10/11/23

Lecture 10: List-decodable learning

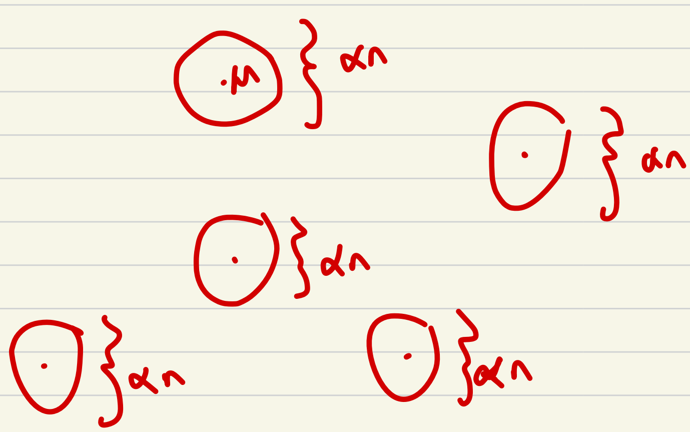
Setup: Let $0 < \alpha < \frac{1}{2}$ be a small constant.

Let q have mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \preceq I_d$,

nature samples $x_1^*, \ldots, x_n^* \sim q$, adversary corrupts arbitrary $\boxed{1-\alpha}$ fraction, we are given the corrupted samples $\{x_1, \ldots, x_n\}$

(# bad points overwhelms # good points!)

One might expect it to be impossible to do anything here. For example, what if the corrupted dataset looked like



where the adversary has created $\frac{1}{\alpha}$ many clusters, each of which is an equally plausible explanation for the data?

At least in this case, could still hope to produce a list of estimates $\hat{\mu}_1, \dots, \hat{\mu}_m$ for $m = O(1/\alpha)$ s.t. $\exists j \in [m]$ s.t. $\|\mu - \hat{\mu}_j\|$ is small.

(note: this generalizes mixture models! i.e. "k" $\approx \frac{1}{\alpha}$)

Amazing fact: this task, "list-decodable mean estimation," is possible in general! (even w/ a practical algorithm)

What is a natural baseline to aim for?

If corrupted dataset looks like a mixture of $k = O(\frac{1}{\alpha})$ bounded-covariance dist's, each "cluster" has radius $\approx \sqrt{d}$. If we project down to the span of the means, then each projected cluster has radius $\approx \sqrt{k}$.

So as long as clusters are $\sqrt{k}$-separated, can hope to cluster and learn all k centers.

So might hope to produce list of estimates s.t. at least one estimate is $O(\sqrt{k}) = O(1/\sqrt{\alpha})$-close to μ.

Thm [Diakonikolas-Kane-Kongsgaard-Li-Tian '20]: For $n = \Omega\left(\frac{d}{\alpha}\right)$, there is an $O\left(\frac{nd}{\alpha}\right)$-time algorithm for list-decodable mean estimation to error $O(1/\sqrt{\alpha})$. (runtime essentially optimal)

Today: "Baby" version of their result that runs in time $\tilde{O}(n^2 d/\alpha)$.

Assumption: there is an $\Omega(\alpha)$ fraction of "good points" $G \subseteq [n]$ s.t.*

$$\left\| \frac{1}{|G|} \sum_{i \in G} (x_i - \mu)(x_i - \mu)^T \right\|_{op} \lesssim 1 \qquad (*)$$

---

* This holds when the adversary is additive, i.e. there are αn i.i.d. draws from q in the dataset.

(Proposition 8.1 from [Charikar-Steinhardt-Valiant '17]).

Obs 1: If we can produce subspace V of dimension $O(1/\alpha)$ s.t. μ close to V, i.e. $\|\Pi_V^\perp \mu\| \lesssim O(1/\sqrt{\alpha})$, then the following alg. solves the task:

1) select $\Theta(1/\alpha)$ points at random from dataset.
2) project these to V and output them

Pf: By ($*$),

$$\frac{1}{|G|}\sum_{i\in G}(\Pi_V x_i - \Pi_V \mu)(\Pi_V x_i - \Pi_V\mu)^T\Big\|_{op} \preceq \Pi_V$$

By taking traces,

$$\Rightarrow \frac{1}{|G|}\sum_{i\in G}\|\Pi_V x_i - \Pi_V\mu\|^2 \leq \dim(V) = O(1/\alpha),$$

so by Markov's, 99% of points in G satisfy $\|\Pi_V x_i - \Pi_V \mu\| \lesssim O(1/\sqrt{\alpha})$. Additionally, $\|\Pi_V u - u\| = \|\Pi_V^\perp u\| \lesssim O(1/\sqrt{\alpha})$. So as long as the $\Theta(1/\alpha)$ points in step 1

contain one of these i's, we are done.
There are $0.99|G| = \Omega(\alpha n)$ such i's,
so we succeed w.h.p. $\square$

So suffices to find $O(1/\alpha)$-dimensional subspace V s.t. μ is $O(\sqrt{\alpha})$-close to V.

In fact, weaker goal suffices:

$$\boxed{\text{find } O(1/\alpha)\text{-dim subspace } V \text{ s.t. we can estimate } \Pi_V^{\perp}\mu \text{ to error } O(1/\sqrt{\alpha}).}$$

(Note: information-theoretically possible: just solve list-decodable mean estimation, and let V be the span of the output vectors.)

Idea: apply iterative filtering, take V to be top-$O(1/\alpha)$ singular subspace of Σ_w, use variant of spectral signatures lemma to argue we can estimate $\Pi_V^{\perp}\mu$ well enough.

Recall notation:

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

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

previously didn't have normalization because $\sum_i w_i$ close to 1, but now $\sum_i w_i$ may be very small.

Throughout, assume $w_1, \ldots, w_n \leq \frac{1}{n}$

(we initialize at $w_i = \frac{1}{n}$ and will only ever decrease these weights)

	$\eta < \frac{1}{2}$ corruption	list-decodable
Invariant on weights	$\sum_{\text{clean } i} \frac{1}{n} - w_i < \sum_{\text{bad } i} \frac{1}{n} - w_i$ more bad weight removed than good	$\sum_{\text{clean } i} w_i \geq \alpha \sqrt{\sum_{\text{all } i} w_i}$ as weights decrease, good mass becomes more and more pronounced.
Termination condition	$\\|\Sigma_w\\|_{op} \lesssim 1$	$\sigma_k(\Sigma_w) \lesssim \frac{1}{\boxed{\sqrt{\sum_i w_i}}}$ for $k = \Theta(1/\alpha)$ kth largest eigenvalue
Scores (τ_i)	$\tau_i = \langle u^T, x_i - \mu_w \rangle^2$, u top eigenvec of Σ_w squared magnitude of residual projected in top eigendirection	$\tau_i = \\|(\Sigma_w^{(k)})^{-1/2} V_w^T (x_i - \mu_w)\\|_2^2$ $\Sigma_w^{(k)} \triangleq V_w^T \Sigma_w V_w$ (projected covariance) V_w: $k \times d$ matrix whose rows are top k eigenvecs of Σ_w squared magnitude of residual projected to top-k eigenspace + whitened using $\Sigma_w^{(k)}$
Condition on $\{\tau_i\}$ needed for invariant to be maintained	$\sum_{\text{clean } i} w_i \tau_i < \frac{1}{2} \sum_{\text{all } i} w_i \tau_i$ average score of clean point less than average score of bad point	$\frac{1}{\sum_{\text{clean } i} w_i} \sum_{\text{clean } i} w_i \tau_i \leq \frac{1}{2 \sum_{\text{all } i}} \sum_{\text{all } i} w_i \tau_i$ "normalized" version
Spectral Signatures lemma	$\\|\mu_w - \mu\\| \lesssim \sqrt{\eta}\left(1 + \sqrt{\\|\Sigma_w\\|_{op}}\right)$	$\\|\mu_w - \mu\\| \lesssim \frac{1}{\sqrt{\alpha}} \cdot \sqrt{1 + \boxed{\sqrt{\sum_i w_i}} \cdot \\|\Sigma_w\\|_{op}}$

① If we hit termination condition, then done

Pf: apply spectral sig lemma to data projected to subspace orthogonal to top-k singular subspace V_w of Σ_w. Then

$$\| \Pi^{\perp} \mu_w - \Pi^{\perp} \mu \| \lesssim \frac{1}{\sqrt{\alpha}} \sqrt{1 + \boxed{\sqrt{\sum_i w_i}} \underbrace{\| \Pi^{\perp} \Sigma_w \Pi^{\perp} \|_{op}}_{\substack{\le \sigma_k(\Sigma_w) \\ \le \frac{1}{\boxed{\sqrt{\sum_i w_i}}}}}}$$

$$\lesssim \frac{1}{\sqrt{\alpha}}.$$

So we have an $O(1/\alpha)$-dim subspace and an $O(\frac{1}{\sqrt{\alpha}})$-accurate estimate of $\Pi^{\perp}\mu$.

---

② If condition on $\{\tau_i\}$ holds, downweighting maintains invariant.

Pf: Recall downweighting rule: $w_i' \leftarrow w_i \left(1 - \frac{\tau_i}{\tau_{max}}\right)$

Suffices to show

$$\frac{\sum_{\text{clean } i} w_i'}{\sum_{\text{clean } i} w_i} \ge \sqrt{\frac{\sum_{\text{all } i} w_i'}{\sum_{\text{all } i} w_i}}$$

$$\text{LHS} = \frac{\sum_{\text{clean } i} w_i \left(1 - \frac{\tau_i}{\tau_{max}}\right)}{\sum_{\text{clean } i} w_i}$$

$$= 1 - \frac{1}{\tau_{max}} \cdot \frac{\sum_{\text{clean } i} w_i \tau_i}{\sum_{\text{clean } i} w_i}$$

$$\geq 1 - \frac{1}{\tau_{max}} \frac{1}{2} \frac{\sum_{\text{all } i} w_i \tau_i}{\sum_{\text{all } i} w_i}$$

$$= 1 - \frac{1}{2}\left(1 - \frac{\sum_{\text{all } i} w'_i}{\sum_{\text{all } i} w_i}\right)$$

$1 - \frac{1}{2}x \geq \sqrt{1-x}$

$$\geq \sqrt{\frac{\sum_{\text{all } i} w'_i}{\sum_{\text{all } i} w_i}} = \sqrt{\text{RHS}}. \qquad \square$$

---

③ If we haven't hit termination condition, Condition on scores holds.

Pf: w.t.s.

$$\frac{1}{\sum_{\text{clean } i} w_i} \sum_{\text{clean } i} w_i \tau_i \leq \frac{1}{2} \frac{1}{\sum_{\text{all } i} w_i} \sum_{\text{all } i} w_i \tau_i \qquad (+)$$

Note: $\tau_i = \left\| \left(\Sigma_w^{(k)}\right)^{-1/2} V_w^T (x_i - \mu_w) \right\|^2$

$$= \mathrm{Tr}\left(\left(\Sigma_w^{(k)}\right)^{-1/2} V_w^T (x_i - \mu_w)(x_i - \mu_w)^T V_w \left(\Sigma_w^{(k)}\right)^{-1/2} \right)$$

$$= \left\langle \left(\Sigma_w^{(k)}\right)^{-1}, V_w^T (x_i - \mu_w)(x_i - \mu_w)^T V_w \right\rangle,$$

so

$$\text{RHS of } (+) = \frac{1}{2} \frac{1}{\sum_{\text{all } i} w_i} \sum_{\text{all } i} w_i \tau_i$$

$$= \frac{1}{2} \left\langle \left(\Sigma_w^{(k)}\right)^{-1}, V_w^T \Sigma_w V_w \right\rangle$$

$$= \frac{1}{2} \mathrm{Tr}(I_{d_k}) = \frac{k}{2}$$

For LHS of $(+)$, split $x_i - \mu_w$ into

$$\Big(x_i - \underbrace{\frac{1}{|G|} \sum_{\text{clean } i} w_i x_i}_{\triangleq\, \mu_{w,G}} \Big) + \Big(\frac{1}{|G|} \sum_{\text{clean } i} w_i x_i - \mu_w \Big)$$

Also define $\Sigma_{w,G} \triangleq \frac{1}{\sum_{\text{clean } i} w_i} \sum_{\text{clean } i} w_i (x_i - \mu_w)(x_i - \mu_w)^T$

We have

$$\tau_i \le \underbrace{2\left\| \left(\Sigma_w^{(k)}\right)^{-1/2} V_w^T (x_i - \mu_{w,G}) \right\|^2}_{\gamma_i} + \underbrace{2\left\| \left(\Sigma_w^{(k)}\right)^{-1/2} V_w^T (\mu_{w,G} - \mu_w) \right\|^2}_{\delta_i}$$

$$\frac{2}{\sum_{\text{clean } i} w_i} \sum_{\text{clean } i} w_i \gamma_i = 2\left\langle \left(\Sigma_w^{(k)}\right)^{-1}, V_w^T \Sigma_{w,G} V_w \right\rangle$$

note $\Sigma_{w,G} \preceq \frac{1}{\sum_{\text{clean } i} w_i} \sum_{\text{clean } i} w_i (x_i - \mu)(x_i - \mu)^T$

$w_i \le \frac{1}{n} \le \frac{\alpha}{|G|}$

$$\preceq \frac{\alpha}{\sum_{\text{clean } i} w_i} \cdot \underbrace{\frac{1}{|G|} \sum_{\text{clean } i} (x_i - \mu)(x_i - \mu)^T}_{\preceq I_d}$$

$$\preceq \frac{\alpha}{\sum_{\text{clean } i} w_i} \cdot I_d$$

so

$$\le \frac{2\alpha}{\sum_{\text{clean } i} w_i} \cdot \mathrm{Tr}\Big(\underbrace{\left(\Sigma_w^{(k)}\right)^{-1}}_{\text{all eigvals } \lesssim \sqrt{\sum_{\text{all } i} w_i}} \Big)$$

$$\lesssim \underbrace{\boxed{\frac{2\alpha \sqrt{\sum_{\text{all } i} w_i}}{\sum_{\text{clean } i} w_i}}}_{\le 1 \text{ by invariant}} \cdot k \le \frac{k}{4}$$

by taking constant factor in termination condition large enough

$vv^T \preceq M \quad \|v\|^2 \le \|M\|_{op} \quad \|v\|^2$

$$\frac{2}{\sum_{\text{clean } i} w_i} \sum_{\text{clean } i} w_i \delta_i = 2 \left\| \left(\Sigma_w^{(k)}\right)^{-1/2} V_w^T (\mu_w - \mu_{w,G}) \right\|_2^2$$

note $\mathbb{E}[X]\mathbb{E}[X] \preceq \mathbb{E}[XX^T]$ by Jensen's, so

$$(\mu_w - \mu_{w,G})(\mu_w - \mu_{w,G})^T \preceq \frac{1}{\sum_{\text{clean } i} w_i} \sum_{\text{clean } i} w_i (x_i - \mu_w)(x_i - \mu_w)^T$$

$$\preceq \frac{1}{\sum_{\text{clean } i} w_i} \sum_{\text{all } i} w_i (x_i - \mu_w)(x_i - \mu_w)^T$$

$$= \frac{\sum_{\text{all } i} w_i}{\sum_{\text{clean } i} w_i} \cdot \Sigma_w$$

so

$$\le \frac{2 \sum_{\text{all } i} w_i}{\sum_{\text{clean } i} w_i} \underset{\text{by invariant}}{\le} \frac{2\sqrt{\sum_{\text{all } i} w_i}}{\alpha} \le \frac{2}{\alpha} \le \frac{k}{4}$$

if we take $k \ge \frac{8}{\alpha}$.

Therefore, LHS of $(+)$ $\le$ RHS of $(+)$ as claimed. $\square$

All that remains is

④ Proof of spectral signature lemma, i.e.

$$\|\mu_w - \mu\| \lesssim \frac{1}{\sqrt{\alpha}} \sqrt{1 + \frac{1}{\sqrt{\sum_i w_i}} \|\Sigma_w\|_{op}}$$

We'll show μ_w and μ are both close

to $\quad \hat{\mu} \triangleq \frac{1}{\langle w, w^* \rangle} \sum_{\text{all } i} w_i w_i^* x_i \quad ,$

where $\quad w_i^* \triangleq \frac{1}{|G|} \cdot \mathbb{1}[i \in G]$.

I) $\|\hat{\mu} - \mu\|^2 = \sup_{u \in S^{d-1}} \langle \hat{\mu} - \mu, u \rangle^2$

$$= \sup_u \left\langle \frac{1}{\langle w, w^* \rangle} \cdot \sum_i w_i w_i^* (x_i - \mu), \, u \right\rangle^2$$

$$\leq \sup_u \frac{1}{\langle w, w^* \rangle} \sum_i w_i w_i^* \langle x_i - \mu, u \rangle^2$$

$$\leq \sup_u \frac{1}{\langle w, w^* \rangle} \frac{1}{n} \underbrace{\sum_i w_i^* \langle x_i - \mu, u \rangle^2}_{= u^T \Sigma_g u \leq 1}$$

$$\leq \frac{1}{n \langle w, w^* \rangle}$$

Note: $\langle w, w^* \rangle = \frac{1}{|G|} \sum_{\text{clean } i} w_i \underset{\substack{\uparrow \\ \text{by invariant}}}{\geq} \frac{1}{|G|} \alpha \cdot \sqrt{\sum_{\text{all } i} w_i}$ ($\clubsuit$)

so

$$\leq \underbrace{\frac{|G|}{n\alpha}}_{=1} \cdot \frac{1}{\sqrt{\sum_{\text{all } i} w_i}}$$

Finally, $\sum_{\text{all } i} w_i \geq \sum_{\text{clean } i} w_i \underset{\substack{\uparrow \\ \text{by invariant}}}{\geq} \alpha \sqrt{\sum_{\text{all } i} w_i}$,

So $\sum_{\text{all } i} w_i \geq \alpha^2$

so

$$\leq \frac{1}{\alpha}$$

II) Similarly, by Jensen's,

$$\|\hat{\mu} - \mu_w\|^2 \leq \sup_{u \in S^{d-1}} \frac{1}{\langle w, w^* \rangle} \sum_{\text{all } i} w_i w_i^* \langle x_i - \mu_w, u \rangle^2$$

$$\leq \sup_u \frac{1}{\langle w, w^* \rangle} \cdot \frac{1}{|G|} \cdot \left(\sum_{\text{all } i} w_i \right) \cdot \underbrace{\frac{1}{\sum_{\text{all } i} w_i} \sum_{\text{all } i} w_i \langle x_i - \mu_w, u \rangle^2}_{= u^T \Sigma_w u \leq \|\Sigma_w\|_{op}}$$

$$\leq \frac{1}{\alpha}\sqrt{\sum_{\text{all } i} w_i} \cdot \|\Sigma_w\|_{op},$$

by (*) from prev. page

So

$$\|\mu - \mu_w\|^2 \leq 2\|\hat{\mu} - \mu\|^2 + 2\|\hat{\mu} - \mu_w\|^2$$

$$\lesssim \frac{1}{\alpha}\left(1 + \sqrt{\sum_{\text{all } i} w_i} \cdot \|\Sigma_w\|_{op}\right)$$

as claimed. $\square$