Lecture 24

Adaptive Data Analysis. (ADA)

- Reusable Holdout
- Sparse Vector Mechanism (SVM)
  - ↳ Application: Synthetic Data for ADA.

Logistics

Today is the last day of "me" lecturing.

Weds : Additional Office Hour for Projects.

Fri : Continue office Hour

Next Week : Project Presentation

Schedule also on Canvas

— May 3rd

Ryan Steed

Shengyuan Hu

Justin Whitehouse

Tianshi Li

— May 5th

Charlie Hou

Zhili Feng

Shuaiqi Wang

Format :

Expect Slides

20 minutes including Q&A.

Final Write-up due later.

Model for Adaptive Data Analysis

Statistical / Linear Queries

$\phi : \mathcal{X} \longrightarrow [0,1]$ "predicate"

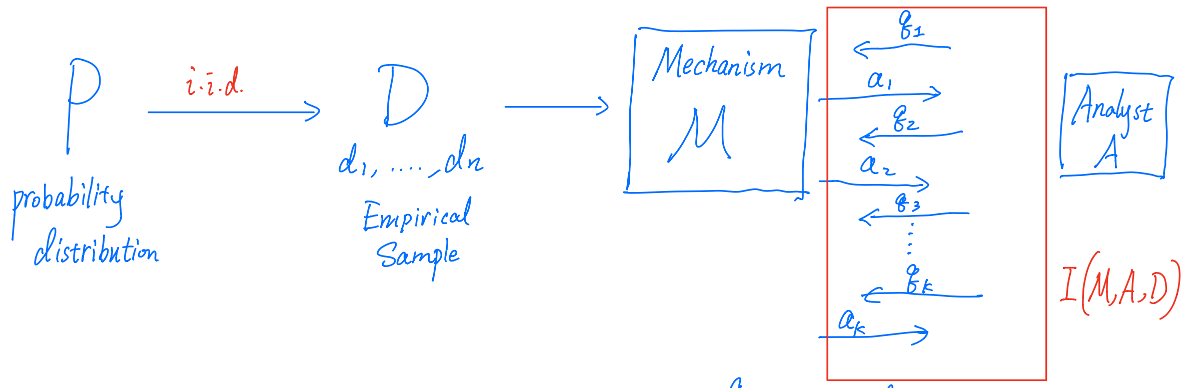
$q_\phi(P) = \mathbb{E}_{x \sim P}[\phi(x)]$ "Population value"

population

$q_\phi(D) = \hat{\mathbb{E}}_{d_i \sim D}[\phi(d_i)] = \frac{1}{n}\sum_{i=1}^{n}[\phi(d_i)]$ "Empirical Average"

$D = (d_1, \ldots, d_n) \in \mathcal{X}^n$

Interaction of Adaptive Data Analysis.



q_j can depend on

$(q_1 \quad , q_2, a_2, \dots, q_{j-1}, a_{j-1})$.

Transfer Theorem. (ε,δ)-version [JLNRSS20]

Suppose $I(M,A,D)$ is (α,β)-sample accurate

& (ε,δ)-DP.

Then for every $c,d>0$, $I(M,A,D)$ is (α',β') distributionally accurate, for

$$\alpha' = \alpha + \underbrace{(e^{\varepsilon}-1)}_{\approx\,\varepsilon} + \underbrace{c}_{\alpha} + 2\underbrace{d}_{\varepsilon}, \quad \beta' = \frac{\beta}{c} + \frac{\delta}{d}$$

$$\alpha' = O(\alpha+\varepsilon) \qquad \beta' = \left(\frac{\beta}{\alpha} + \frac{\delta}{\varepsilon}\right)$$

Answering k adaptive queries.

Gaussian Mechanism V.S. Sample Splitting

$\alpha' \lesssim \sqrt{\frac{k^{\frac{1}{2}}}{n}}$ $\qquad$ $\alpha' \le \sqrt{\frac{k}{n}}$

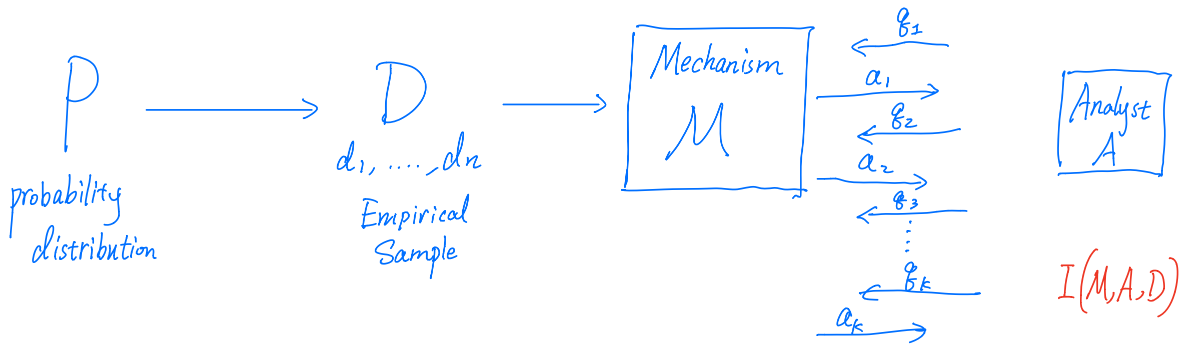
Can answer $k \approx O(n^2)$ queries $\qquad$ $k \le O(n)$.

Can we do better?

→ Output Compression. (Sparse Vector)

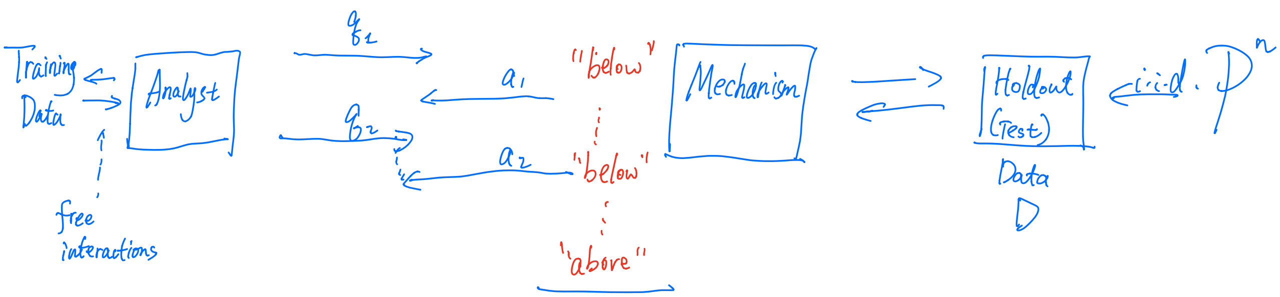
→ SV + Private Multiplicative Weights (Synthetic Data Method)

Model for ADA



q_j can depend on

$(q_1, a_1, q_2, a_2, \ldots, q_{j-1}, a_{j-1})$.



Example: $q_1(D) = \frac{1}{n} \sum_{i=1}^{n} \left[\mathbb{1}[h(x_i) = y_i] \right]$

Reusable Holdout.

Sparse Vector

Input: an adaptive sequence $q_1, q_2, \ldots$ (Δ-sensitive)

Dataset D. Threshold T.

Noisy Threshold: $\tilde{T} = T + Z_0$, $Z_0 \sim Lap\left(\frac{2\Delta}{\varepsilon}\right)$

For every query i:

$\tilde{a}_i = q_i(D) + Z_i$, $Z_i \sim Lap\left(\frac{4\Delta}{\varepsilon}\right)$

If $\tilde{a}_i < \tilde{T}$, return $b_j = \perp$ "Below"

Else return $b_j = T$; Break.

"Above"; Stop.

Binary coding: $\vec{b} = (\underbrace{0\,0\,0 \ldots 0}_{\text{"Below"}}\ \underbrace{1}_{\text{"Above"}})$

Accuracy.

Given $q_1, q_2, \ldots\ldots$

Return $b_1, b_2, \ldots\ldots$

"Empirical Error"

If $b_i =$ "Below", $\max(0, \underbrace{q_i(D) - T}_{\text{Want to be small}})$.

If $b_i =$ "Above", $\max(0, \underbrace{T - q_i(D)}_{\text{Want small.}})$

Theorem. With probability $1-\beta$, when run over a sequence of k queries. SV has empirical error

$$\alpha \le \frac{6 \cdot \Delta \ \ln\left(\frac{k+1}{\beta}\right)}{\varepsilon}$$

Error scales logarithmicaly in k.

Proof Sketch. Union Bound Laplace Noise $|z_i|$.

Composing m runs of SV.

Population error. $\max(0, \underbrace{q_i(P) - T}_{\text{Want to be small}})$. for "Below"

$\max(0, \underbrace{T - q_i(P)}_{\text{Want small.}})$ for "Above"

$$\alpha' \le O\Big(\underbrace{\frac{\ln k}{n\varepsilon}}_{\substack{\text{Empirical error}\\ \text{for answering}\\ k \text{ queries in total}}} + \underbrace{\varepsilon\sqrt{m}}_{\substack{\text{Generalization error}\\ \text{for } (\varepsilon\sqrt{m},\,\delta)\text{-DP}}}\Big)$$

Transfer Theorem :

$$\le O\left(\frac{m^{\frac{1}{4}} \ln^{\frac{1}{2}} k}{\sqrt{n}}\right).$$

m : # "Above" $\longleftarrow$ "Interesting Events"

k : $\approx$ # "Below"

Privacy.

Theorem. SparseVector is $(\varepsilon, 0)$-D.P.

Proof. Sketch.

Fix any neighbors D & D',

any output $\vec{b} = (\perp)^{(k-1)}(\top)$

Given the output, $q_1, \dots, q_k$ are also fixed.

Furthermore, fix noise values

$$Z_1 = z_1,\ Z_2 = z_2,\ \dots\dots,\ Z_{k-1} = z_{k-1}.$$

Let $$g(D) = \max_{j=1}^{k-1} q_j(D) + z_j$$

The output is $\vec{b} = (\perp)^{(k-1)}(\top)$ if and only if

$$g(D) < \tilde{T} \leq q_k(D) + z_k. \quad \text{for } D$$

and $$g(D') < \tilde{T} \leq q_k(D') + z_k \quad \text{for } D'$$

The only randomness we consider:

$$(\tilde{T}, \quad Z_k).$$

For every realization (τ, z),

we construct (τ', z') :

$$\tau' = \tau + g(D') - g(D)$$

$$z' = z + g(D') - g(D) + g_k(D) - g_k(D')$$

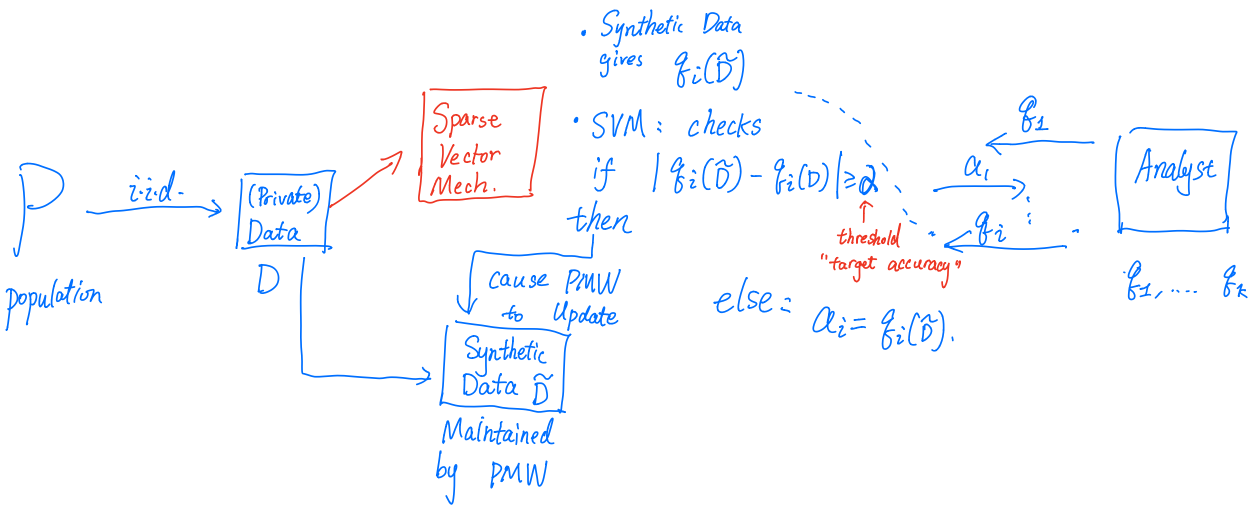
$$\mathbb{P}\left[\left(\tilde{T}, Z_k\right) = (\tau, z)\right] \le e^{\varepsilon}\, \mathbb{P}\left[\left(\tilde{T}, Z_k\right) = (\tau', z')\right]$$

$$\Downarrow \qquad\qquad\qquad\qquad \Downarrow$$

$$M(D) = \vec{b} \qquad\qquad\qquad\qquad M(D') = \vec{b}.$$

$\square$

Private Multiplicative Weights (PMW)



Hope: Answer most queries using the synthetic data.
$q_i(\tilde{D})$

PMW : adaptive analog of MWEM
(MW w/ Exp Mech)

PMW: runs in time linearly in $|\mathcal{X}|$.
$(\mathcal{X} = \{0,1\}^d)$

Sample Complexity / Accuracy.

Non-adaptive Queries.

- Take empirical averages : $a_j = q_j(D)$

$$\max_j |a_j - q_j(P)| \lesssim \sqrt{\frac{\log(k)}{n}}$$

Adaptive Queries

- Sample Splitting Method: $D_1, \dots, D_k$, $a_j = q(D_j)$

$$\max_j |a_j - q_j(P)| \leq \sqrt{\frac{k}{n}}$$

Use Transfer Theorem

- Gaussian Mechanism

$$\max_j |a_j - q_j(P)| \leq \sqrt{\frac{k^{\frac{1}{2}}}{n}}$$

- Private Multiplicative Weights + Sparse Vector

$$\underline{\qquad\qquad} \lesssim \min\left\{ \left(\frac{\ln(k)\sqrt{\ln|\mathcal{X}|}}{n} \right)^{\frac{1}{3}}, \sqrt{\frac{k^{\frac{1}{2}}}{n}} \right\}$$

Remark: · Runtime is Linear in $|\mathcal{X}|$ (Infinite or $\exp(d)$)

- The error bound depends on $\log|\mathcal{X}|$. $\approx d$.

dimension of data.

Lower Bound. Statement

" If either the algorithm is $\underbrace{\text{polytime}(d)}_{\text{efficient}}$

or $\ln|\mathcal{X}| \geq \Omega(n^2)$,

then # queries $\underline{k} \leq \Theta(n^2)$.

(before analyst comes up w/ a high-error query) "

Satisfies.

Gaussian Mech $k \approx n^2$.