Binary Tree Mechanism

- Problem formulation: gallery release
- Algorithm: How to add "smart" noise?
- Applications: quantiles, CDF, continual release.

Algorithms So far ...

Answering Numeric Queries

- Laplace Mechanism
- Randomized Response

more general ↓

less noise ↑

Focus.

Selection Problem ←

- Exponential Mechanism
- Report Noisy Max.

Query Release Problem.

Example: Census Bureau

Focus: Linear / counting queries

Given $x = (x_1, \dots, x_n) \in \mathcal{X}^n$

$f(x) = \sum_{i=1}^{n} \varphi(x_i)$ for $\varphi : \mathcal{X} \longmapsto \{0,1\}$

↑ Specify some property

e.g., (Asian, 30's, male)

Goal: Release answers $\{a_1, \dots, a_k\}$ for queries $\{f_1, \dots, f_k\}$ with accuracy α & ε-DP.

$$\max_{j=1}^{k} |a_j - f_j(x)| \leq \alpha n$$

↑ e.g., 1%

Applying Laplace Mechanism.

Define $F : \mathcal{X}^n \longmapsto \mathbb{R}^k$ change some x_i

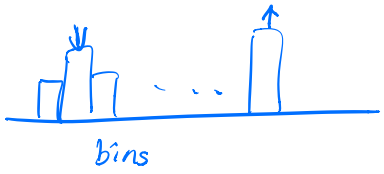
s.t. $F(x) = \begin{pmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_k(x) \end{pmatrix}$ $\leftarrow \pm 1$, $\leftarrow \pm 1$

ℓ_1 – sensitivity $GS_F := \max_{x \sim x'} \|F(x) - F(x')\|_1$

$\boxed{GS_F = k \quad \text{in worst case.}}$

Lap Mech : $f_j(x) + Z_j$, $Z_j = Lap\left(\frac{k}{\varepsilon}\right)$

max error $\leq O\left(\frac{k \log k}{\varepsilon}\right)$. $\leftarrow$ all n.

Histogram : 

Structure

$GS_F \simeq 2$.

Interval / Threshold Queries.

Given dataset $x = (x_1, \ldots, x_n) \in X$

Data Universe "Domain" $X = \{1, \ldots, D\} \equiv [D]$

↳ e.g., height, time arrival, salary, ... ($D \gg n$)

Interval Queries: given $1 \le s \le t \le D$

$$f_{s,t}(x) = \#\{i \mid s \le x_i \le t\}$$

Threshold Queries: given $t \in [D]$

$$f_t(x) = \#\{i \mid x_i \le t\}.$$

Examples:

① CDF (cumulative distribution function): $\Phi(t) = \#\{i \mid x_i \le t\} / n$

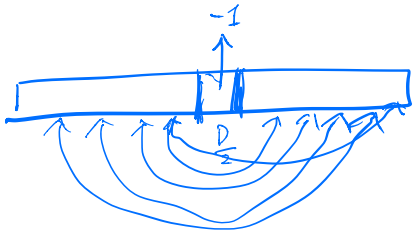
② q-quantile: threshold t such that $\Phi(t) = q$. $(\Phi^{-1}(q))$.

Baseline Laplace Mechanism

Answering all $\binom{D}{2} = O(D^2)$ interval queries

$$F(x) = \begin{pmatrix} f_{s,t}(x) \\ \vdots \end{pmatrix}_{1 \le s \le t \le D}$$

what is GS_F?



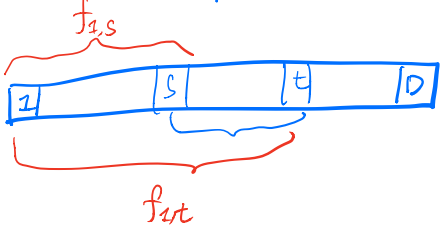
affect $O(D^2)$

$\forall (s,t)$:

$$a_{s,t} = f_{s,t}(x) + Z_{s,t}, \quad Z_{s,t} \sim Lap\left(\frac{GS_F}{\varepsilon}\right) \leftarrow O\left(\frac{D^2}{\varepsilon}\right)$$

$$\mathbb{E}\left[\max_{(s,t)} \left| a_{s,t} - f_{s,t}(x) \right|\right] = O\left(\frac{D^2 \log D}{\varepsilon}\right) \leftarrow \text{error}$$

Simple Improvement over Basic Laplace.



$$f_{s,t} = f_{1,t} - f_{1,s}$$

$f_{1,t} \leftarrow a_{1,t}$, $f_{1,s} \leftarrow a_{1,s}$.

$$\forall\ 1 \le t \le D$$

$$f_{1,t} + Z_{1,t}, \quad Z_{1,t} \sim Lap\left(\frac{D}{\varepsilon}\right)$$

$$\mathbb{E}[\text{max-error}] \le O\left(\frac{D \log D}{\varepsilon}\right) \leftarrow \text{previously } O\left(\frac{D^2 \log D}{\varepsilon}\right)$$

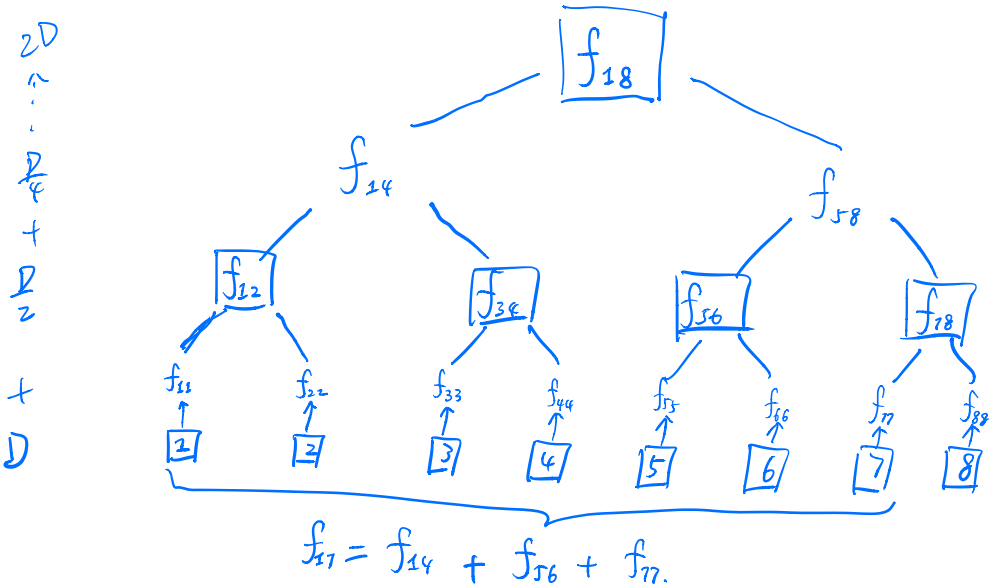
Answer an easier set of sufficient statistics.

$\rightarrow$ post-process to get answers.

Binary Tree Mechanism

error $\longrightarrow \frac{1}{\varepsilon}\log^2 D$

Wlog $D = 2^m$



$2D \approx \frac{D}{2} + \frac{D}{4} + \cdots + D$

Toy: $D = 8$

① Low Sensitivity

② Sufficient statistic.

$f_{17} = f_{14} + f_{56} + f_{77}$.

$$\mathbf{T} = \left\{ (s,t): \; s = j \cdot 2^{\ell-1} + 1 \quad \text{and} \quad t = (j+1)\cdot 2^{\ell-1} \right\}$$

for $1 \le \ell \le \log_2 D$ and $1 \le j \le D/2^{\ell-1}$.

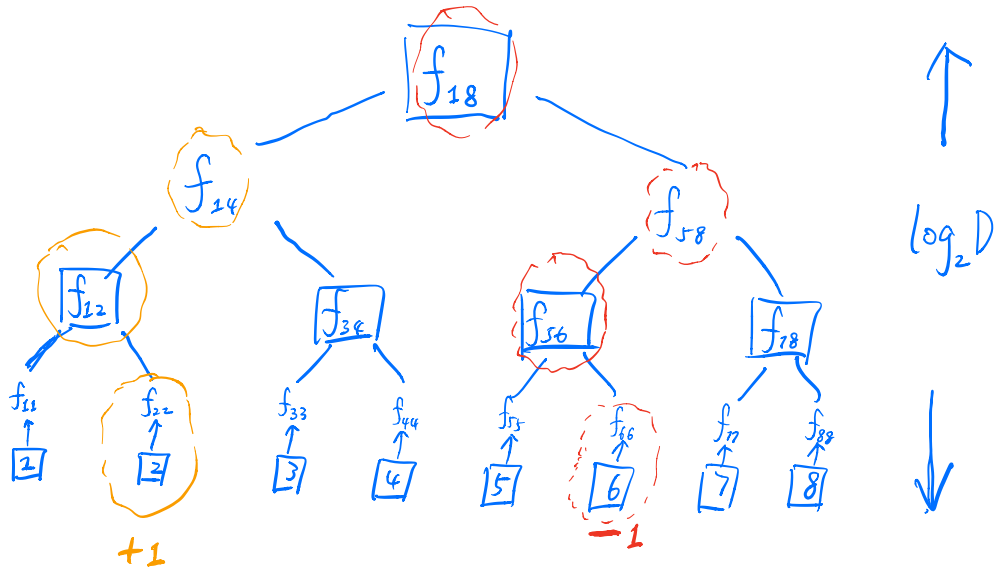
What is the sensitivity $F_{\mathbf{T}}$? $\approx \log_2 D$

$$|\mathbf{T}| = \sum_{\ell=1}^{\log_2 D} \frac{D}{2^{\ell-1}} = 2D - 1.$$

Claim: Let $F_T(x) = (f_{s,t}(x))_{(s,t)\in T}$

GS_F is at most $2\log_2 D$.

Proof.



Binary Tree Mechanism

Release tree node intervals.

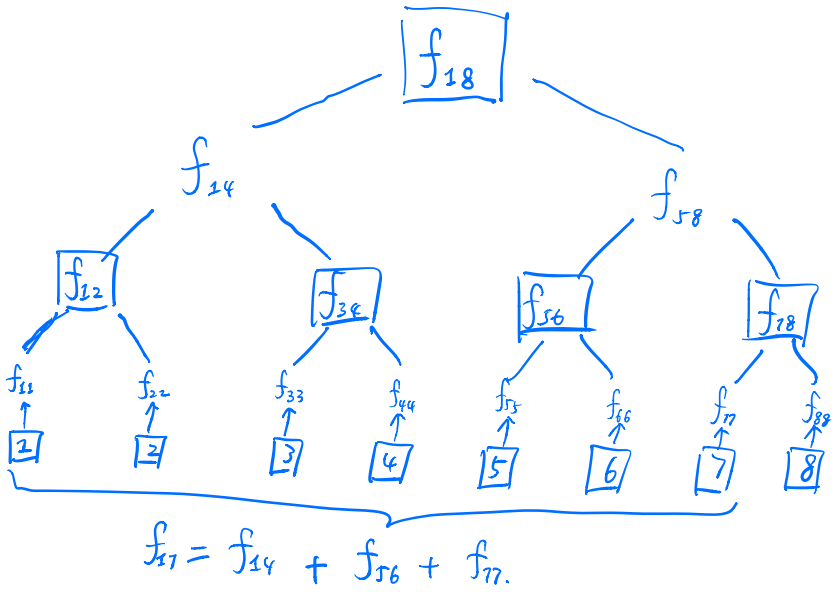
for every $(s,t) \in T$

$$a_{s,t} = f_{s,t}(x) + Z_{s,t},$$

$$Z_{s,t} \sim Lap\left(\frac{2\log_2 D}{\varepsilon}\right) \quad \text{per-coor.} \quad \varepsilon\text{-}O\left(\frac{\log D}{\varepsilon}\right)$$

$$\mathbb{E}\left[\max_{(s,t)\in T} |Z_{s,t}|\right] = O\left(\frac{\log^2(D)}{\varepsilon}\right).$$

Recovering the Answers.



Claim. For every $1 \le t \le D$, there exists $S \subseteq T$ of size $|S| \le \log_2 D$ such that

$$f_{1,t} = \sum_{(u,v) \in S} f_{u,v}. \longrightarrow \approx \sum a_{u,v}$$

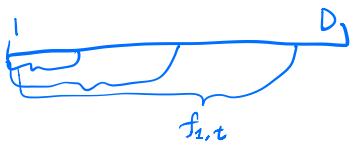
Privacy + Accuracy Guarantees.

Theorem. Bin Tree Mechanism is ε-DP and releases $(a_{s,t})_{1 \le s \le t \le D}$ such that

$$\mathbb{E}\left[\max_{(s,t)} |a_{s,t} - f_{s,t}(x)|\right] \le O\left(\frac{1}{\varepsilon} \log^3 D\right).$$

Proof Sketch: ε-DP $\Leftarrow$ Lap Mech w/ $2\text{Log}(D)$ sensitivity

Accuracy: ① Suffices to release $f_{1,t}$ $\forall t \in [D]$



② $f_{1,t} = \sum_{(u,v) \in S} f_{u,v}$

$$\mathbb{E}\left[\max_{(s,t) \in T} |a_{s,t} - f_{s,t}|\right] \le \alpha\left(\frac{\log^2 D}{\varepsilon}\right)$$

$$\mathbb{E}\left[\max_{t \in [D]} |a_{1,t} - f_{1,t}(x)|\right] \le \log D \cdot O\left(\frac{\log^2 D}{\varepsilon}\right).$$

$$= O\left(\frac{\log^3 D}{\varepsilon}\right) \qquad \square$$

Remark:

① $O\left(\frac{1}{\varepsilon} \log^{2.5} D\right)$ better bound.

② $f_{16} = f_{14} + f_{56}$

$= f_{18} - f_{78}$

Construct multiple estimates & Average.

③ $f_{1,1}, f_{1,2} \cdots f_{1,t}$ monotone.

post-processing →

Domain Size. D

Release $f_{s,t}(x) \quad \forall 1 \le s \le t \le D.$

$x_i \in \{1, \ldots, D\}$

$x_i \in [0, B]$

"Practical" solution : Discretize

$\{0, \gamma, 2\gamma, \ldots, B\}$ ← Domain Size $O\left(\frac{B}{\gamma}\right)$.

↑ Choosing γ is tricky.

"Impossibility" Result.

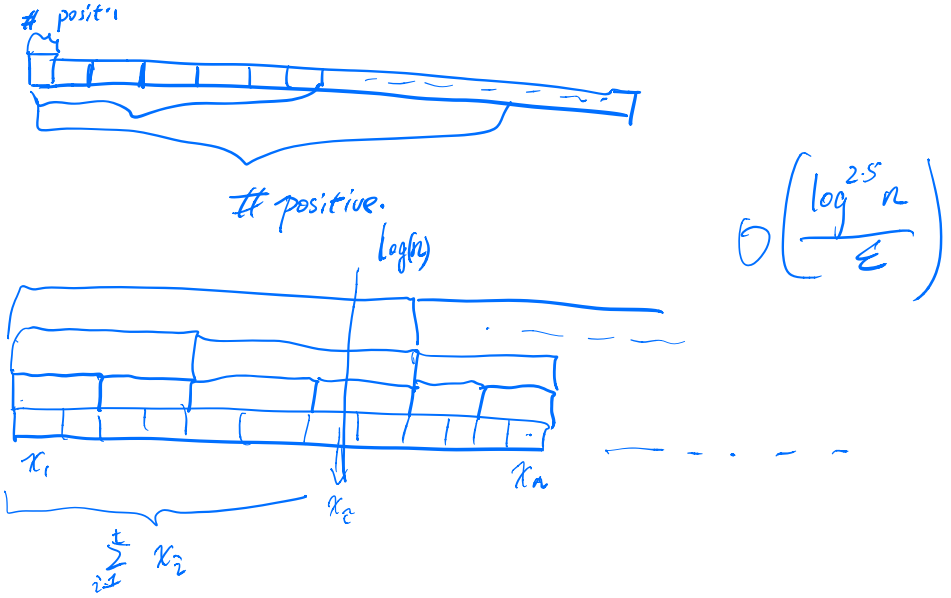
max error $> \frac{1}{\varepsilon} \log^* D.$ ← # take log so that D becomes ≤ 1

$\log(\ldots \log(D)) \le 1.$

Continual Release.

Stream: $x_1, x_2 \ldots\ldots, x_n \in \{0,1\}$

Example: Covid Tests.



$$\Theta\left(\frac{\log^{2.5} n}{\varepsilon}\right)$$