10/1/23

Lecture 9: Robust Stats, Iterative Filtering

Setup: q has mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \preceq I_d$, nature samples $x_1^*, \ldots, x_n^* \sim q$, adversary corrupts arbitrary η fraction, we are given the corrupted samples $\{x_1, \ldots, x_n\}$

Decompose $\{x_1, \ldots, x_n\}$ into

$$\boxed{S_{good} \cup S_{bad} \setminus S_r}$$

where $|S_{bad}| = |S_r| = \eta n$,

$S_{good} = \{x_1^*, \ldots, x_n^*\}$ are the original i.i.d draws from q, S_r are the points which were corrupted, and S_{bad} are the points they've been replaced with.

we'll use shorthand:

"$\sum_{\text{clean } i} x_i$" : $\sum_{x \in S_{good} \setminus S_r} x$

"$\sum_{\text{bad } i} x_i$" : $\sum_{x \in S_{bad}} x$

Assume for now that for

$$\mu_g \triangleq \frac{1}{|S_{good}|} \sum_{x \in S_{good}} x$$

$$\Sigma_g \triangleq \frac{1}{|S_{good}|} \sum_{x \in S_{good}} (x - \mu_g)(x - \mu_g)^T,$$

① $\|\mu_g - \mu\|_2 \lesssim \sqrt{\varepsilon}$ and ② $\|\Sigma_g\|_{op} \lesssim 1$

(we will see in pset 3 why this is a valid assumption. note: a little subtle b/c we don't assume the higher moments of η are bounded)

We will maintain weights $w = \{w_i\}_{i \in [n]}$ for the dataset that indicate how confident we are that x_i is clean.

$$0 \le w_i \le \frac{1}{n} \quad \forall\, i \in [n] \qquad (*)$$

Ideally we would want $w_i = \frac{1}{n} \cdot \mathbb{1}[i \in S_{good}]$.

NOTE: w_i is like "$\frac{1}{n} a_i$" from SoS analysis.
But these are actual #'s now, not SoS variables

Define the weighted mean and weighted covariance by

$$\mu_w \triangleq \frac{1}{\sum_i w_i} \sum_i w_i x_i$$

$$\Sigma_w \triangleq \frac{1}{\sum_i w_i} \sum_i w_i (x_i - \mu_w)(x_i - \mu_w)^T.$$

In general, $\{w_i\}$ can be viewed as "soft" indicators for a subset. We want to pick out a large subset of clean points, so we care about w_i's that satisfy

$$\sum_i w_i \geq 1-\eta \qquad (**)$$

Note: set of w satisfying $(*)$ and $(**)$ is convex hull K_η of $\left\{ \frac{1}{n} \cdot \mathbb{1}_S : S \subseteq [n] \text{ s.t. } |S| \geq 1-\eta \right\}$

We will maintain that our $w \in K_\eta$ throughout the algorithm.

<u>Main lemma</u>: (if μ_w wrong, $\|\Sigma_w\|$ large):
For any $w \in K_\eta$,

$$\|\mu_g - \mu_w\|_2 \lesssim \sqrt{\eta}\left(1 + \sqrt{\|\Sigma_w\|_{op}}\right)$$

First, let's see what to do with this....

Given current weights w, define scores

$$\tau_i \triangleq \langle v, x_i - \mu_w \rangle^2,$$

where v is top eigvec of Σ_w. Let $\tau_{max} = \max_{i: w_i > 0} \tau_i$

higher score $\tau_i \iff$ more likely x_i is corrupted

Lemma: Consider update rule

$$w_i' \leftarrow \left(1 - \frac{\tau_i}{\tau_{max}}\right) w_i$$

Then if $\sum_{\text{clean } i} w_i \tau_i < \sum_{\text{bad } i} w_i \tau_i$, then

"Safety condition"

$$\sum_{\text{clean } i} (w_i - w_i') < \sum_{\text{bad } i} (w_i - w_i')$$

(more bad mass removed than good mass).

"progress condition"

and $\mathrm{nnz}(w') < \mathrm{nnz}(w)$.

↑

\# nonzero entries

Pf:
$$\sum_{\text{clean } i} (w_i - w_i') = \sum_{\text{clean } i} \frac{\tau_i w_i}{\tau_{max}}$$

$$\text{and} \quad \sum_{\text{bad } i} (w_i - w_i') = \sum_{\text{bad } i} \frac{\tau_i w_i}{\tau_{max}}$$

"Progress condition" immediate from def. $\square$

We will maintain invariant that

$$\sum_{\text{clean } i} \left(\frac{1}{n} - w_i\right) < \sum_{\text{bad } i} \left(\frac{1}{n} - w_i\right), \qquad \text{(INV)}$$

i.e. less good mass removed than bad mass.

As long as $\|\Sigma_w\|$ still large, we update the weights via the rule in the Lemma.

Obs 1: If (INV) always maintained, then algorithm runs for at most $2\eta n$ iterations.

Pf: After $2\eta n$ iterations, we must have removed at least 2η mass total, thus $\geq \eta$ mass from bad points (by INV), resulting in $\sum_{\text{bad } i} w_i = 0$. Because INV always maintained, and there is no bad mass left, alg. must terminate. $\square$

Obs 2: If $\|\Sigma_w\| \lesssim 1$, then safe to output μ_w by Main Lemma.

Obs 3: If (INV) always maintained, we $\mathcal{K}_{2\eta}$.

Pf: $\sum_{\text{clean } i} (\frac{1}{n} - w_i) < \sum_{\text{bad } i} (\frac{1}{n} - w_i) \leq \eta$, so $\sum_{\text{all } i} w_i > 1 - 2\eta$. □

By Obs's 1, 2, suffices to show:

Lemma: Suppose $\|\Sigma_w\| \gg 1$, and (INV) holds, then $\sum_{\text{clean } i} w_i \tau_i < \sum_{\text{bad } i} w_i \tau_i$.

Pf: Note

$$\sum_{\text{all } i} w_i \tau_i = \sum_i w_i \langle v, x_i - \mu_w \rangle^2$$

$$= v^T \left[\sum_i w_i (x_i - \mu_w)(x_i - \mu_w)^T \right] v$$

$$= v^T \Sigma_w v = \|\Sigma_w\|_{op}$$

so suffices to show $\sum_{\text{clean } i} w_i \tau_i \leq \frac{1}{2} \|\Sigma_w\|_{op}$.

Note $\sum_{\text{clean } i} w_i \tau_i \le \frac{1}{n} \sum_{x \in S_{good}} \langle v, x - \mu_w \rangle^2$

$$= \frac{1}{n} \sum_{x \in S_{good}} \langle v, (x - \mu_g) + (\mu_g - \mu_w) \rangle^2$$

$$\le \underbrace{\frac{2}{n} \sum_{x \in S_{good}} \langle v, x - \mu_g \rangle^2}_{A} + \underbrace{\frac{2}{n} \sum_{x \in S_{good}} \langle v, \mu_g - \mu_w \rangle^2}_{B}$$

$A \le 2 v^T \left(\frac{1}{|S_{good}|} \sum_{x \in S_{good}} (x - \mu_g)(x - \mu_g)^T \right) v \lesssim 1$ by initial assumption on P. 1.

$$B \le 2 \langle v, \mu_g - \mu_w \rangle^2$$

$$\le 2 \| \mu_g - \mu_w \|^2$$

(Main Lemma)

$$\lesssim \sqrt{\eta} \left(1 + \sqrt{\| \Sigma_w \|_{op}} \right)$$

Provided η sufficiently small constant,
if $\|\Sigma_w\| \geq C$ for sufficiently large constant,

$$Ⓐ + Ⓑ \leq 2 + O\left(\sqrt{\eta}\left(1 + \sqrt{\|\Sigma_w\|_{op}}\right)\right)$$

$$\leq \frac{\|\Sigma_w\|_{op}}{2}. \qquad \square$$

Algorithm:

- $w_i \leftarrow \frac{1}{n} \quad \forall\, i \in [n]$
- while $\|\Sigma_w\|_{op} \geq C$:
  - $v \leftarrow$ top eigenvec of Σ_w
  - $\tau_i \leftarrow \langle v, x_i - \mu_w \rangle^2 \quad \forall\, i \in [n]$
  - $\tau_{max} \leftarrow \max_{i: w_i > 0} \tau_i$
  - $w_i' \leftarrow w_i \left(1 - \frac{\tau_i}{\tau_{max}}\right)$
- output μ_w

Remains to prove Main Lemma:

<u>Pf</u>: (will look like an SoS proof)

(for $x \in S_{good} \setminus [n]$, define $w_i := 0$)

$$\boxed{\left(\sum_{\text{all } i} w_i\right) \|\mu_w - \mu_g\|^2}$$

$$= \sum_{\text{all } i} w_i \langle \mu_w - \mu_g, \mu_w - \mu_g \rangle$$

$$= \sum_{\text{all } i} w_i \langle x_i - \mu_g, \mu_w - \mu_g \rangle$$

$$= \sum_{\text{bad } i} \text{"} \qquad \text{"} + \sum_{i \in S_{good}} \text{"} \qquad \text{"}$$

$$= \boxed{\sum_{\text{bad } i} w_i \|\mu_w - \mu_g\|^2} + \underbrace{\sum_{i \in S_{good}} \frac{1}{n} \langle x_i - \mu_g, \mu_w - \mu_g \rangle}_{=0} \qquad ([illegible])$$

$$+ \underbrace{\sum_{\text{bad } i} w_i \langle x_i - \mu_w, \mu_w - \mu_g \rangle}_{\text{I}} + \underbrace{\sum_{i \in S_{good}} \left(w_i - \frac{1}{n}\right) \langle x_i - \mu_g, \mu_w - \mu_g \rangle}_{\text{II}}$$

$$\text{I}^2 = \left(\sum_{\text{bad } i} \underbrace{w_i}_{\sqrt{w_i} \cdot \sqrt{w_i}} \underline{\langle x_i - \mu_w, \mu_w - \mu_g \rangle} \right)^2$$

$$\leq \left(\sum_{\text{bad } i} w_i\right) \cdot \left(\sum_{\text{bad } i} w_i \langle x_i - \mu_w, \mu_w - \mu_g\rangle^2\right)$$

$$\leq \eta \cdot \sum_{\underline{\text{all } i}} \text{''} \qquad \text{''}$$

$$= \eta \cdot (\mu_w - \mu_g)^T \underbrace{\left[\sum_{\text{all } i} w_i (x_i - \mu_w)(x_i - \mu_w)^T\right]}_{\Sigma_w} (\mu_w - \mu_g)$$

$$\leq \eta \, \|\Sigma_w\|_{op} \cdot \|\mu_w - \mu_g\|^2$$

So $\quad \text{I} \leq \sqrt{\eta} \cdot \sqrt{\|\Sigma_w\|_{op}} \cdot \|\mu_w - \mu_g\|$

$$\text{II}^2 = \left(\underbrace{\sum_{i \in S_{good}} n\left(w_i - \frac{1}{n}\right)^2}\right) \cdot \left(\underbrace{\frac{1}{n}\sum_{i \in S_{good}} \langle x_i - \mu_g, \mu_w - \mu_g\rangle^2}_{(\mu_w - \mu_g)^T \Sigma_g (\mu_w - \mu_g)}\right)$$

$|n(w_i - \frac{1}{n})| \leq 1$, and $w_i = 0$ for $i \in S_{good} \setminus [n]$

$$\leq \sum_{\text{clean } i} \left|w_i - \frac{1}{n}\right| + \sum_{i \in S_{good} \setminus [n]} \frac{1}{n}$$

$$\lesssim \|\mu_w - \mu_g\|^2$$

$\leq 2\eta$ (b/c $w \in K_\eta$ and $|S_{good} \setminus [n]| \leq \eta n$)

So $\quad \text{II} \leq \sqrt{2\eta}\, \|\mu_w - \mu_g\|$

Substituting into $(\star\spadesuit\spadesuit)$, rearranging, and dividing by $\|\mu_w - \mu_g\|$, we get

$$\Big(\sum_{\text{clean } i} w_i\Big)\|\mu_w - \mu_g\| \lesssim \sqrt{\eta}\Big(1 + \sqrt{\|\Sigma_w\|_{op}}\Big). \qquad \square$$