CS 7545: Machine Learning Theory **Fall 2021**

Lecture 6: Robust Estimation

Instructor: Santosh Vempala *Lecture date: 09/15/2021*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

6.1 Introduction

There are many examples of learning from data. What if not all the data is generated by the model? Consider the case when $(1-\epsilon)$ fraction of data points are from the model and ϵ fraction are arbitrary (adversarial). Is it still possible to estimate the model parameters?
Low-degree sample moments are not robust estimators. For example,

- Mean: Adding a point really far from the samples will significantly change the mean.
- Singular vectors/eigenvectors: Consider samples with e_1 as the top singular vector. Adding a point $(0, Me_2), M \gg 1$, will significantly move the top singular vector.

Consider the problem of estimating a single Gaussian. For a 1-d Gaussian distribution, $N(\mu, \sigma^2)$, the median of the sample points, $\hat{m}$ is a robust estimator of the mean. $|\mu - \hat{m}| = O(\epsilon)\sigma$ w.h.p. This is the best possible estimate in 1-d. In dimension d, what is a robust estimate for the mean such that the error does not grow with d?
Tukey Ellipsoid: Tukey ellipsoid is the minimum volume ellipsoid that contains half of the data points. The center of Tukey ellipsoid is a good estimate for the mean. However, it is NP-hard to compute.
[2016] For a large class of distributions, the mean and covariance can be estimated to within error of information theoretic bound.

$$\left\|\Sigma^{-1/2}(\bar{\mu} - \mu)\right\|_2 = O(\epsilon\sqrt{\log(1/\epsilon)}).$$

Two algorithms for robust mean estimation:

1. Iterative Filtering
2. Recursive Dimension Halving

Lemma 6.1. *For an ϵ-corrupted Gaussian, $N(\mu, I)$ with additive corruptions, if the sample covariance, $\hat{\Sigma}$ satisfies $\left\|\hat{\Sigma}\right\|_2 \le 1 + \epsilon$, then $\|\hat{\mu} - \mu\|_2 = O(\epsilon)$.*

Proof. Without loss of generality, let $\mu = 0$. Let $S = G \cup B$ where G is set of samples from the Gaussian and B is added corrupted points. With enough samples, the sample mean of G, $\mu_G \approx \mu$ and the sample variance of G, $\Sigma_G \approx \Sigma$.

Let $\mu_B = \sum_{x \in B} x/|B|$ and $\Sigma_B = \sum_{x \in B} xx^\top/|B| - \mu_B\mu_B^\top$. The sample mean $\hat{\mu}$ and the sample covariance $\hat{\Sigma}$, Then

$$\hat{\mu} = \epsilon\mu_B.$$
$$\hat{\Sigma} = (1-\epsilon)I + \epsilon\Sigma_B + (\epsilon - \epsilon^2)\mu_B\mu_B^\top. \tag{6.1}$$

For $v = \mu_B / \|\mu_B\|$,

$$1 + \epsilon \ge v^\top\hat{\Sigma}_B v \ge 1 - \epsilon + (\epsilon - \epsilon^2)\|\mu_B\|^2.$$

Therefore, $\|\mu_B\| \le 2/(1-\epsilon) = O(1)$ and $\|\hat{\mu}\| = O(\epsilon)$. □

For general noise, $\|\hat{\mu} - \mu\| = O(\epsilon\sqrt{\log(1/\epsilon)})$.

Due to Gaussian concentration, $\|x - \mu\| \leq C\sqrt{d\log(N/\tau)}$ for all sample points $x \in G$ w.h.p. So, we can remove all points x_i with $\|x_i - \mu\| \geq C\sqrt{d\log(N/\tau)}$.

Lemma 6.2. *After removing all points x with $\|x\| \geq C\sqrt{d\log(N/\tau)}$ from the sample, $\lambda_{\min}(\hat{\Sigma}) \geq 1 - \epsilon$ and $\mathrm{Tr}\hat{\Sigma} = (1 + O(\epsilon))d$.*

Proof. For any $v \in \mathbb{R}^d$ with $\|v\| = 1$, $v^\top \hat{\Sigma} v \geq (1 - \epsilon)$, therefore $\lambda_{\min}(\hat{\Sigma}) \geq (1 - \epsilon)$.

After removing points with $\|x\|_2 \geq C\sqrt{d\log(N/\tau)}$,

$$\begin{aligned}\mathrm{Tr}\hat{\Sigma} &= (1-\epsilon)d + \epsilon\left(\mathrm{Tr}\Sigma_B + (1-\epsilon)\mu_B\mu_B^\top\right)\\ &\leq (1-\epsilon)d + \epsilon\sum_{x\in B}\|x\|_2^2/|B|\\ &\leq (1-\epsilon)d + \epsilon C^2 d = (1+O(\epsilon))d.\end{aligned}$$

□

$\lambda_{\min}(\hat{\Sigma}) \geq 1 - \epsilon$ and $\mathrm{Tr}\hat{\Sigma} = (1 + O(\epsilon))d$ imply $\lambda_{d/2}(\hat{\Sigma}) = 1 + O(\epsilon)$. So, in the span of the bottom $d/2$ eigenvectors of $\hat{\Sigma}$, sample mean is a good approximation of the true mean.

Algorithm 1: Recursive Dimension Halving

Given corrupted samples S:

1. Let m = coordinate-wise median($\{x : x \in S\}$).
2. Remove all points, x with $\|x - m\|_2 \geq C\sqrt{d\log(N/\tau)}$ from the samples.
3. Find eigendecomposition of $\hat{\Sigma}$. Let W be the span of bottom $d/2$ eigenvectors and V be the span of top $d/2$ eigenvectors. Then $\|\hat{\mu}_W - \mu_W\|_2 \leq O(\epsilon)$.
4. Recurse on V.

In step 2, we don't know the true mean μ but $\|\mu - m\|_2 \leq O(\epsilon\sqrt{d\log(N/\tau)})$ w.h.p. So, for any point $x \in G$, $\|x - \mu\| \leq \|x - m\| + \|\mu - m\| = O(\sqrt{d\log(N/\tau)})$. There are $\log(d)$ levels of recursion and the total error is $O(\epsilon\sqrt{\log(d)})$.

Idea 2: Remove points so that $\|\hat{\Sigma}\|_2$ is close to 1.
Let $N = |S|$. Using a union bound, w.p. $\tau/3$, for all $x \in G$ we have

$$\|x - \mu\|_2 = O(\sqrt{d\log(N/\tau)}).$$

The number of samples N is at least $\Omega(\frac{d^2}{\epsilon^2}\log(d/\epsilon\tau))$. Let $\alpha = \frac{1}{\log(d\log(\frac{d}{\epsilon\tau}))}$. The next 2 lemmas prove the correctness of Algorithm 2.

Lemma 6.3. *If $\left\|\hat{\Sigma}\right\|_2 \geq 1 + C\epsilon\sqrt{\log(1/\epsilon)}$, then there exists $v \in \mathbb{R}^d$, $\|v\| = 1$ and $t > 0$ such that*

$$\Pr_S(|v^\top x - v^\top \mu| > t + 2) > 8e^{-t^2/2} + \frac{8\epsilon\alpha}{t^2}. \tag{6.2}$$

Proof. Without loss of generality, let $\mu = 0$. Let v be the top eigenvector of $\hat{\Sigma}$. If $\Pr_S(|v^\top x| > t+2) \leq 8e^{-t^2/2}$ for all $t > 0$, then

$$v^\top \Sigma_B v = \mathbb{E}_B[(v^\top(x - \mu_B))^2] = \mathbb{E}_B[(v^\top x)^2] - (v^\top \mu_B)^2$$

$$\leq 2\int_{t=0}^{\infty} t \Pr_B(|v^\top x| \geq t)\, dt = 2\int_{t=0}^{O(\sqrt{d\log(N/\tau)})} t \Pr_B(|v^\top x| \geq t)\, dt.$$

We can restrict to $|v^\top x| \leq \|x\| \leq \sqrt{d\log(N/\tau)}$ after naively pruning. Note that $\Pr_B(|v^\top x| \geq t) \leq \frac{|S|}{|B|}\Pr_S(|v^\top x| \geq t)$.

$$\begin{aligned}
v^\top \Sigma_B v &\leq 2\int_{t=0}^{O(\sqrt{\log(1/\epsilon)})} t\, dt + \frac{2}{\epsilon}\int_{O(\sqrt{\log(1/\epsilon)})}^{O(\sqrt{d\log(N/\tau)})} t \Pr_S(|v^\top x - \mu| \geq t)\, dt \\
&\leq \log(1/\epsilon) + \frac{16}{\epsilon}\int_{O(\sqrt{\log(1/\epsilon)})}^{O(\sqrt{d\log(N/\tau)})} t e^{-\frac{(t-2)^2}{2}}\, dt + 16\int_{O(\sqrt{\log(1/\epsilon)})}^{O(\sqrt{d\log(N/\tau)})} \frac{\alpha}{t}\, dt \\
&\leq \log(1/\epsilon) + \frac{16}{\epsilon}\int_{O(\sqrt{\log(1/\epsilon)})}^{O(\sqrt{d\log(N/\tau)})} (t-2) e^{-\frac{(t-2)^2}{2}}\, dt + \frac{32}{\epsilon}\int_{O(\sqrt{\log(1/\epsilon)})}^{O(\sqrt{d\log(N/\tau)})} e^{-\frac{(t-2)^2}{2}}\, dt + O(1) \\
&\leq \log(1/\epsilon) + O(\epsilon) + O(1)
\end{aligned}$$

Using equation (6.1),

$$v^\top \hat{\Sigma} v \leq (1-\epsilon) + \epsilon\log(1/\epsilon) + O(\epsilon),$$

which is a contradiction. □

The idea is to remove all points with $|v^\top x| > t+2$ from the sample and iterate. The next lemma implies that at least half of the removed points are from B. In the end, $\|\hat{\Sigma}\|_2$ is small and at most 2ϵ fraction of the points are removed.

Lemma 6.4. *For all unit vectors $v \in \mathbb{R}^d$ and $t > 0$,*

$$\Pr_G(|v^\top x - \mu^\top v| > t) \leq 2e^{-t^2/2} + \frac{\epsilon\alpha}{t^2} \tag{6.3}$$

with probability at least $1-\tau$.

Proof. Wlog, let $\mu = 0$. Let $\delta = \epsilon\alpha$. Let $n = |G| \geq (1-\epsilon)N$. We will prove that for any $v \in \mathbb{R}^d$ with $\|v\| = 1$ and $t > 0$,

$$|\Pr_G(|v^\top x| > t) - \Pr_{x\sim N(0,1)}(|v^\top x| > t)| \leq \frac{\delta}{t^2}.$$

Since the VC-dimension of the set of all halfspaces is $d+1$, if $t < \sqrt{C\log(1/\delta)}$, this bound is true with probability at least $1-\tau/3$ if we have more than $\Omega(\frac{d\log(1/\delta)^2}{\delta^2})$ samples using the VC inequality from [devroye2012combinatorial].

We only need to consider the case when $t \geq \sqrt{C\log(1/\delta)}$. Let E_i denote the event that $|v^\top x_i| > t$. Then $\Pr(E_i) \leq 2e^{-\frac{t^2}{2}}$ and E_i's are mutually independent. Note that $\Pr_G(|v^\top x| > t) = \sum_i 1_{E_i}/n$. Therefore,

$$\mathbb{E}[e^{\frac{T^2 n}{3}\Pr_G(|v^\top x|>t)}] \leq (1+e^{-t^2/2}e^{\frac{t^2}{3}})^n = (1+e^{-\frac{t^2}{6}})^n \leq (1+\delta^2)^n \leq e^{\delta^2 n}.$$

Using Markov's inequality,

$$\Pr(\Pr_G(|v^\top x| > t) \geq \delta/T^2) \leq \frac{\mathbb{E}[e^{\frac{T^2 n}{3}\Pr_G(|v^\top x|>t)}]}{e^{\frac{\delta n}{3}}} \leq e^{n\delta^2 - \frac{n\delta}{3}} \leq e^{-\frac{n\delta}{6}}.$$

Let $\mathcal{C}$ be a 1/2-net for unit vectors in $\mathbb{R}^d$. Then $|\mathcal{C}| = 2^{O(d)}$. From equation , $|v^\top x| \leq O(\sqrt{d\log(n/\tau)})$. Let $R = c\sqrt{d\log(n/\tau)}$ for some large constant c and let D be set of all powers of 2 between $\sqrt{C\log(1/\delta)}$ and R. Since $n = \Omega(\frac{d}{\epsilon}\log(1/\tau))$, for any $v' \in \mathcal{C}$ and $t' \in D$, with probability at least

$$1 - e^{-\frac{n\delta}{6}}|\mathcal{C}|\cdot|D| \geq 1-\tau/2$$

we have

$$\Pr_S(|v'^\top x| > t') \leq \frac{\delta}{t'^2}.$$

For any unit vector $v \in \mathbb{R}^d$ and $t \in [\sqrt{C \log(1/\delta)}, R)$. Then, there exists $t' \in D$ such that $t \leq t' \leq 2t$ and $v' \in C$ such that $|v^\top x| \leq 2|v'^\top x|$. So, $|v^\top x| > t$ implies $|v'^\top x| > t'$, and

$$\Pr_S(|v^\top x| > t) \leq \Pr_S(|v'^\top x| > t') \leq \frac{\delta}{Ct'^2} \leq \frac{\delta}{Ct^2}. \qquad \square$$

From equations (6.2) and (6.3), in each iteration, we remove more corrupted than uncorrupted points.

Algorithm 2: Iterative Filtering

1. Naively prune points with large norms.
2. If $|\hat{\Sigma}|_2 \leq 1 + C\epsilon\sqrt{\log(1/\epsilon)}$, output $\hat{\mu}$.
3. Let v be the top eigenvector of $\hat{\Sigma}$ and $m = \text{median}(\{v^\top x : x \in S\})$.
4. Find $t > 0$ such that
$$\Pr_S(|v^\top x - m)| > t + 2) > 8e^{-t^2/2}.$$
5. Recurse on $S' = \{x \in S : |v^\top x - m| \leq t + 2\}$

In step 2, we don't know the value of $v^\top \mu$ but $|v^\top \mu - m| \leq O(\epsilon)$ w.h.p. So, for any point $x \in G$,

$$|v^\top x - v^\top \mu| < |v^\top x - m| + |v^\top \mu - m| = O(1).$$

References

[1] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation.* Springer Science & Business Media, 2012.