14.1 Overview

14.1.1 Last Time

Robust Statistics: The given samples did not necessarily all come from the distribution. About $\epsilon$-fraction of them are corrupted. Looked at different models on how to incorporate this, but they either took time in exponential in dimension or the error guarantees had polynomial dependence on dimension.

- Robust Statistics model
- Classical Approaches
- Stability Condition

14.1.2 Today

We will look at the weighted version of stability condition and get better run-time/error guarantee algorithms.

- Weighted Stability
- Algorithms

14.2 Some Definitions

**Stability Condition Def:** A set $S$ is $(\epsilon, \delta)$ - stable (wrt $G$) if for any $T \subset S, |T| \geq (1 - \epsilon)|S|$ and for any unit vector $v$-

\[
\left| \frac{1}{|T|} \sum_{x \in T} v(x - \mu_G) \right| < \delta \\
\left| \frac{1}{|T|} \sum_{x \in T} v(x - \mu_G)^2 - 1 \right| < \delta^2 / \epsilon
\]  

(14.1)  

(14.2)

**Proposition:** If $S$ is $(\epsilon, \delta)$ - stable, if $T$ s.t. $|S \cap T| \geq (1 - \epsilon)|S|, (1 - \epsilon)|T|$ and if the maximum eigen value of $\sum_T$ is at most $(1 + \lambda)$ then,

\[
|\mu_G - \mu_T|_2 = O(\delta + \sqrt{\lambda \epsilon})
\]  

(14.3)

This means that if we know with high probability the good samples that we have formed a stable set, then using the proposition we have sample mean is close to the original distribution.

So, if covariance matrix is bounded we have a way of certifying that we have the correct answer.
14.3 Weighted Stability Condition

We slightly adapt the definition of the stability condition to hold for probability distributions instead of just sets. Essentially, we replace summation with expectation, etc. as below:

**Modified Def:** A distribution $S$ is $(\epsilon, \delta)$ stable (wrt $G$) if for any event $T$, $\Pr(T) \geq (1-\epsilon)|S|$ and for any unit vector $v$-

$$|\mathbb{E}_{x \sim (S|T)} v(x) - \mu_G| < \delta$$

$$|\mathbb{E}_{x \sim (S|T)} v(x)^2 - 1| < \delta^2/\epsilon$$

**Modified Proposition:** If $S$ is $(\epsilon, \delta)$ stable, if $T$ s.t. $d_{TV}(S, T) < \epsilon$ and if the maximum eigen value of $\sum_T$ is at most $(1 + \lambda)$ then,

$$|\mu_G - \mu_T|^2 = O(\delta + \sqrt{\lambda \epsilon})$$

14.3.1 Example 1

If $S = \mathcal{N}(\mu, I)$ which is $(\epsilon, O(\sqrt{\log(1/\epsilon)})$-stable. Taking a unit vector, $v$ and $v.S$ is a standard normal distribution. Now, we want an event $T$ that either moves the mean or changes the covariance (so it’s not stable). In general it’s important to note that you move the mean/variance as much as possible by throwing out the epsilon-tails.

In particular, to move the mean we remove the $\epsilon$ tail (point lies $\approx \sqrt{\log(1/\epsilon)}$ far from origin). Now the question arises by how much does it change the mean? It changes it by- $O(\epsilon \sqrt{\log(1/\epsilon)})$.

Looking at the second condition in the stability condition, i.e. the change in variance in terms of upper bound- $\text{var}(v.T) \leq 1/(1-\epsilon)$. This makes sense as the variance cannot be increased by throwing away points from the distribution. The lower bound case is slightly trickier as removing mass from $S$ can actually lower the variance- $\text{var}(v.T) \geq 1 - O(\epsilon \log(1/\epsilon))$

What did we finally learn from this analysis? If we take the continuous Gaussian distribution and remove $\epsilon$ mass, it will be stable. Later we will see that this holds even in the discretized versions.

14.3.2 Example 2

What happens if $\text{cov}(S) = O(I)$, variance in any direction is atmost some universal constant.

$S$ is $(\epsilon, O(\sqrt{\epsilon})$-stable. We want to answer what happens if we throw away some tail. $L$ is the point where we cut the tail off.

$$\mathbb{E}_L [v(x) \mu] << 1/\sqrt{\epsilon} \text{ using Jenson’s inequality}$$

$$\mathbb{E}_L [v.(x-\mu)^2] << 1/\epsilon$$

So, the mean would look something like- $\mathbb{E}_{S \setminus L} [v.(x-\mu)] = O(\sqrt{\epsilon})$

The variance would be- $\text{Cov}(v.T) < O(I)$, same argument as before. And $\text{Cov}(v.T) > 0$ because they are always positive semi-definite.
Basically, if you have the full distribution, then it’s about the tail bounds. Nice tighter tail bounds like the one in Gaussian guarantees the stability condition. But if you only know the bounds on say the covariance matrix, then you will still be stable with $\delta = \sqrt{\epsilon}$.

### 14.4 Algorithms

Suppose our good samples, $S$, are $(\epsilon, \delta)$-stable. But we do not see this set of good samples. Instead, we see this other set $T$, which is $\epsilon$ fraction corrupted. Can’t apply the proposition directly, we try to do some outlier removal.

We try to find a set $T' \subset T$ s.t.-

- $T'$ is close to $S$
- $\text{Cov}(T')$ bounded

If we can find this set $T'$, we can apply the proposition to $T'$, sample mean over $T'$ will be close to our true distribution.

**How to search for $T'$?**

Given $T, \epsilon$, define $W$ (set of weights) = \{ $w_x : x \in T \mid w_x \geq 0, \sum w_x = 1, w_x \leq 1/(1-\epsilon)|T|$ \}. This set is a nice convex set with desirable properties.

**Important things to note:**

1. For any $T'$ in $W$, $d_{TV}(T', S) = O(\epsilon)$

\[
d_{TV}(T', S) = \sum_{x \in T} \max(0, w_x - \frac{\mathbb{1}_{x \in S}}{|S|})
= \sum_{x \in T \setminus S} w_x + \sum_{x \in T \cap S} \max(0, w_x - \frac{1}{|S|}) , \text{ counting only the places where val}(T') > \text{val}(X)
\leq \frac{\epsilon |T|}{(1-\epsilon)|T|} + \frac{\epsilon |T|}{(1-\epsilon)|T|}
= O(\epsilon)
\]

2. If we find $w \in W$ s.t. $\text{Cov}(w) \leq (1 + \lambda)I$, then by the proposition $|\mu_G - \mu_W|_2 = O(\delta + \sqrt{\epsilon})$

3. If we take $w^* = \text{uniform distribution over } S \cap T$, by 14.5, $\text{Cov}(w^*) \leq (1 + O(\delta^2/\epsilon))I$. Now, we know that there is a solution to this and if we find this solution, it will give a good approximation to the mean.

We have successfully reduced robust statistics problem into a computational problem.

**Problem:** Given $T, \epsilon, \delta$, find some $w \in W$ s.t. $\text{Cov}(W) \leq (1 + O(\delta^2/\epsilon))I$. If we find this, then $|\mu_W - \mu_G| = O(\delta + \sqrt{\delta^2/\epsilon}) = O(\delta)$

**Solutions:** Several techniques. We will be covering two such techniques.

#### 14.4.1 Convex Programming

$W$ is convex, set is that are weights need to be in. For any $v$, which is a unit vector,

\[
\sum_x w_x (|v.(x - \mu_w)|^2) < 10\delta^2/\epsilon
\]  

(14.7)
This constraint is very close to being a set of linear equations. If we had a fixed constant instead of the $\mu_w$ term in the above equation, we could have solved this problem. Now we have to do a little more work before we can use convex programming technique. We need a separation oracle to separate solution that doesn’t work and the ones which do work.

**Lemma 14.1** Given $W$ with Cov$(w) \not\leq O(1 + O(\delta^2/\epsilon))I$, find a linear relation $L$ s.t. $L(w) < L(w^*)$. Using this as a separator we can make use of linear programming algorithms.

If Cov$(w) > O(1 + O(\delta^2/\epsilon))I$, then there is a unit vector, $v$ s.t. Cov$(w)$ in its direction is too big-

$$\sum w_x v.(x - \mu_w)^2 > 1 + 10\delta^2/\epsilon$$

(14.8)

and

$$\sum w_x^* v.(x - \mu_w^*)^2 \leq 1 + 10\delta^2/\epsilon$$

(14.9)

We want to look at the difference of the mean. We know that $d_{TV}(w, w^*) = O(\epsilon)$. It will be-

$$|v.\mu_w - v.\mu_w^*| = O(\sqrt{var(v.w) + var(v.w^*)})$$

(14.10)

using Cauchy-Schwartz.

Unless you change the variance by lot, you can’t change the mean by that much. Now, let’s assume that the variance given by 14.8 is the highest possible-

$$\sum w_x v.(x - \mu_w)^2 = 1 + \lambda$$

$$|\mu_w - \mu_G| = O(\delta + \sqrt{\lambda \epsilon})$$

$$|\mu_w^* - \mu_w| = O(\delta + \sqrt{\lambda \epsilon})$$

Because $w$ and $w^*$ are close in mean.

$$\sum w_x^* v(x - \mu_w)^2 \leq 1 + 5\delta^2/\epsilon + O(\delta^2 + \lambda \epsilon)$$

(14.11)

$$\sum w_x v(x - \mu_w)^2 > 1 + \lambda$$

(14.12)

If $\lambda > 10\delta^2/\epsilon$ then, 14.12 > 14.11 and this implies that there is a separation oracle.

$L(W') = \sum w_x^* v(x - \mu_w)^2$ Also, if we use the $w$, whose covariance is large then we use 14.11, 14.12 to separate $w$ from the $w^*$. Either it satisfies the condition or we can find the linear separator.

**Theorem 14.2** If $S$ is a $(\epsilon, \delta)$-stable set. $\exists$ an algorithm (polytime) given $T$, an $\epsilon$-corrupted version of $S$, computes $\mu_S$ to an error $O(\delta)$.

Current algorithm uses convex program, which is a polynomial time but not a great polynomial.

### 14.4.2 Outlier Removal Filter Algorithm

Assuming that we have a set that didn’t have a bounded covariance matrix. If cov$(T)$ not bounded, then in some direction $v$ s.t. var$(v.T)$ is large.
There are a bunch of outliers that lie together and causing the disruption. Assume not only is $S$ stable, but $S$ has Gaussian tail bounds. For any $v \& T$, $Pr_{x\sim S}(|v.(x-\mu_S)| > T) < 2e^{-ct^2}$.

Looking at the projection- $\mathbb{E}[|v.S|^2] \approx 1$ and $\mathbb{E}[|v.T|^2] > 1 + \delta^2/\epsilon$. We also know that $T \& S$ are not that different from each other, only on $\epsilon$ points.

**What could we have possibly done to get the variance of $T$ that far?** We took an $\epsilon$-mass and moved it far away from the origin, a decent multiple of $\sqrt{lg(1/\epsilon)}$. Almost all of these far away points are bad.

**Idea:** If $\text{Cov}(T) \not< (1 + \delta^2/\epsilon).I$.

Then:

- Find the largest eigen vector.
- Project points on this line.
- Remove the outliers (atleast as many bad points as good ones). (this step is the filter)

Basically, we are reducing the size of symmetric difference between the two sets, $|S\triangle T|$, every time we apply the filter.

**Algorithm:**

```plaintext
while (cov(T) not bounded):
    apply filter;
return $\mu_T$
```

It will eventually terminate and the covariance matrix will be bounded. At every step, since we decrease the size of the symmetric difference, when we reach the stability condition, we can apply the proposition.

**How to make the filter work?** Basically, we want to pick some threshold and throw the points beyond the threshold. We assume tail bounds and as long as the number of points we throw away are at least twice away from the tail bound, we are good.

Find some $t$ s.t. $Pr_{x\sim T}(|v.(x-\mu)| > t) > 4e^{-ct^2}$.

**Know:** Contribution of the bad points to $\mathbb{E}[|v.x|^2] >> \epsilon lg(1/\epsilon)$, assuming $\mu = 0$, can be written as:

$$\int_0^{\inf} Pr(|v.x| > t, t \text{ is bad})2tdt$$

(14.13)

Probability in 14.13 is atmost $\epsilon$. If no threshold,

$$10\epsilon lg(1/\epsilon) \geq \int_{0}^{\max(\epsilon, 4e^{-ct^2})}2tdt$$

$$= \int_{0}^{\text{const}\sqrt{lg(1/\epsilon)}}2tdt + \int_{\text{const}\sqrt{lg(1/\epsilon)}}^{8t.e^{-ct^2}}8t.e^{-ct^2}dt$$

$$= O(\epsilon lg(1/\epsilon)) + O(\epsilon)$$

If tail points were satisfied, then it could not have contributed this much. Basically, cut-off point is at a multiple of $O(\sqrt{lg(1/\epsilon)})$ gives contradiction. Find the tail bound they violated and then throw away points beyond that. You keep doing this and you will be guaranteed that after a certain point you won’t be able to do it and then you have your stability condition being satisfied. To note, it is a lot faster than convex program.