\,\mathbb{E}_{Z_i}\left[\max_u \left\{\sum_{i=1}^{K}\left(|Z_i\cdot u| - \mathbb{E}_{Z_i'}\left[|Z_i'\cdot u|\right]\right) + \sum_{i=1}^{K}\mathbb{E}_{Z_i'}\left[|Z_i'\cdot u|\right]\right\}\right]$$

$$\leq \frac{1}{r}\,\mathbb{E}_{Z_i}\left[\max_u \sum_{i=1}^{K}\left(|Z_i\cdot u| - \mathbb{E}_{Z_i'}\left[|Z_i'\cdot u|\right]\right)\right] + \frac{1}{r}\,\sum_{i=1}^{K}\max_u\,\mathbb{E}_{Z_i'}\left[|Z_i'\cdot u|\right]$$

$$\leq \frac{1}{r}\,\mathbb{E}_{Z_i,Z_i'}\left[\max_u \sum_{i=1}^{K}\left(|Z_i\cdot u| - |Z_i'\cdot u|\right)\right] + \frac{1}{r}\,K\max_u\,\mathbb{E}_{Z_i'}\left[|Z_i'\cdot u|\right]$$

Using random signs - $\sigma_i \in \{\pm 1\}^K$ - we can set a further upper bound:

$$\leq \frac{1}{r}\,\mathbb{E}_{Z_i,Z_i'}\left[\mathbb{E}_{\sigma_i}\left[\max_u \sum_{i=1}^{K}\sigma_i\left(|Z_i\cdot u| - |Z_i'\cdot u|\right)\right]\right] + \frac{K}{r}\,\max_u\,\mathbb{E}_{Z_i'}\left[|Z_i'\cdot u|\right]$$

Since Z_i and Z_i' are copies of an identical random variable:

$$\leq \frac{2}{r}\,\mathbb{E}_{Z_i}\left[\mathbb{E}_{\sigma_i}\left[\max_u \sum_{i=1}^{K}\sigma_i\left(Z_i\cdot u\right)\right]\right] + \frac{K}{r}\,\max_u\,\mathbb{E}_{Z_i'}\left[|Z_i'\cdot u|\right]$$

The quantity

$$\mathbb{E}_{Z_i}\left[\mathbb{E}_{\sigma_i}\left[\max_u \sum_{i=1}^{K}\sigma_i(Z_i\cdot u)\right]\right]$$

is also known as the **Rademacher complexity**.
Moving back to the specifics of our problem:

$$= \frac{2}{r}\,\mathbb{E}_{Z_i}\left[\mathbb{E}_{\sigma_i}\left[\underbrace{\max_u(\sum_{i=1}^{K}\sigma_i Z_i)\cdot u}_{\text{maximized when the two terms aligned}}\right]\right] + \frac{K}{r}\,\max_u\,\mathbb{E}_{Z_i'}\left[|Z_i'\cdot u|\right]$$

$$= \frac{2}{r}\,\mathbb{E}_{Z_i,\sigma_i}\left[\|\sum_{i=1}^{K}\sigma_i Z_i\|\right] + \frac{K}{r}\,\max_u\,\mathbb{E}_{Z_i'}\left[|Z_i'\cdot u|\right]$$

□

As an example of a well-known theorem where similar techniques are applied, we examine the well-known *Glivenko-Cantelli theorem:*

Theorem 8. *(Glivenko - Cantelli)*
We wish to estimate a distribution $\mathcal{D}$ over $\mathbb{R}^d$ via the histogram from a sample $Z_1, \ldots, Z_n \sim \mathcal{D}$.

The histogram estimate of the density for any interval $I = [a, b]$ *is* $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{Z_i \in I\}$.

Its deviation is bounded as:

$$\mathbb{E}_Z \left[\max_I \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{Z_i \in I\} - \mathbb{E}_{Z\sim\mathcal{D}} \left[\mathbb{1}\{Z \in I\} \right] \right\} \right] \leq \mathcal{O}\left(\sqrt{\frac{\log n}{n}} \right)$$

This is an example of more general expressions of the type

$$\mathbb{E}_{X_1,\ldots,X_n} \left[\max_{f\in\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right]$$

, where $\mathcal{F}$ is a class of functions.

Lecture 3

Ellipsoid algorithm and its applications

Let's first briefly recap where we left off last time. We want to do heavy-tailed mean estimation in $\mathbb{R}^d$ for large d. The clever median-of-means algorithm in the $d = 1$ case led us to consider bucketing algorithms, but unfortunately we have to do a little more work to generalize the notion of a median to higher dimensions. That led us to the notion of a (r, ϵ)-central point (see Definition 1). Indeed, the Lugosi-Mendelson estimator admits the following extremely succinct description.

1. Divide input data into k buckets Z_i, $i \in [k]$.

2. Output a $(r, 0.01)$–center of the Z_i.

Here $r = \theta\left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right)$ is the gold standard for confidence δ tradeoff (viz. the Gaussian tradeoff).

Last time we didn't actually finish the proof of Lemma 6. Here's the rest of the calculation for clarity.

Proof. Last time we managed to show that

$$\mathbb{E}_{Z_i}\left[\max_u \|\mathrm{Far}_u\|\right] \leq \frac{2}{r}\mathbb{E}\max_u \left|\left(\sum_{i=1}^k \sigma_i Z_i\right)\cdot u\right| + \frac{k}{r}\max_u \mathbb{E}|Z\cdot u|.$$

Let's first upper bound the first term. First, the magnitude of the dot product is evidently maximized when u is aligned with $\sum_i \sigma_i Z_i$, in which case the dot product becomes $\|\sum_i \sigma_i Z_i\|_2$. In general, when we see 2-norms floating around it behooves us to try to apply Cauchy-Schwarz. Following our nose here and churning out the algebra

yields

$$\begin{aligned}\mathbb{E}_{Z,\sigma}\left\|\sum_i \sigma_i Z_i\right\|_2 &\leq \left(\mathbb{E}_{Z,\sigma}\left\|\sum_i \sigma_i Z_i\right\|_2^2\right)^{1/2} \\ &= \mathbb{E}_{Z,\sigma}\sum_{i,j}\sigma_i\sigma_j(Z_i \cdot Z_j) \\ &= \sum_i \mathbb{E}_{Z_i}\|Z_i\|^2.\end{aligned}$$

In the last line we use the fact that $\mathbb{E}\sigma_i\sigma_j = \delta_{ij}$, where δ is the Kronecker delta function (i.e. 1 when $i = j$, 0 otherwise).

From here it is just a standard calculation, but Prasad thinks it's worth working it out the first time you see it. So for completeness' sake, we'll include it here. Recall that $Z_i = \frac{1}{n/k}\sum_{j\in B_i} X_j$. So computing out the expectation we get

$$\frac{k^2}{n^2}\sum_{i,j}\mathbb{E}X_iX_j.$$

The cross terms vanish because the X_i are i.i.d. and mean zero. On the other hand because $\mathrm{Cov}(\mathcal{D}) = I$, $\mathbb{E}[\|X\|^2] = d$ (we can be even more thorough and compute this explicitly by comparing to the covariance matrix). Returning back to the original term we find that

$$\mathbb{E}\max_u\left|\left(\sum_{i=1}^k \sigma_i Z_i\right)\cdot u\right| \leq k\sqrt{\frac{d}{n}}.$$

Now let's look at the other term. Again we want to use Cauchy-Schwarz because of the pesky absolute values; we have

$$\begin{aligned}\max_u \mathbb{E}|Z\cdot u| &\leq \max_u\left(\mathbb{E}(Z\cdot u)^2\right)^{1/2} \\ &= \max_u(u^\top\mathbb{E}[ZZ^\top]u)^{1/2} \\ &= \left\|\frac{1}{n/k}\mathrm{Cov}(\mathcal{D})\right\|_{\mathrm{op}}^{1/2} \\ &= \sqrt{\frac{k}{n}}.\end{aligned}$$

In the second line we have just expanded out the square and pulled the expectation through the u's (which are deterministic). In the remaining lines we have used the definition of the operator norm and the fact that $\mathrm{Cov}(\mathcal{D}) = I$.

Putting it together, recalling that $k \triangleq \log(1/\delta)$, we have

$$\begin{aligned}\mathbb{E}_{Z_i}\left[\max_u\|\mathrm{Far}_u\|\right] &\leq \frac{2}{r}\cdot k\sqrt{\frac{d}{n}} + \frac{k}{r}\cdot\sqrt{\frac{k}{n}} \\ &= \frac{1}{r}(k\cdot\Theta(r)) \\ &= \Theta(k).\end{aligned}$$

□

Remark. Here are a couple takeaways from this algebra.

1. Absolute value of 2-norm should ring a bell in our head to try to use Cauchy-Schwarz. It passes things to somethign more algebraic.
2. Cross terms vanish because of centered assumption and independence.
3. Whenever we see this $\max_u \mathbb{E}[(Z \cdot u)^2]$ we should think of the operator norm of the covariance matrix.

We also have the following elementary property about central points.

Lemma 9. *Every point x which is $(r, 0.01)$-central relative to $Z_1, \ldots, Z_k$ is also close to the mean. In particular it satisfies $\|x - \mu\| = O(r)$.*

Proof. Prasad chose to do a proof by picture. In lieu of that I'll try to describe it in (at most) a thousand words. Simply pick the direction u to be the one joining x and μ. Since 99% of the buckets are r close to x and μ, by Pigeonhole principle there's some point Z_i which is r close to both in the direction of u. So that means that the distance between x and μ is at least $2r$! We can see that this same argument would work for any $(r, \frac{1}{2} + \epsilon)$ center. □

Computing the Lugosi-Mendelson estimator

Now that we've shown that the true mean is a $(r, 0.01)$ central point for the buckets, we need to actually figure out how to compute such a central point. Obviously this might not be the actual μ, otherwise we'd be done already!

The algorithm that we're going to introduce is slow but the main point is just to introduce the so-called *ellipsoid algorithm*. To that end, let's try to design an algorithm from scratch.

First fix a direction u and project all the Z_i onto u. This gives an ordering of the projected Z_i, and based on the top 1% and bottom 1% of the points we get some bounding lines. In essence this gives bounding lines θ_u and θ'_u, a valid central point would have to lie within this slab. To that end, we define the notation

$$\text{Slab}_u \triangleq \{x | x \cdot u \in [\theta_u, \theta'_u]\}.$$

It isn't so important what θ_u and θ'_u exactly are how to compute them, so we won't dwell on this point. Now we need to find a point in

$$\mathcal{K} \triangleq \bigcap_{u \in S^{d-1}} \text{Slab}_u.$$

This set $\mathcal{K}$ is convex, as an intersection of convex sets. This is where the ellipsoid algorithm comes in. Before we introduce the ellipsoid algorithm, we need to define the notion of a *separation oracle*.

Definition 2. Given a convex set $\mathcal{K}$, a separation oracle is a black box algorithm that does the following:

Given a point x as input:

1. if $x \in \mathcal{K}$, output YES.
2. if $x \notin \mathcal{K}$, output NO, as well as a halfspace h such that $h(x) < 0$ but $h(y) > 0$ for all $y \in \mathcal{K}$.

With that out of the way, let's explain how the ellipsoid algorithm works.

Ellipsoid algorithm

Input: Positive real numbers $r < R$ such that there is some x_0 with $B(x_0, r) \subseteq \mathcal{K} \subseteq B(0, R)$ and a separation oracle for $\mathcal{K}$. Note that we don't need to know what x_0 is.

1. First initialize the ellipsoid $E_1 = B(0, R)$ and the center $C_1 = 0$.
2. Query if $C_1 \in \mathcal{K}$. If yes, we're done. Otherwise, the oracle gives a halfspace h_1 which separate C_1 from $\mathcal{K}$.
3. Now construct the next ellipsoide E_2 which contains h_1 as well as $\mathcal{K}$ and set C_2 to be the center of E_2.
4. Repeat steps 2-3 until we get some center which is in $\mathcal{K}$.

At first, it might be unclear why this algorithm should actually terminate. But the point is that by using the separating hyperplane we're essentially binary searching. In fact, the volume of the ellipsoid must shrink by a factor of at most $1 - \frac{1}{d}$. Since we maintain the invariant that our ellipsoid always contains $\mathcal{K}$, the algorithm terminates whenever we reach a volume smaller than $\text{Vol}(B(0, r))$. This takes $O(\log R/r)$ steps.

The following theorem is actually true, although we didn't prove it in lecture.

Theorem 10. *The ellipsoid algorithm finds $x \in \mathcal{K}$ using $O(d^2 \log(R/r))$ oracle calls.*

Applying the ellipsoid algorithm

The power of the ellipsoid algorithm is not necessarily in its runtime, but rather that it is a very powerful and flexible hammer. If one wishes to study some computational problem, it is a good idea to try to use the ellipsoid algorithm to show that it is tractable. Assuming that it can be solved with the ellipsoid algorithm, with a bit more ingenuity, there is some hope of finding more efficient algorithms.

To embody that ethos, let's try to use the ellipsoid algorithm to compute a $(r, 0.01)$ center. What we need to do is to make sure we can provide the required inputs to the ellipsoid algorithm, namely a bound on r, R, and a separation oracle. For R we handwave it away by appealing to machine precision. For r we know that μ is $(r^*, 0.01)$ central but we don't care about constants so we can just double r^* and this gives us wiggle room around μ. The major loose strand is the separation oracle.

Naively, one could try to do an ϵ-net argument over $O(2^d)$ directions to try to upper bound this $\max_u |\text{Far}(u)|$. Obviously we should try to do avoid doing this (and also it's messy). A better idea is to try to exhaustively search over the possible values of $\text{Far}(u)$. Using this idea with the ellipsoid algorithm will give us a $\text{poly}(d, 1/\delta)$ runtime algorithm, which is secretly a $\text{poly}(d, 2^k)$ runtime algorithm, which is still not great. [2]

The idea here is that if some candidate z is not $(r, 0.01)$-central, then there is some $u \in \mathbb{R}^d$ with $z \notin \text{Slab}_u$. We can enumerate over all subsets of buckets, and let L_u and R_u be the set of failures (the top and bottom most 1% when we project onto u). We can try all $(2^k)^2$ possiblities for (L_u, R_u) and try to infer what u is.

Say $L_u = \{Z_{i_1}, \ldots, Z_{i_\ell}\}$. We know that z is r-far from L_u along some u. We can thus construct the following quadratic program. We need to find u of unit norm (hence quadratic) such that for each $Z_j \in L_u$, we have $(Z_j - z) \cdot u \geq r$. This can be solved in $\text{poly}(d, 2^k)$ time. Or we could recurse and use ellipsoid again, but this was mostly just a joke.

Anyway, we can do the same for R_u and so solving these QPs for every possible combination of L_u and R_u will give us a feasible u. Note that this is guaranteed to work because we know that $\mathcal{K}$ is nonempty ($\mu \in \mathcal{K}$).

[2] Someone asked in class why ϵ-nets don't work but Rademacher magically does. Prasad explained this by appealing to the following thought experiment. Let's suppose we took $\mathbb{E} \max_{u \in U} f(u)$ where U has size 2^d. These can be arbitrary so basically anything can happen, and we really do need a union bound argument. But when we maximize over the function class $\mathcal{F} = \{\phi_u(z) = |Z \cdot u| : u \in S^{d-1}\}$, even though $\mathcal{F}$ is infinite, its complexity measure (according to Rademacher) is small! So we avoid using union bound altogether, which as a rule of thumb is probably good when you have an uncountably large set...

Lecture 4

Robust Mean Estimation

Introduction

In this class, we continue our work on estimating the mean of a distribution, but we would like our estimates to be *robust*. By robust, we mean that the samples we obtain from our distribution $\mathcal{D}$ may be corrupted. Our goal is to estimate the mean of the true distribution, and not of the mixed one.

Let's first see some contamination models that capture how our samples may get corrupted.

Huber Contamination Model. A $(1-\epsilon)$ fraction of our samples is obtained from $\mathcal{D}$. The remaining ϵ fraction of our data is sampled from a 'corrupted' distribution $\mathcal{N}$.

Our goal is to estimate

$$\mu(\mathcal{D}) = \mathbb{E}_{X\sim\mathcal{D}}[X]$$

Strong Corruption Model. We first obtain n samples $X_1, \ldots, X_n$ from the original distribution $\mathcal{D}$. The adversary is allowed to look at the samples and corrupt at most ϵn of them. That is, they are allowed to remove up to ϵn of the X_i's and replace them with their own.

The final set of samples (including the corrupted ones) is $\{\tilde{X}_1, \ldots, \tilde{X}_n\}$. Our goal is the same, given $\{\tilde{X}_i\}$, can we estimate $\mu(\mathcal{D})$?

Additional Motivation for Robust Statistics. We have been thinking of corruptions as error introduced by an adversary, but corrupted data may also appear due to model deviation. For example, while studying a real-life phenomenon we may make the assumption that the underlying distribution behaves like a Gaussian. Even though this may not be the case, using a robust estimator will allow us to estimate the properties we want, as long as the true distribution is close to the Gaussian.

Robust Properties of Distributions

Definition 3 ((Δ, ϵ)-Resilience)**.** A distribution $\mathcal{D}$ on $\mathbb{R}^d$ is (Δ, ϵ)-resilient if deleting an ϵ-fraction [3] of $\mathcal{D}$ does not change $\mu(\mathcal{D})$ by more than Δ.

[3] An intuitive way to think about deleting an ϵ-fraction of the distribution is to consider the PDF of $\mathcal{D}$. An adversary is allowed to remove at most ϵ mass from the PDF. Note that what we get is not exactly a PDF, since it is not normalized.

It is easy to see that resilience is a desirable property, otherwise the adversary could corrupt only an ϵ-fraction of our distribution and significantly affect the mean.

An alternative definition of (Δ, ϵ)-resilience is that $\forall$ events E with $\mathbb{P}[E] \geq 1 - \epsilon$

$$\left\| \mathbb{E}_{x \sim \mathcal{D}}[x \mid E] - \mathbb{E}_{x \sim \mathcal{D}}[x] \right\|_2 \leq \Delta$$

This definition is actually very useful; it will give us a strategy to do robust mean estimation.

Robust Mean Estimation Algorithms

Strategy Overview

Let the original set of samples before corruption be S, and assume that S is $(\Delta, 10\epsilon)$-resilient [4].

[4] If a distribution $\mathcal{D}$ is (Δ, ϵ)-resilient, then its empirical distribution (for a sufficiently large number of samples) is also close to (Δ, ϵ)-resilient.

The adversary removes ϵ-fraction of the samples and adds its own ϵ-fraction of 'bad' samples to obtain $\tilde{S} = S_{good} \cup S_{bad}$. Here $S_{good} \subseteq S$ is the subset of original samples that were not removed by the adversary.

One natural direction is to identify the 'good' samples of $\tilde{S}$ and take their mean. But how can we identify them?

Claim 11. *S_{good} is a $(2\Delta, \epsilon)$-resilient set*

The intuition for the above claim is that S_{good} and S share at least a $(1 - \epsilon)$-fraction of their samples. Thus, if S_{good} were not $(2\Delta, \epsilon)$-resilient then an adversary could remove an ϵ-fraction of samples from S to obtain S_{good} and then remove an additional ϵ-fraction from S_{good} to move the mean 2Δ away from $\mu(\mathcal{D})$. This would imply that S is not $(2\Delta, 2\epsilon)$-resilient, which is a contradiction, as we assumed that S is $(\Delta, 10\epsilon)$-resilient.

Goal. Find a $(2\Delta, \epsilon)$-resilient set $T \subseteq \tilde{S}$ of size at least $(1 - \epsilon)n$. T may include samples from S_{bad}.

Claim 12. *The mean of T is close to the mean of S, i.e.*

$$\left\| \mu(T) - \mu(S) \right\|_2 \leq 3\Delta$$

Proof. Consider the intersection of $S \cap T$.

We can obtain $S \cap T$ by deleting a $O(\epsilon)$-fraction of the samples from S. This means that their means are close since S is resilient

$$\left\| \mu(S) - \mu(S \cap T) \right\|_2 \leq \Delta$$

Similarly, we can obtain $S \cap T$ by deleting a $O(\epsilon)$-fraction of samples from T. This again means that their means are close since T is resilient

$$\left\|\mu(T) - \mu(S \cap T)\right\|_2 \leq 2\Delta$$

Triangle inequality completes the proof

$$\left\|\mu(T) - \mu(S)\right\|_2 \leq 2\Delta + \Delta = 3\Delta$$

□

Verifying Resilience

The current definition of resilience makes it intractable to verify. Turns out that a bound on the variance of the distribution implies resilience.

Theorem 13. *Suppose $\mathcal{D}$ on $\mathbb{R}$ has $Var[\mathcal{D}] = 1$. Then $\mathcal{D}$ is $\left(\frac{\sqrt{\epsilon}}{1-\epsilon}, \epsilon\right)$-resilient $\forall \epsilon \geq 0$*

Proof. Intuition. Say we are the adversary and we want to corrupt a small number of points to affect the mean a lot. If we introduce a new point that is $O(c)$-far from the mean, its contribution to the variance is $O(c^2)$. Thus having a bound on the variance limits the control of the adversary on the mean.

Formally, pick an event E such that $\mathbb{P}[E] = 1 - \epsilon$. We want to prove that

$$\left| \underbrace{\mathbb{E}[X \mid E]}_{\mu_E} - \mathbb{E}[X] \right| \leq \frac{\sqrt{\epsilon}}{1-\epsilon}$$

Rewrite the true mean as

$$\begin{aligned}
\mu &= \mathbb{P}[E] \cdot \mu_E + \mathbb{P}[E^c] \cdot \mu_{E^c} \\
&= (1-\epsilon) \cdot \mu_E + \epsilon \cdot \mu_{E^c} \\
\Rightarrow \mu(E) - \mu &= \frac{\epsilon}{1-\epsilon} \cdot (\mu - \mu_{E^c}) \\
&= \frac{1}{1-\epsilon} \cdot \mathbb{E}[(X-\mu)\mathbb{1}[E^c]] \\
&\leq \frac{1}{1-\epsilon} \cdot \left(\underbrace{\mathbb{E}[(X-\mu)^2]}_{Var[\mathcal{D}]=1} \right)^{1/2} \cdot \left(\underbrace{\mathbb{E}[\mathbb{1}[E^c]]}_{=\epsilon} \right)^{1/2} \\
&\leq \frac{\sqrt{\epsilon}}{1-\epsilon}
\end{aligned}$$

Where the last inequality is due to Cauchy-Schwarz. □

Question. Is μ_{E^c} even finite?
Write

$$\begin{aligned}\mu_{E^c} &= \mathbb{E}[x \mid E^c] \\ &= \frac{\mathbb{E}[X \cdot \mathbb{1}[E^c]]}{\mathbb{P}[E^c]} \\ &\leq \frac{\left(\mathbb{E}[X^2]\right)^{1/2} \cdot \left(\mathbb{E}[\mathbb{1}[E^c]]\right)^{1/2}}{\epsilon}\end{aligned}$$

Resilience in Higher Dimensions

Theorem 14. *Suppose $\mathcal{D}$ on $\mathbb{R}^d$ has $Cov[\mathcal{D}] \preceq I$ [5]. Then $\mathcal{D}$ is $\left(\frac{\sqrt{\epsilon}}{1-\epsilon}, \epsilon\right)$-resilient $\forall \epsilon \geq 0$*

[5] For two matrices A, B, we say that $A \preceq B$ if $\lambda_{min}(B - A) \geq 0$. If $B = \sigma I$, then $A \preceq B \Leftrightarrow \lambda_{max}(A) \leq \sigma$

Proof. We want to bound the ℓ_2 norm

$$\|\mu(E) - \mu\|_2 = \max_{\|u\|=1} \langle u, \mu(E) - \mu \rangle$$

Note that we just reduced a high-dimensional problem to a single dimension. Use the formula from before

$$\begin{aligned}\|\mu(E) - \mu\|_2 &= \max_{\|u\|=1} \left\langle u, \frac{1}{1-\epsilon} \cdot \mathbb{E}[(X - \mu)\mathbb{1}[E^c]] \right\rangle \\ &= \max_{\|u\|=1} \frac{\mathbb{E}[\langle (X - \mu), u \rangle \cdot \mathbb{1}[E^c]]}{1-\epsilon} \\ &\leq \frac{\left(\max_{\|u\|=1} \mathbb{E}\left[\langle x - \mu, u \rangle^2\right]\right)^{1/2} \cdot \left(\mathbb{E}[\mathbb{1}[E]]\right)^{1/2}}{1-\epsilon} \\ &\leq \frac{\sqrt{\epsilon}}{1-\epsilon}\end{aligned}$$

□

Now it remains to find such subset efficiently

Problem. Given $X_1, \ldots, X_n \in \mathbb{R}^d$. Find $|T| \geq (1-\epsilon)n$ points such that $Cov[T] \preceq 2I$.

First Idea. Let's think of the situation in 1 dimension. Start with the empirical mean of the entire set. Remove points from the far left and right as an effort to bound the variance. Recompute the mean of the remaining points and repeat.

Since the mean changes with each iteration, this operation is not linear and it is hard to argue that the procedure will stop before eliminating every point. In particular, the procedure may eliminate more good points than bad points in each iteration.

Define $w_i = \frac{1}{|T|}$ if $x_i \in T$ and 0 otherwise. Then $\mu(T) = \sum_i w_i X_i$. The variance is then equal to

$$\sum_i w_i (X_i - \mu(T))^2 = \sum_i w_i (X_i - \sum_j w_j X_j)^2$$

This quantity that we are optimizing over the w_i's is not convex.

Filter Algorithm

We will try to make this first idea work. One obstacle was that removing a point is maybe a bit too harsh. We can instead give each point a weight c_i. Whenever a point is far from the mean, instead of deleting it, we will discount its weight [6].

[6] Note that there may be some dependency on ϵ missing in the definition of the algorithm

---

Input: $x_1, \ldots, x_n \in \mathbb{R}^d$
Initialize: $c_1 = \cdots = c_n = 1$

1. Compute the empirical mean
$$\mu_c = \frac{\sum_i c_i X_i}{\sum_i c_i}$$
2. Compute the empirical covariance
$$\Sigma_c = \frac{\sum c_i (X_i - \mu_c)(X_i - \mu_c)^T}{\sum_i c_i}$$
3. If $\|\Sigma_c\|_{op} \leq 1000$, terminate. We have found a subset with bounded covariance and thus it is resilient.
4. Compute the top eigenvector v of Σ_c.
5. Define
$$\tau_i = \langle X_i - \mu_c, v \rangle^2$$
6. Set
$$c_i \leftarrow c_i \left(1 - \frac{\tau_i}{\tau_{max}}\right)$$
7. Repeat

Note that all of the weights decrease at each iteration.

Claim 15. *At all steps*
$$\sum_{i \in S_{good}} (1 - c_i) \leq \sum_{i \in S_{bad}} (1 - c_i)$$

We know there are at most ϵn points. So the RHS is at most ϵn. The weight of the good points starts off as $(1 - \epsilon n)$. From the RHS, we know that the weight of the good points can lose at most another ϵn, so the total weight of the good points is at least $1 - 2\epsilon n$

This implies that the total variation distance between the uniform distribution over S_{good} and the distribution of weights C is small

$$TV(S_{good}, C) \leq O(\epsilon)$$

Therefore when the algorithm terminates we have a distribution C with bounded covariance matrix, which implies that the distribution C is $(O(\sqrt{\epsilon}), \epsilon)$-resilient.

We conclude that $\|\mu_C - \mu(S_{good})\|_2 \leq O(\sqrt{\epsilon})$

Remarks about Robust Mean Estimation

Robust Mean Estimation can be used as a primitive to do other things. One such example is Linear Regression. We would like to fit a line through a set of sample points $\{x_i, y_i\}$. This can be expressed as the following optimization problem

$$\min_{\ell \in \mathcal{F}} \sum_i loss(\ell(x_i) - y_i)$$

Even in Linear Regression, we could imagine the same statistical problem occurring: an ϵ-fraction of our sample points is corrupted.

Before we get to the robust case, how do we solve the non-robust one? One way is to use gradient descent, by expressing the gradient of the problem as

$$\sum_i \nabla_\ell loss(\ell(x_i) - y_i)$$

Note that the gradient is nothing more than the *average* of the gradient of the loss over the samples.

So in the robust case, when some samples are corrupted, robustly estimating the average could be used to get an approximation of the gradient!

Lecture 5

Sum-of-squares semidefinite programming

Claim 16. *Let A be a symmetric matrix in $\mathbb{R}^{d\times d}$. The following are equivalent:*

1. *All eigenvalues of A are ≥ 0*
2. *$v^\intercal Av \geq 0$ for all $v \neq 0$ in $\mathbb{R}^d$*
3. *There exists a matrix B such that $A = BB^\intercal$ (Cholesky decomposition)*

Definition 4. A *positive semidefinite* (PSD) matrix A is a symmetric matrix that satisfies any of the above conditions. This is denoted $A \succeq 0$.

A consequence of the second part of claim 16 is:

Claim 17. *The set of all PSD matrices is a convex set in $\mathbb{R}^{d\times d}$.*

Definition 5. Let $A = (A_{ij})$ and $B = (B_{ij})$ be two symmetric matrices of the same dimension. The Frobenius or Hilbert-Schmidt inner product is

$$\langle A, B\rangle = \sum_{ij} A_{ij}B_{ij}.$$

Definition 6. Let C, X and $(A_i)_{k=1}^m$ be $d \times d$ matrices. A *semidefinite program* (SDP) is an optimisation problem of the form

$$\begin{aligned} &\text{minimise } C, X \\ &\text{subject to } X \succeq 0 \\ &\qquad A_i, X \geq b_i \text{ for } i = 1, \ldots, m. \end{aligned}$$

Suppose X is not PSD. Then $\lambda_{\min}(X) < 0$, so there exists a $v \in \mathbb{R}^d$ with $v^\intercal Xv < 0$. Define a separating hyperplane $H(Z) = v^\intercal Zv$. Note that $H(X) < 0$ and $H(B) \geq 0$ for all PSD matrices B.

We can solve SDPs efficiently using a separation oracle: given $X \in \mathbb{R}^{d\times d}$,

- Check if $\langle A_i, X \rangle \geq b_i$ for all $i = 1, \ldots, m$.
- Compute $\lambda_{\min}(X)$ and check if it is < 0. Then take the separating hyperplane defined by smallest eigenvector v: $H(Z) = v^\intercal Z v$.

Using this separation oracle, we can apply the ellipsoid algorithm. Note that any minimsation problem can be converted to a feasibility problem by setting $\texttt{objective} \leq \theta$ and performing binary search on θ.

Remark. LPs can be solved exactly (if the inputs are rational then so is the solution) but SDPs cannot (the solution may not even be algebraic).

Introducing sum-of-squares

Consider the following optimisation problem:

$$\min_{x \in \mathbb{R}} p(x) = p_0 + p_1 x + p_2 x^2 + \cdots + p_d x^d.$$

This function is not convex. How can we convexify it? A general principle is as follows:

$$\text{a set} \longrightarrow \text{probability distributions on the set}$$

Now consider minimising $\mathbb{E}_{X \sim \mathcal{D}}[p(X)]$ over probability distributions $\mathcal{D}$ on $\mathbb{R}$. Note that

$$\mathbb{E}_{X \sim \mathcal{D}}[p(X)] = \sum_{i=0}^{d} p_i \cdot \underbrace{\mathbb{E}_{X \sim \mathcal{D}}[X^i]}_{M_i}$$

so we only care about the moments of the distribution. So our optimisation problem is now

$$\begin{aligned} &\text{minimise } \sum_{i=0}^{d} p_i M_i \\ &\text{subject to } (M_0, \ldots, M_d) \in \mathcal{M}^{d+1} \end{aligned}$$

where $\mathcal{M}^{d+1}$ is the set of all tuples $(M_0, \ldots, M_d)$ that are the moments of some probability distribution $\mathcal{D}$ on $\mathbb{R}$. Clearly this is a convex subset of $\mathbb{R}^{d+1}$ (think about the moments of mixture distributions). Also $M_0 = 1$. What other constraints on $\mathcal{M}^{d+1}$ are there? Here is a simple one: $M_2 = \mathbb{E}_{X \sim \mathcal{D}}[X^2] \geq 0$ for all $\mathcal{D}$. More generally, $\mathbb{R}_{X \sim \mathcal{D}}[q^2(X)] \geq 0$ for all $\mathcal{D}$ and for all polynomials q. So our optimisation problem is now

$$\begin{aligned} &\text{minimise } \sum_{i=0}^{d} p_i M_i \\ &\text{subject to } \sum_{a,b} q_a q_b M_{a+b} \geq 0 \end{aligned}$$

for all polynomials $q(x) = \sum q_i x^i$ with $\deg(q) \geq d/2$ and $M_0 = 1$. Note that there are infinitely many constraints. [7]

[7] Every convex program can be thought of as a linear program with infinitely many constraints since every convex set is the intersection of halfspaces.

We can rewrite these constraints to make the problem in to an SDP:

$$\begin{aligned} &\text{minimise } \sum_{i=0}^{d} p_i M_i \\ &\text{subject to } M' \succeq 0 \end{aligned}$$

where $M' \in \mathbb{R}^{d/2 \times d/2}$ with $M'_{a,b} = M_{a+b}$ and $M_0 = 1$. Another way to write this is

$$\begin{aligned} &\text{minimise } \sum_{i=0}^{d} p_i M'_{0,i} \\ &\text{subject to } M' \succeq 0 \\ &\qquad\qquad M'_{a,b} = M'_{c,d} \text{ if } a+b = c+d \\ &\qquad\qquad M_{0,0} = 1 \end{aligned}$$

This is known as the *sum-of-squares* SDP for this problem. Does this give the right answer? To answer this question we need to introduce some more notation.

Definition 7. A *pseudo-expectation* operator $\tilde{E}$ is a linear functional

$$\begin{aligned} \tilde{E} : \mathbb{R}[x] &\to \mathbb{R} \\ r(x) &\to \sum r_i M_i \end{aligned}$$

that satisfies $\tilde{E}[1] = 1$. Here $\mathbb{R}[x]$ denotes the polynomial ring over $\mathbb{R}$. For a tuple $(M_0, \ldots, M_d)$, the associated pseudo-expectation operates on the polynomials of degree $\leq d$.

We can give yet another formulation of the SDP:

$$\begin{aligned} &\text{minimise } \tilde{E}[p(X)] \\ &\text{subject to } \tilde{E}[q^2(X)] \geq 0 \text{ for all } q(x) \text{ with } \deg(q) \leq d/2 \end{aligned}$$

where the minimisation is over linear functionals $\tilde{E}$.

Example.

1. $p(x) = r^2(x)$. Then $\tilde{E}[p(X)] = \tilde{E}[r^2(x)] \geq 0$.
2. $p(x) = \sum_i r_i^2(x)$. Then $\tilde{E}[p(X)] = \sum_i \tilde{E}[r_i^2(x)] \geq 0$.
3. $p(x) = \beta + \sum_i r_i^2(x)$. Then $\tilde{E}[p(X)] = \beta \cdot \tilde{E}[1] \sum_i \tilde{E}[r_i^2(x)] \geq \beta$.

Hence a sum-of-squares proof (in the algebraic sense) that $p(x) \geq \beta$ implies that the SDP returns a value $\geq \beta$. So this provides a dual certificate for the SDP.

Claim 18. *Let $p(x) > 0$ for all $x \in \mathbb{R}$ and $\deg(p)$ is even. Then $p(x) = \sum_i r_i^2(x)$ for some polynomials $r_i(x)$.*

So every non-negative polynomial of even degree is a sum of squares.

Proof. The zeroes of p are of the form $a_l \pm b_l i$ for $l = 1, \ldots, d/2$ where $\deg(p) = d$. Then factor p as

$$p(x) = C^2 \prod_{l=1}^{d/2} (x - (a_l + b_l i))(x - (a_l - b_l i)) = C^2 \prod_{l=1}^{d/2} \left((x - a_l)^2 + b_l^2 \right)$$

and multiply out the outer parentheses on the right hand side. □

Now suppose that $\min_{x \in \mathbb{R}} p(x) = \beta$. Define $\tilde{p}(x) = p(x) - \beta \geq 0$. Then we can write $p(x) = \beta + \sum r_j^2(x)$. We can therefore use the sum-of-squares representation to find the minimum of a univariate polynomial. Hence the SDP always gives the right answer.

So for univariate polynomials, we can prove any algebraic fact about them using sum-of-squares techniques. However for multivariate polynomials this is not the case.

Lecture 6

Lecture 7

Lecture 8

Lecture 9

Lecture 10

Lecture 11

Lecture 12

Lecture 13

Lecture 14

Lecture 15

Lecture 16

Lecture 17

Lecture 18

Lecture 19

Lecture 20

Lecture 21

Lecture 22

Lecture 23

Lecture 24

Lecture 25

Lecture 26

Lecture 27

Lecture 28

Lecture 29

Lecture 30

Bibliography