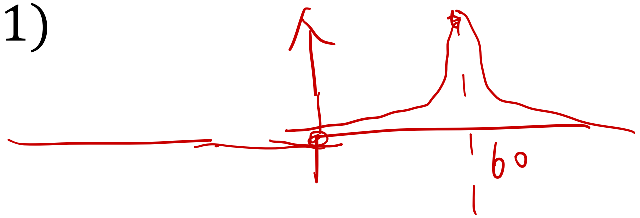
CS541 Artificial Intelligence
Guest Lecture on Mean Estimation

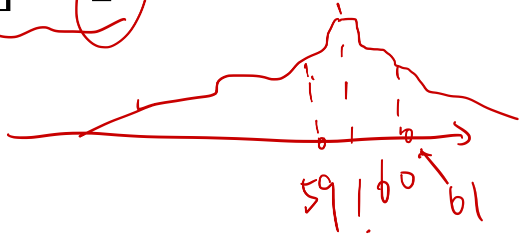
Lecturer: Shiwei Zeng

Estimating Average Height

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



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



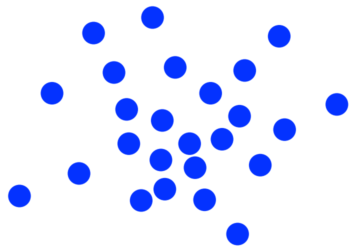
- Estimator $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$x_i \sim D$ i.i.d.

$$E[\hat{\mu}] = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[x_i] = \mu$$

$$\mathrm{Var}[\hat{\mu}] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}[x_i] = \frac{1}{n}$$

ME in Higher Dimension



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

$\boldsymbol{D}$

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

When Data is Noisy

$O(\cdot)$

$\varepsilon < \frac{1}{2}$ $\quad \varepsilon = 30\%$

$\varepsilon \cdot c \quad c > 0$

- 1-dimensional: (a lower bound)

$\|\hat{\mu} - \mu\|_2^2 \overset{\text{"="}}{\geq} \Omega(\varepsilon)$ if corrupt ε-fraction.

$D_1 = N(\mu_1, 1) \quad D_2 = N(\mu_2, 1)$ construct Q_1 and Q_2 s.t.

$\|\mu_1 - \mu_2\| = \Omega(\varepsilon)$ and

$$\boxed{D_\varepsilon} = (1-\varepsilon) D_1 + \varepsilon Q_1 = (1-\varepsilon) D_2 + \varepsilon Q_2 \qquad (1)$$

Let ϕ_1 be pdf of D_1, ϕ_2: pdf of D_2. Let μ_1, μ_2 be s.t. the total variance distance between D_1, D_2 is

$$\frac{1}{2} \int |\phi_1 - \phi_2| \, dx = \frac{\varepsilon}{1-\varepsilon} \implies \|\mu_1 - \mu_2\| \geq \frac{2\varepsilon}{1-\varepsilon}.$$

Q_1: $\frac{1-\varepsilon}{\varepsilon} (\phi_2 - \phi_1) \cdot \mathbb{1}_{\phi_2 \geq \phi_1}$ and Q_2: $\frac{1-\varepsilon}{\varepsilon} (\phi_1 - \phi_2) \mathbb{1}_{\phi_1 \geq \phi_2}$.

$D_1 = N(\mu_1, 1) \quad D_2 \; N(\mu_2, 1)$

$\|\mu_1 - \mu_2\| = \Omega(\varepsilon)$ and

$$\boxed{D_\varepsilon} = (1-\varepsilon) D_1 + \varepsilon Q_1 = (1-\varepsilon) D_2 + \varepsilon Q_2 \qquad (1)$$

Let ϕ_1 be pdf of D_1, ϕ_2: pdf of D_2. Let μ_1, μ_2 be s.t. the total variance distance between D_1, D_2 is

$$\frac{1}{2}\int |\phi_1 - \phi_2| dx = \frac{\varepsilon}{1-\varepsilon} \implies \|\mu_1 - \mu_2\| \geq \frac{2\varepsilon}{1-\varepsilon} \geq \Omega(\varepsilon)$$

$4\varepsilon \geq \quad \geq 2\varepsilon$

$\varepsilon \in (\frac{1}{2}, 1)$

$Q_1 : \frac{1-\varepsilon}{\varepsilon}(\phi_2 - \phi_1) \cdot 1_{\phi_2 \geq \phi_1}$ and $Q_2 : \frac{1-\varepsilon}{\varepsilon}(\phi_1 - \phi_2) 1_{\phi_1 \geq \phi_2}$.

$$(1-\varepsilon)\phi_1 + \varepsilon \cdot \frac{1-\varepsilon}{\varepsilon}(\phi_2 - \phi_1) \cdot 1_{\phi_2 \geq \phi_1}$$

$$1_{\phi_2 \geq \phi_1} = \begin{cases} 1 & \phi_2 \geq \phi_1 \\ 0 & \phi_2 < \phi_1 \end{cases}$$

$$= (1-\varepsilon)\phi_1 + (1-\varepsilon)(\phi_2 - \phi_1) \cdot 1_{\phi_2 \geq \phi_1}$$

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

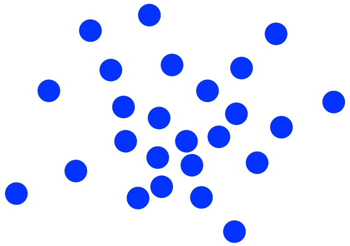
$$(1-\varepsilon)\phi_2 + \varepsilon \cdot \frac{1-\varepsilon}{\varepsilon}(\phi_1 - \phi_2) 1_{\phi_1 \geq \phi_2}$$

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

$\phi_1 = \phi_2$

Robust Mean Estimation

Mean Estimation

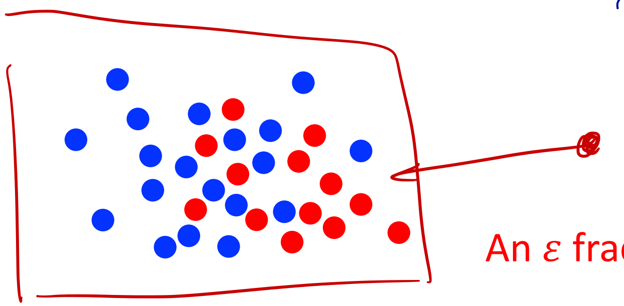


D

E[*D*] = ?

ε-robust Mean Estimation

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



An ε fraction is corrupted

D* + *D'

E[*D*] = ?

Natural approaches

- Learn each coordinate separately

$\hat{\mu}$ s.t. $\|\hat{\mu}_i - \mu_i\| \geq \Omega(\varepsilon)$.

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

$\|\hat{\mu} - \mu\|_2 \geq \sqrt{n} \cdot \varepsilon$

$n = 10^6$

$\sqrt{n} \cdot \varepsilon = 10^3 \cdot \varepsilon \qquad \varepsilon\,(0, \tfrac{1}{2})$

Natural approaches

e^n

- Maximum Likelihood Estimator

Negative Log Likelihood

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

Assume F is Gaussian : $F(x_i) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{\|x_i - \mu\|_2^2}{2}}$

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

$$= \min_{\mu} \boxed{-N \cdot \log \frac{1}{\sqrt{2\pi}}} + \sum_{i=1}^{m} + \frac{\|x_i - \mu\|_2^2}{2}$$

$$\Longrightarrow \min_{\mu} \underline{\frac{1}{2} \sum_{i=1}^{m} \|x_i - \mu\|_2^2}$$

$\hat{\mu}$: empirical mean

$= \frac{1}{m} \sum_{i=1}^{m} x_i$

Efficient Algorithm – Convex Programming

~ 7:30

Output a $\hat{w} = (w_1, w_2, \cdots w_m)$ such that

$$\hat{\mu} = \frac{1}{m} \sum_{i=1}^{m} w_i x_i \quad \text{is close to } \mu .$$

by solving a convex program.

$O(n^6)$

$D = N(\underline{\mu}, I)$ $\forall v \in \mathbb{R}^n$ $Pr_{x\sim D}[|v \cdot (x-\mu)| \geq t] \leq \exp(-\frac{t^2}{2})$ ☆ $\varepsilon(0, \frac{1}{2})$ $\tau = 0.01$ 0.99

Efficient Robust Mean Estimation - Filter

T: corrupted dataset.

$Pr_{x\sim D}[|p(x) - E[p(x)]| \geq t] \leq \boxed{}$ $\frac{1}{m}\sum_{i=1}^{m} x_i$ $\frac{1}{m}\sum_{i=1}^{m}(x_i-\mu_T)(x_i-\mu_T)^T$

1. Compute empirical mean and covariance $\mu_T, \Sigma_T = I$
2. Compute largest eigenvalue λ^* of $\boxed{\Sigma_T - I}_{=0}$, and eigenvector v^*
3. If λ^* is small, return μ_T
4. Otherwise, find $t > C_1$ such that $C_1, C_2, C_3 > 0$ τ

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

$= \delta$ dependent on m !

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

$\sqrt{n}$, n-dim

$\hat{\mu} \to \mu$

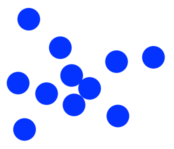
$\|\hat{\mu} - \mu\|_2 \leq O(\varepsilon) \leq 1$

t t $\hat{\mu}$

List-decodable Mean Estimation

T : corrupted data set.

Mean Estimation

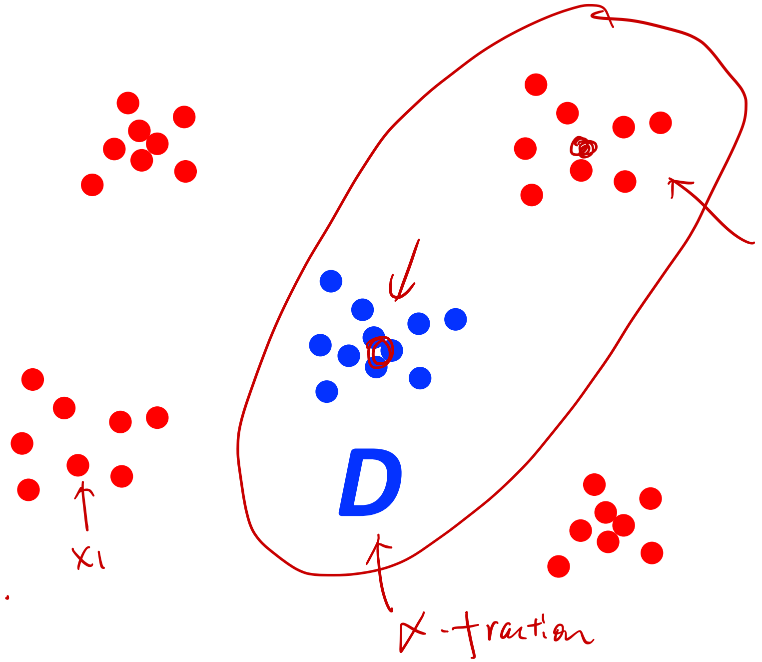


D

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

List-Decodable Mean Estimation

α : fraction of clean samples.



k golden queries

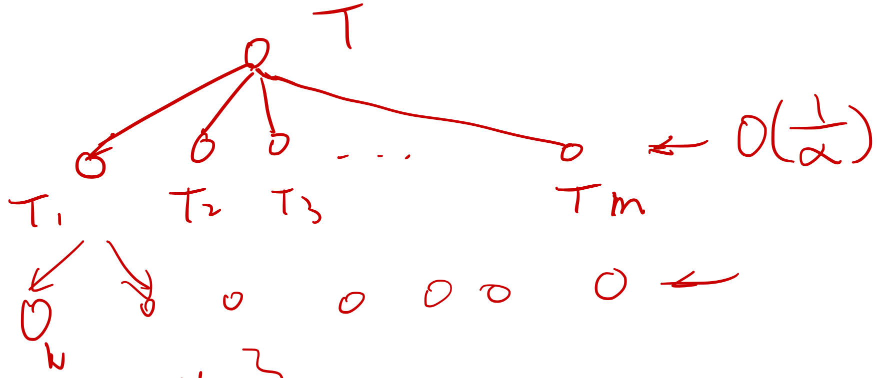
→ a list of e^k estimated meas.

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

log(m) queries → $m = O\left(\frac{1}{\alpha}\right)$

T_i is α-good: α-fraction of T_i are clean.

Algorithm: Multi-filtering



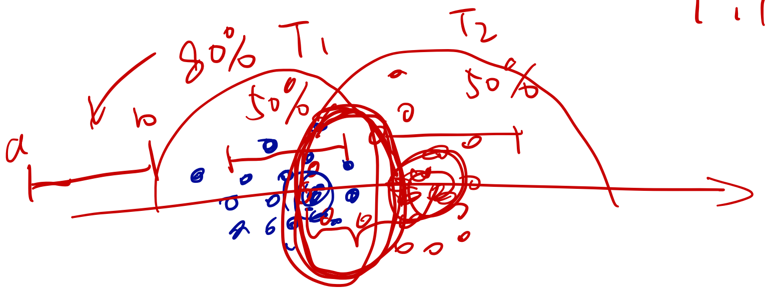
- A tree of subsets T_i's,

list ← $\{\mu_1 \dots \mu_m\}$

$T \to T_1, T_2$

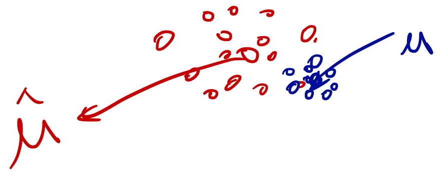
- Iterate through each node
  (1) Create a leaf node, an estimate $\hat{\mu}_i$ ←
  (2) Create child nodes, subsets T_i's
    a. One node, cleaner set
    b. Two nodes, overlapping subsets
  (3) Delete if it can't be α-good.

$T_1 \cap T_2$ = ball with radius $\sqrt{n}$,



α-good T_i: α-fraction of T_i.

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



$O\left(\frac{1}{\alpha^{\frac{1}{2}}}\right)$

↓

$\text{poly}\left(\frac{1}{\alpha}\right) \Longrightarrow O\left(\frac{1}{\alpha^{\frac{1}{2d}}}\right)$ d: degree

$\min_{i \in \{\mu_1 \dots \mu_m\}} \|\hat{\mu}_i - \mu\|_2 = O\left(\frac{1}{\sqrt{\alpha}}\right) = 1000$

$\alpha = 10^{-6}$