CS 270 Algorithms **Spring 2021**

Lecture 7: Robust Mean Estimation and Tensors

Lecturer: Prasad Raghavendra   Scribe: Vishal Raman, Tor Nitayanont, Varun Menon

7.1 Robust Mean Estimation

We have a Dataset $\mathcal{D} = \mathcal{D}_{in} \cup \mathcal{D}_{out}$, where $\mathcal{D}_{in}$ are the inliers and $\mathcal{D}_{out}$ are the outliers. $\mathcal{D} = \{x_1, \dots, x_N\}$ and $\mathcal{D}_{in}$ contains $(1-\epsilon)N$ elements and an adversery has inserted ϵn outliers to $\mathcal{D}_{out}$. The goal is to estimate the mean.

Definition 7.1. *A dataset $\mathcal{D} = \{x_1, \dots, x_n\}$ is (ϵ, Δ)-stable if for any subset $\mathcal{S} \subset \mathcal{D}$ with $|\mathcal{S}| \geq (1-\epsilon)N$,*

$$\left\| \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} x_i - \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} x_i \right\| \leq \Delta$$

We assume that $\mathcal{D}_{in}$ are (ϵ, Δ)-stable. We find a subset $\mathcal{S}$ which is also (ϵ, Δ)-stable, with $|S| = (1-\epsilon)N$.
Claim: This implies that $\|\mu(S) - \mu(\mathcal{D}_{in})\| \leq 2\Delta$.

Proof. Let $T = \mathcal{S} \cap \mathcal{D}_{in}$. By (ϵ, Δ)-stability of $\mathcal{D}_{in}$, $\|\mu(\mathcal{D}_{in}) - \mu(T)\| \leq \Delta$. Similarly, by (ϵ, Δ)-stability of $\mathcal{S}$, $\|\mu(S) - \mu(T)\| \leq \Delta$. It follows that

$$\|\mu(S) - \mu(\mathcal{D}_{in})\| \leq \|\mu(\mathcal{D}_{in}) - \mu(T)\| + \|\mu(\mathcal{T}) - \mu(S)\| < 2\Delta.$$

□

We begin by defining a more general notion of stability.

Definition 7.2. *An ϵ-filtering of a set S is any set T such that $|T| \geq (1-\epsilon)|S|$.*

Definition 7.3. *A distribution θ_{in} is an ϵ-filtering of a distribution θ over $\mathbb{R}^n$ if for all $x \in \mathbb{R}^n$,*

$$\theta_{in}(x) \leq \theta(x)(1+\epsilon).$$

Definition 7.4. *A distribution θ is (ϵ, Δ)-stable if for all ϵ-filterings θ_{in},*

$$\|\mu(\theta) - \mu(\theta_{in})\| \leq \Delta.$$

Lemma 7.5. *A distribution θ over $\mathbb{R}^n$ is (ϵ, Δ) stable for*

$$\Delta = \sqrt{\epsilon \|cov[\theta]\|_{op}},$$

where the operator norm $\|cov[\theta]\|_{op}$ is the largest eigenvalue of the covariance matrix.

Proof. Suppose θ_{in} is an ϵ-filtering of θ. It suffices to show that $\|\mu(\theta_{in}) - \mu(\theta)\| \leq \Delta$.

We first write $\theta = (1-\gamma)\theta_{in} + \gamma\theta_{out}$ where $\gamma = \frac{\epsilon}{1+\epsilon}$, hence, we have

$$\theta_{out}(x) = \frac{(1+\epsilon)\theta(x) - \theta_{in}(x)}{\epsilon}.$$

Then,

$$\mu(\theta) = (1-\gamma)\mu(\theta_{in}) + \gamma\mu(\theta_{out}),$$

and

$$\text{cov}[\theta] = (1-\gamma)^2\text{cov}[\theta_{in}] + \gamma^2\text{cov}[\theta_{out}] + 2\gamma(1-\gamma)(\mu[\theta_{in}] - \mu[\theta_{out}])(\mu[\theta_{in}] - \mu[\theta_{out}])^\intercal.$$

In the second equation, applying $v^\intercal v$ for $v = \frac{\mu[\theta_{in}] - \mu[\theta_{out}]}{\|\mu[\theta_{in}] - \mu[\theta_{out}]\|}$, we find that the left hand side is

$$v^T \text{cov}[\theta] v \leq \|\text{cov}[\theta]\|_{op},$$

and the right hand side is

$$v^\intercal(1-\gamma^2)\text{cov}[\theta_{in}]v + v^\intercal\gamma^2\text{cov}[\theta_{out}]v + 2\gamma(1-\gamma)\|\mu[\theta_{in}] - \mu[\theta_{out}]\|^2 \geq 2\gamma(1-\gamma)\|\mu[\theta_{in}] - \mu[\theta_{out}]\|^2,$$

where the inequality follows from the fact that the Covariances are positive-semidefinite.

It follows that

$$\|\mu[\theta_{in}] - \mu[\theta_{out}]\| \leq \sqrt{\frac{\|\text{cov}[\theta]\|_{op}}{2\gamma(1-\gamma)}}.$$

Furthermore, we have

$$\|\mu[\theta] - \mu[\theta_{in}]\| = \|\gamma(\mu[\theta_{in}] - \mu[\theta_{out}])\| \leq \sqrt{\frac{\gamma\|\text{cov}[\theta]\|_{op}}{2(1-\gamma)}} = \sqrt{\epsilon\|\text{cov}[\theta]\|_{op}},$$

substituting $\epsilon = \frac{\gamma}{2(1-\gamma)}$. $\square$

Now, given the dataset $\mathcal{D} = \{x_1, \ldots, x_N\}$, we wish to find weights $w_1, \ldots, w_N$ so that the following hold:

- $0 \le w_i \le \frac{1+\epsilon}{N}$ $\quad$ ($w = (w_1, ..., w_n)$ is an ϵ-filtering of $\mathcal{D}$),
- $\sum w_i = 1$,
- $\mathrm{cov}[w] = \sum w_i(x_i - \mu(w))(x_i - \mu(w))^\intercal < \lambda \cdot I_d$ where $\mu(w) = \sum w_i x_i$

Consider the weights $w^* = (w_1^*, ..., w_n^*)$ where

$$w_i^* = \begin{cases} 1/|\mathcal{D}_{in}| & \text{if } x_i \in \mathcal{D}_{in} \\ 0 & \text{else} \end{cases} \quad .$$

We can see that w^* satisfies all the conditions above.

Next, We find $\{w_i\}$ via the ellipsoid algorithm. Notice that the w_i lie within the n-simplex. Given a point w, note that it satisfies

$$\|\sum w_i(x_i - \mu(w))(x_i - \mu(w))^\intercal\|_{op} < \lambda.$$

Suppose $\|\mathrm{cov}[w]\|_{op} = \lambda$ for some λ and let v be the top eigenvector: $v^T \mathrm{cov}[w] v = \lambda$. Define a linear function

$$L[y] = v^\intercal \left(\sum y_i[x_i - \mu(w)][x_i - \mu(w)]^\intercal\right) v.$$

Then, $L[w] = \lambda$ so if we show $L[w^*] < O(\epsilon\lambda)$, it follows that $L[y] \le \lambda$.

Lemma 7.6. $L[w^*] = v^T \left(\sum w^*[x_i - \mu(w)][x_i - \mu(w)]^\intercal\right) v \le O(\epsilon\lambda)$.

Proof. We have the following identity:

$$\begin{aligned} L[w^*] &= v^\intercal \left(\sum w_i(x_i - \mu[w^*])(x_i - \mu[w^*])^\intercal\right) v + v^T((\mu(w^*) - \mu(w))(\mu(w^*) - \mu(w))^\intercal) v \\ &\le v^\intercal \mathrm{cov}[w^*] v + \|\mu(w^*) - \mu(w)\|^2 \end{aligned}$$

The first term, $v^\intercal \mathrm{cov}[w^*] v$, is $O(1)$. Hence,

$$\begin{aligned} \|\mu(w^*) - \mu(w)\|^2 &\le 2\|\mu(w^*) - \mu(w \cap \mathcal{D}_{in})\|^2 + 2\|\mu(w \cap \mathcal{D}_{in}) - \mu(w)\|^2 \\ &\le O(\epsilon\lambda) + O(\epsilon\lambda) = O(\epsilon\lambda). \end{aligned}$$

Note that $w \cap \mathcal{D}_{in}$ is an $O(\epsilon)$-filtering of w^* and w. $\square$

7.2 Tensors

We set T to be a higher dimensional array of nonzero 3-dimensional tensors: $T \in \mathbb{R}^{m \times n \times p}$. For a matrix M, we can define the bilinear form

$$M(x, y) = \sum_{i,j} M_{ij} x_i y_j.$$

We can similarly define a trilinear form

$$T(x, y, z) = \sum_{i,j,k} T_{ijk} x_i y_j z_k.$$

These show up in the moments of distributions: Recall that $\mu(d) = E_{x \sim D}[x]$, and $\mathrm{cov}(D)[(x - \mu)(x - \mu)^\intercal]$. We could define higher moments,

$$T_{ijk} = E[(x_i - \mu_i)(x_j - \mu_j)(x_k - \mu_k)] \approx E[x_i x_j x_k].$$

In other words, we use tensors to encode higher order correlations of random variables.