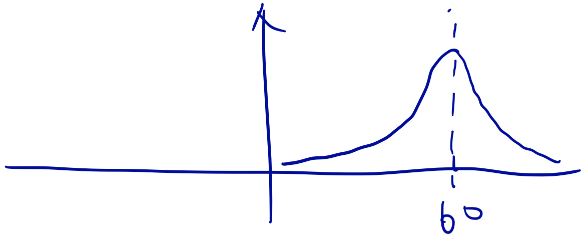
CS541 Artificial Intelligence
Guest Lecture on Mean Estimation

Lecturer: Shiwei Zeng

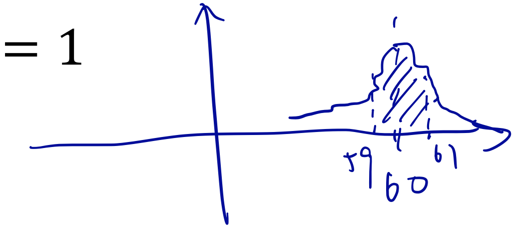
szeng4@stevens.edu

Estimating Average Height

- Assume $D = N(60,1)$



- Assume $\mathrm{E}[D] = 60, \mathrm{Var}[D] = 1$



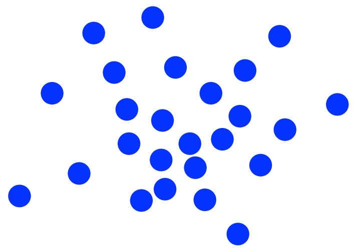
- Estimator $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\quad S \sim D^m$

$$E[\hat{\mu}] = E_{S\sim D^m}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} E[x_i] = 60 = \text{ground truth.}$$

$$\mathrm{Var}_{S\sim D^m}[\hat{\mu}] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[x_i] = \frac{1}{n^2}\cdot n = \frac{1}{n} \qquad n \nearrow$$

ME in Higher Dimension



$\boldsymbol{x} \in \mathbb{R}^{\boldsymbol{n}}$

D

$$\mathbf{E}[\boldsymbol{D}] = ?$$

When Data is Noisy

Adversary : corrupt. ε- fraction, $\varepsilon < \frac{1}{2}$

Total variation distance D_1, D_2

$\frac{1}{2}\int |\phi_1 - \phi_2| dx = \frac{\varepsilon}{1-\varepsilon}$

- 1-dimensional: (a lower bound)

$O(\varepsilon)$

$\phi_1 \quad D_1 = N(\mu_1, 1) \qquad D_2 = N(\mu_2, 1) \quad \phi_2$

$|\mu_1 - \mu_2| \geq \Omega(\varepsilon)$

$Q_1, Q_2 \qquad D_\varepsilon = (1-\varepsilon) D_1 + \varepsilon \cdot Q_1 = (1-\varepsilon) D_2 + \varepsilon Q_2$

$$Q_1 = \frac{1-\varepsilon}{\varepsilon}(\phi_2 - \phi_1) \cdot \mathbb{1}_{\phi_2 \geq \phi_1}$$

$$Q_2 = \frac{1-\varepsilon}{\varepsilon}(\phi_1 - \phi_2) \cdot \mathbb{1}_{\phi_1 \geq \phi_2}$$

Verify: $q_1 = (1-\varepsilon)\cdot\phi_1 + \cancel{\varepsilon} \cdot \frac{1-\varepsilon}{\cancel{\varepsilon}}(\phi_2 - \phi_1)\cdot \mathbb{1}_{\phi_2 \geq \phi_1}$

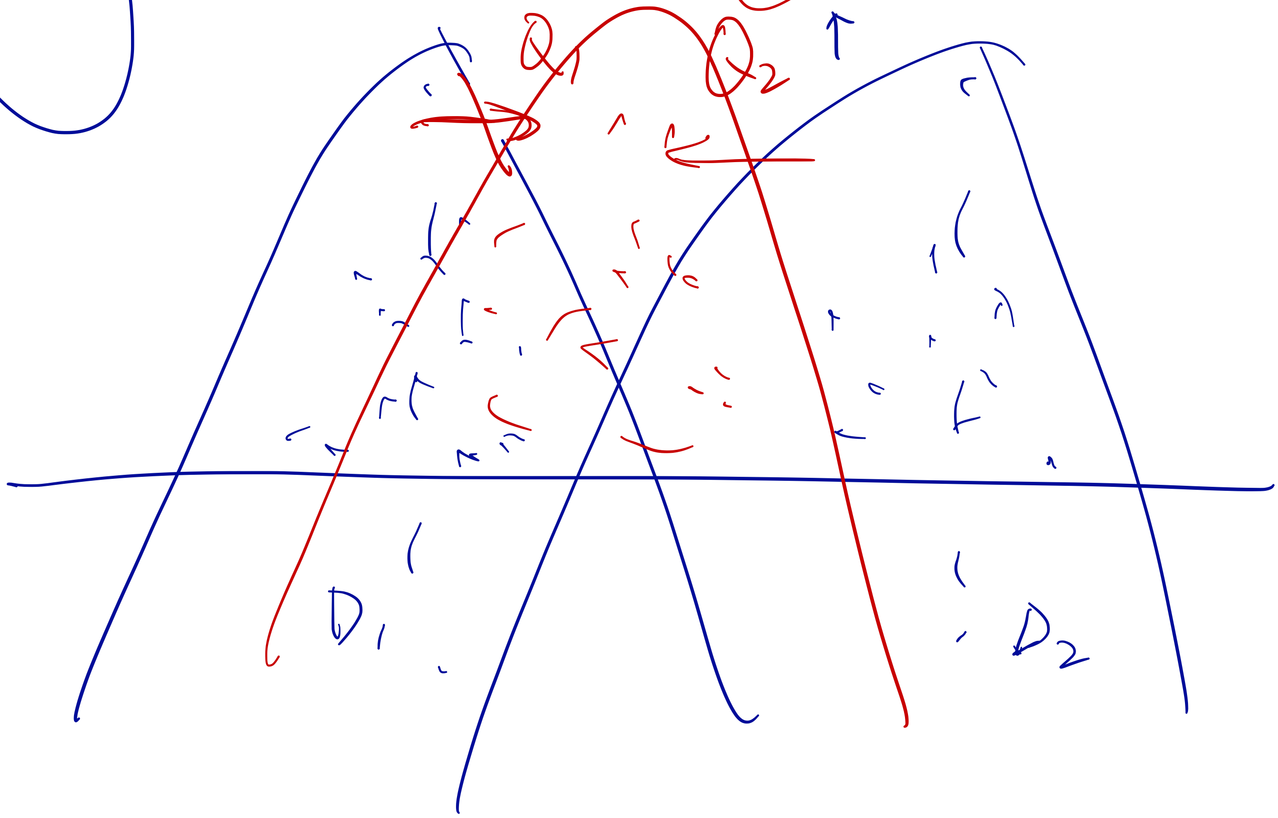
$$= \begin{cases} (1-\varepsilon)\cdot\phi_2 & \phi_2 \geq \phi_1 \\ (1-\varepsilon)\cdot\phi_1 & \phi_2 < \phi_1 \end{cases}$$

$q_2 = (1-\varepsilon)\phi_2 + \varepsilon \frac{1-\varepsilon}{\varepsilon}(\phi_1 - \phi_2 \ldots$

$= \{$

$$D_\epsilon = (1-\epsilon) D_1 + \epsilon Q_1$$

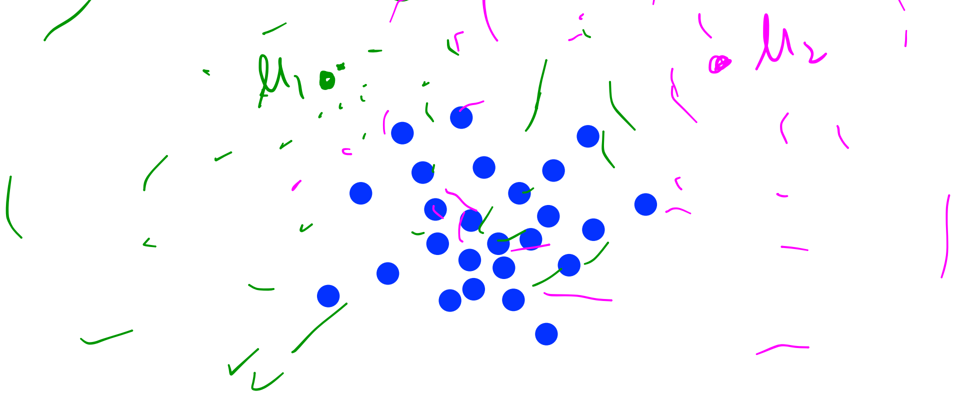
$$D_\epsilon = (1-\epsilon) D_2 + \epsilon Q_2$$



Robust Mean Estimation

$[\frac{1}{2}, 1)$

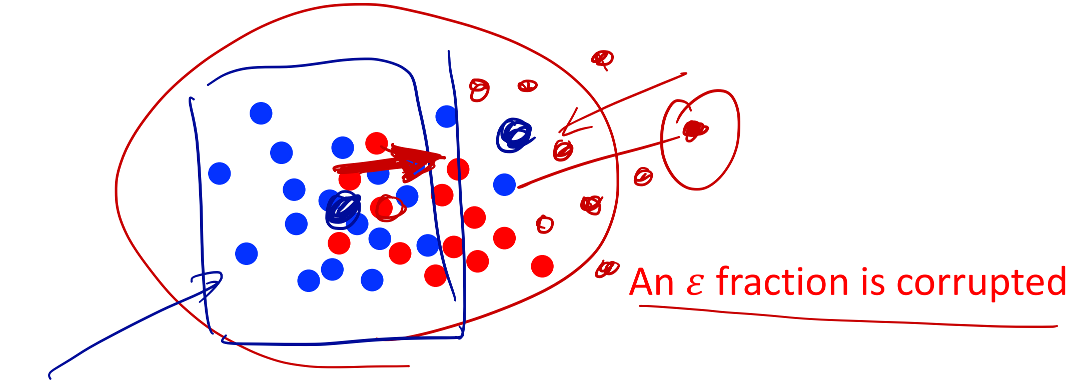
Mean Estimation



D

E[*D*] = ?

ε-robust Mean Estimation



D* + *D'

$\Omega(\varepsilon)$

bf 2016

$O(\varepsilon \cdot \sqrt{n})$

$1000 \cdot \varepsilon$

$n = 10^6$

E[*D*] = ?

$\varepsilon = (0, \frac{1}{2})$

$\|\hat{\mu} - \mu\|_2 \geq 500$

Natural approaches

- Learn each coordinate separately

$$|\hat{\mu} - \mu| \geq \Omega(\varepsilon)$$

in n-dimension : $\|\hat{\mu} - \mu\|_2^2 = \sum_{i=1}^{n} |\hat{\mu}_i - \mu_i|^2 \geq n \cdot \Omega(\varepsilon)^2 \geq \Omega(n\varepsilon^2)$

$$\underline{\Omega(\sqrt{n} \cdot \varepsilon)}$$

$$e^n$$

Natural approaches

- Maximum Likelihood Estimator

Negative Log likelihood = NLL

$$\min NLL(F, x_1, \cdots, x_m) = -\sum_{i=1}^{m} \log F(x_i)$$

$$F(x_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\|x_i-\mu\|_2^2}{2}} \leftarrow \quad Var = 1$$

$$\min -\sum_{i=1}^{m} \log\left(\frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{\|x_i-\mu\|_2^2}{2}}\right)$$

$$\Rightarrow \min_{\mu \in \mathbb{R}^n} -\sum_{i=1}^{m}\left(\underline{\log\left(\frac{1}{\sqrt{2\pi}}\right)} - \frac{\|x_i-\mu\|_2^2}{2}\right)$$

$$\Rightarrow \arg\min \frac{1}{2}\sum_{i=1}^{m} \|x_i-\mu\|_2^2 = \hat{\mu} \leftarrow$$ empirical mean.

$\downarrow$

Can be quite bad.

Alg is not robust.

~ 7:30

Efficient Algorithm – Convex Programming

weight vector $\boxed{\hat{w}} = (w_1, w_2, \cdots, w_m)$

Goal: output $\hat{w}$, $\sum_{i=1}^{m} \hat{w}_i \cdot x_i = \hat{\mu} \rightarrow \mu$.

min empirical variance.

s.t. $\hat{w} \in \mathcal{W}$

$O(n^6)$

$p(x) = x \cdot v^*$

$O(\sqrt{n})$ $\quad D = N(\mu, I_n)$

Efficient Robust Mean Estimation - Filter

1. Compute empirical mean and covariance μ_T, Σ_T $\quad T$: corrupted data set.
2. Compute largest eigenvalue λ^* of $\Sigma_T - I$, and eigenvector v^*
3. If λ^* is small, return μ_T $\quad \lambda^* \Sigma = v^* -$
4. Otherwise, find $t > C_1$ such that $\leq$

$$\Pr_{X \in T}[|v^* \cdot (X - \mu_T)| > t] > C_2 e^{-t^2/2} + \frac{C_3 \varepsilon}{t^2 \log(n \log \frac{n}{\varepsilon \tau})}$$

5. Remove X such that $|v^* \cdot (X - \mu_T)| > t$, go back to step 1.

$O(\varepsilon)$

λ^* : eigenvalue $\longrightarrow$ variance.

$$\lambda^* v^* = v^* \Sigma_T$$

$$\boxed{\operatorname*{Var}_{x \sim N(0,I)}[x \cdot v^*]} = E[(x \cdot v^*)^2]$$

$$= \mathop{E}_{x \sim D}\left[(v^{*T} x)(x^T \cdot v^*)\right]$$

$$= v^* \cdot \underbrace{E[x x^T]}_{= \Sigma_T} v^*$$

$$= v^* \cdot \Sigma_T \cdot v^*$$

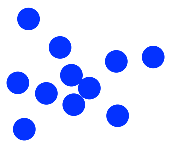
$$= \lambda^* \underbrace{v^* \cdot v^*}_{= 1}$$

$$\boxed{= \lambda^*}$$

$\varepsilon \geq \frac{1}{2}$ $\alpha = 1 - \varepsilon$

List-decodable Mean Estimation

Mean Estimation



D

$$E[D] = ?$$

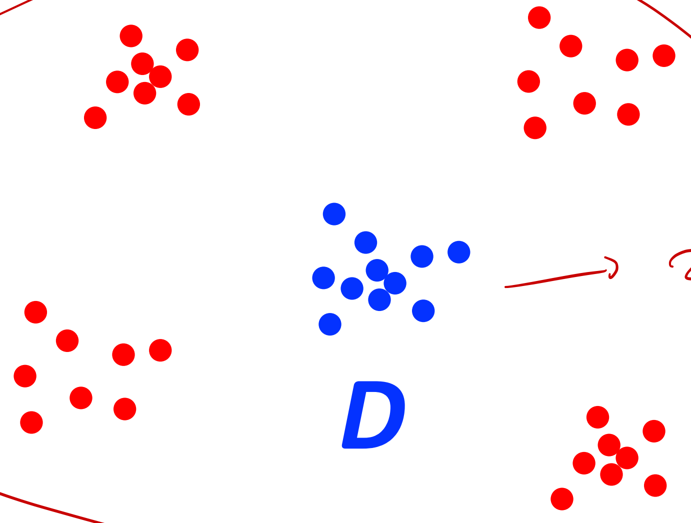
List-Decodable Mean Estimation

$10^6 \times 20\%$

$= 0.2 \times 10^6$

200000

T



$\rightarrow 20\% = \alpha$

D

$$E[D] \in \{\mu_1, \ldots, \mu_m\}$$

Gaussian Annulus Theorem.

$\sqrt{n}$.

Algorithm: Multi-filtering

- A tree of subsets T_i's,   ① Clustering

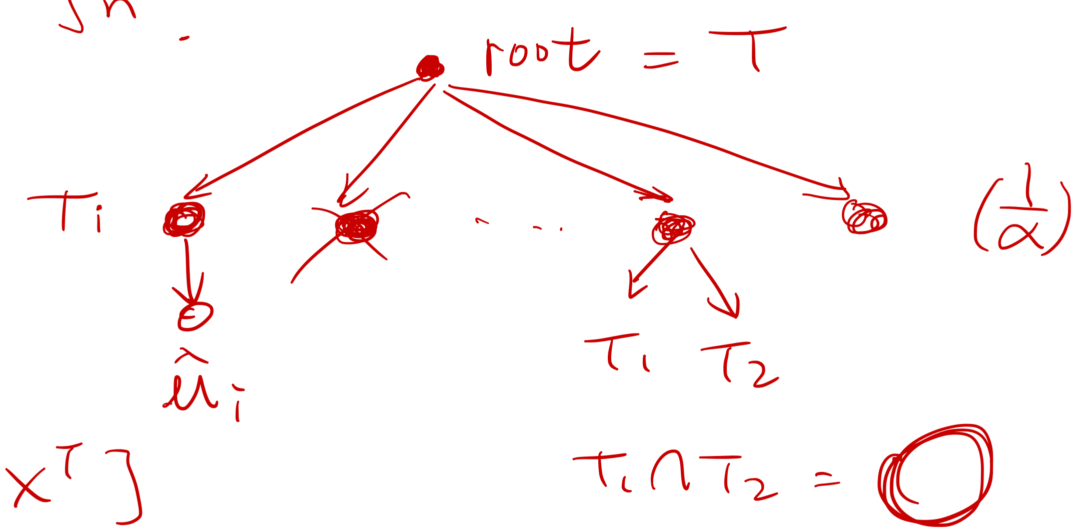
x_6^6

$E[xx^T]$

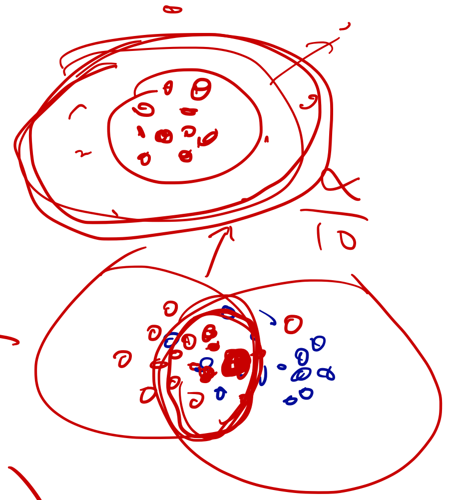
② - Iterate through each node x_1^2 $x_1x_2x_3$
    (1) Create a leaf node, an estimate $\hat{\mu}_i$
    (2) Create child nodes, subsets T_i's
        a. One node, cleaner set
        b. Two nodes, overlapping subsets
    (3) Delete if it can't be α-good.

x_1x_1 x_1x_2 x_1x_3 ...

$O(\log\frac{1}{\alpha})$



α-fraction



- No more filtering, then return all $\hat{\mu}_i$'s

$O\left(\frac{1}{\alpha^{\frac{1}{2d}}}\right)$ $d \in \mathbb{N}^+$, d: degree of polynomial

$err = \min_i \|\hat{\mu}_i - \mu\|_2 \le O\left(\frac{1}{\sqrt{\alpha}}\right)$

T_i: $\alpha T_i \to$ good samples.