Lecture: Private Release of CDFs and the Binary Tree Mechanism

using notation from [Rameshwar, Tandon, & Sharma (2025)]

We shall continue our study of a particular "meta-statistic" derived from (and giving rise to) interval queries: the CDF. As before, consider a dataset $\underline{x} = (x_1, \dots, x_N) \in [K]^N$ and let

$$c_j := c_{j-1,j} = \#\{t : j-1 \leq x_t \leq j\}.$$

We are interested in the statistic

$$\underline{F} := \underline{F}(\underline{x}) \triangleq \frac{1}{N}.(c_1, \dots, c_K)$$

that is the CDF of $\underline{x}$.

HW: Let $\underline{c} = (c_1, \dots, c_K)$ and let $\underline{\bar{c}} \triangleq (\bar{c}_1, \dots, \bar{c}_K) = (c_{1,1}, \dots, c_{1,K})$.
Argue that $\Delta_{\underline{c}}$ and $\Delta_{\underline{\bar{c}}}$ are, resp., 2 and $K-1$, in the "replace" model.

We first discuss two simple mechanisms for releasing $\underline{F}$ privately.

① DP CDFs via range queries

We construct the vector $M_{\underline{\bar{c}}}(\underline{x}) = (M_{\underline{\bar{c}},1}, \dots, M_{\underline{\bar{c}},K})$ of estimates such that

$$M_{\underline{\bar{c}},K} = N$$

and for $1 \leq j \leq N-1$,

$$M_{\underline{c},j} = \bar{c}_j + Z_j,$$

where $Z_j \overset{iid}{\sim} \mathrm{Lap}\left(\frac{K-1}{\epsilon}\right)$. Our mechanism then returns $\underline{F}^{range}(\underline{x}) = \frac{1}{N} \cdot M_{\underline{c}}$.

__Lemma 1:__ $\underline{F}^{range}$ is ϵ-DP. [Why?]

Our interest is in evaluating the utility of an estimate $\hat{\underline{F}}$ of $\underline{F}$ via the ℓ_2-error:

$$\mathcal{E}_2(\hat{F}) = \mathbb{E}\left[\|\hat{\underline{F}} - \underline{F}\|_2^2\right].$$

__Prop 1:__ We have that $\mathcal{E}_2(\underline{F}^{range}) = \frac{2(K-1)^3}{N^2\epsilon^2}$.

__Proof:__ Via the fact that the $(Z_j)_{j\geq 1}$ RVs are iid, we get

note!

$$\mathcal{E}_2(\underline{F}^{range}) = \frac{1}{N^2} \cdot \sum_{j=1}^{K-1} \mathcal{E}_2(M_{\underline{c},j})$$

$$= \frac{1}{N^2} \cdot (K-1) \cdot \mathrm{Var}(Z_j) = \frac{2(K-1)^3}{N^2\epsilon^2}. \quad \blacksquare$$

② __DP CDFs via histogram queries__

The previous mechanism obtained an ℓ_2-error that was $\Theta\left(\frac{(K-1)^3}{N^2\epsilon^2}\right)$.

We shall now discuss another technique to obtain an ϵ-DP CDF, using this time the counts $\underline{c}$.

Let $M_{\underline{c}}(\underline{x}) = (M_{\underline{c},1}, \ldots, M_{\underline{c},k})$ be an ϵ-DP estimate of $\underline{c}$, obtained by setting

$$M_{\underline{c},i} = c_i + Z_i',$$

where $Z_i' \sim \mathrm{Lap}\left(\frac{2}{\epsilon}\right)$.

We then release $\underline{F}^{hist} \triangleq \frac{1}{N} \cdot \overline{M}_{\underline{c}} = \frac{1}{N} \cdot (\overline{M}_{\underline{c},1}, \cdots, \overline{M}_{\underline{c},k})$, where

$$\overline{M}_{\underline{c},i} = \sum_{j \leq i} M_{\underline{c},j}$$

as an ϵ-DP estimate of $\underline{F}$ [via post-processing].

<u>Prop. 2</u>: We have that $E_2(F^{hist}) = \frac{4k(k-1)}{N^2 \epsilon^2}$. $\left[= \Theta\left(\frac{k^2}{N^2\epsilon^2}\right)\right]$

<u>Proof</u>: Observe that

$$E_2(F^{hist}) = \frac{1}{N^2} \cdot \sum_{i=1}^{k-1} \mathbb{E}\left[\left(\sum_{j \leq i} Z_j'\right)^2\right]$$

$$= \frac{1}{N^2} \cdot \sum_{i=1}^{k-1} \sum_{j \leq i} \mathbb{E}\left[(Z_j')^2\right] \qquad \text{[Why?]}$$

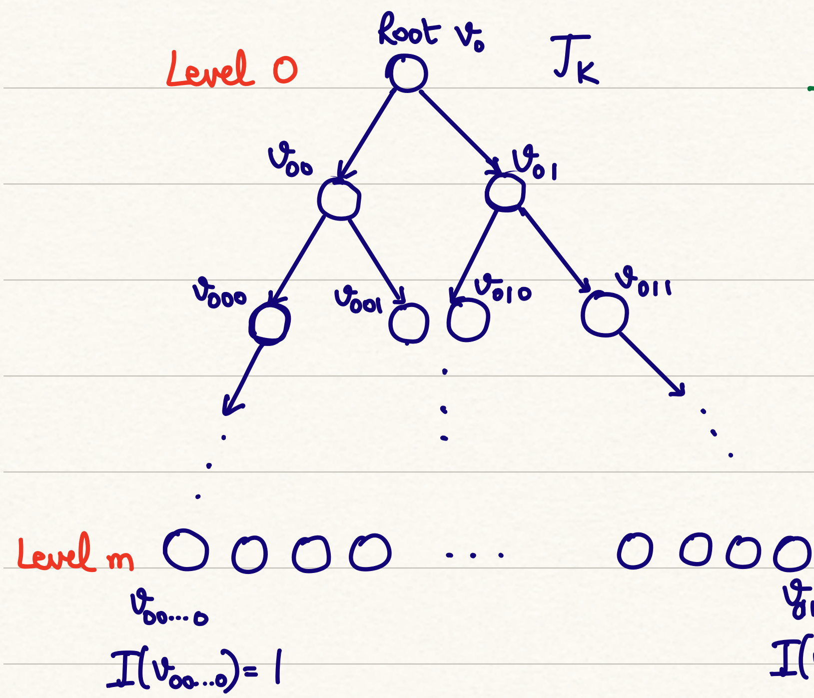
$$= \frac{8}{N^2\epsilon^2} \cdot \sum_{i=1}^{k-1} i = \frac{4k(k-1)}{N^2\epsilon^2}. \qquad ▨$$

In the next section, we shall discuss a very clever mechanism, called the binary tree mechanism (BT), which results in an ℓ_2-error that is $\Theta\left(\frac{K(\log K)^3}{N^2\epsilon^2}\right)$.

While the BT mechanism as stated is applied to CDF release, one can extend it very straightforwardly for the release of histograms, interval queries, etc.; futhermore, this versatile mechanism also sees applications in local DP settings, in DP-"Follow-the-Regularized-Leader" (DP-FTRL), and many more.

③ DP CDFs via the BT mechanism

Let $K = 2^m$, for some $m \geq 1$. We induce a BT $\mathcal{T}_K$ on the bins $[K]$ that are treated as the leaves of $\mathcal{T}_K$.



To any non-root node $u \in \mathcal{T}_K$, add $\mathrm{Lap}(1/\epsilon)$ noise to the count in its assoc. interval:

$$Y(u) = X(u) + \underbrace{Z_u}_{\sim \mathrm{Lap}\left(\frac{2m}{\epsilon}\right)}.$$

Construct

$$\left(M_\epsilon^{BT}(x)\right)_i = \sum_{\underbrace{u \in C(i)}_{\text{covering of } i \in \{0,1,\ldots,K-1\}}} Y(u).$$

Glossary :

① $v_{b_1 b_2 \dots b_\ell}$: An arbitrary node of $\mathcal{T}_K$ at level $0 \le \ell \le m$.

Note that level ℓ has 2^ℓ nodes.

② The $K = 2^m$ leaves 'u' have associated with them intervals $I(u)$ that correspond to the individual bins $j \in [K]$.

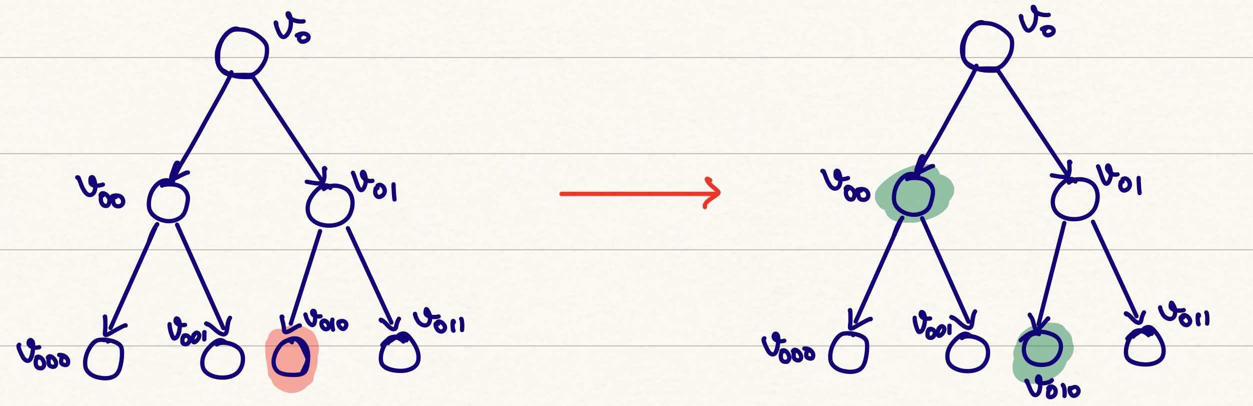
For any non-leaf node v, $I(v) = \bigcup_{u = \text{child}(v)} I(u)$. Clearly, $I(r) = [K]$.

③ Associated to the leaf nodes u, we have counts

$$X(u) = \#\{j : x_j \in B_u\},$$

where B_u is the interval $I(u)$.

④ For every leaf node u, we associate the "cumulative" interval $\vec{I}(u) = \bigcup_{\substack{w \le u, \\ w \text{ leaf node}}} I(w)$. The covering of $\vec{I}(u)$, also written as $C(u)$, is the smallest collection of nodes in $\mathcal{T}_K$ the union of whose intervals equal $\vec{I}(u)$. For example, in the tree below:



the covering $C(v_{010})$ on the left is the collection of green nodes on the right.

We finally release $\underline{F}^{BT} \triangleq \frac{1}{N} \cdot M_\epsilon^{BT}(\underline{x}) = \frac{1}{N} \cdot \left(M_{\epsilon,1}^{BT}(\underline{x}), \dots, M_{\epsilon,k}^{BT}(\underline{x})\right)$ as an estimate of $\underline{F}$.

Claim: $\underline{F}^{BT}$ is ϵ-DP.

Proof: It suffices to show that the estimates $Y(u)$ satisfy ϵ-DP. This holds since changing a single sample can change at most $2m$ node counts $X(u)$, $u \in J_k$. $\boxtimes$

Prop. 3: We have $E_2(\underline{F}^{BT}) = \frac{4km^3}{N^2\epsilon^2} = \frac{4k(\log k)^3}{N^2\epsilon^2}$.

Proof: Note that

$$E_2(\underline{F}^{BT}) = \frac{1}{N^2} \cdot \sum_{u \text{ leaf node}} E_2\Big(\sum_{v \in C(u)} Y_v\Big)$$

$$= \frac{1}{N^2} \cdot \sum_{u \text{ leaf node}} \sum_{v \in C(u)} \mathrm{Var}(Z_v) \quad \text{—(1)}$$

We assume that N is publicly known; hence, $\mathrm{Var}(\bar{Z}_{\mathfrak{r}}) = 0$, since we set $\bar{Z}_{\mathfrak{r}} = 0$. Let s_v be the # of times any non-root node v appears in the summation in (1). Further, let $\eta(v)$ denote the number of leaves in the subtree with v as the root.

By the structure of $C(u)$, we have that if v is in level $0 \le i \le m$,

$$s_v = \begin{cases} 2^{m-i}, & \text{if } v = v_{0b_1b_2\ldots b_{i-1}0}, \\ 0, & \text{if } v = v_{0b_1b_2\ldots b_{i-1}1}. \end{cases}$$

Plugging this back in (1), we see that

$$N^2 \cdot E_2(\underline{F}^{BT}) = \sum_{v \in J_K} s_v \cdot \mathrm{Var}(\bar{Z}_v)$$

$$= \sum_{i=1}^{m} \frac{8m^2}{\epsilon^2} \cdot 2^{i-1} \cdot \left[0 + 2^{m-i}\right]$$

non-root nodes

$$= \sum_{i=1}^{m} \frac{8m^2}{\epsilon^2} \cdot 2^{m-1}$$

$$= \frac{4K(\log K)^3}{\epsilon^2}. \quad \left[\text{using } K = 2^m\right] \qquad \blacksquare$$