CS839: Differential Privacy and Learning Feburary 04, 2026

Lecture 5: Linear Queries & the Binary Tree Mechanism

Instructor: Gavin Brown *Scribe:* Oliver Jing

Disclaimer: This document is intended as an informal supplement to in-class note-taking. It has not been given the level of scrutiny expected in polished lecture notes, let alone that reserved for peer-reviewed publications.

1 Linear Queries / Counting

Setup: data universe $\mathcal{U}$. Given $x_1, .., x_n \in \mathcal{U}$, k queries, and $f_j(x) = \frac{1}{n}\sum_{i=1}^n \varphi_j(x_i)$ where $\varphi_j : \mathcal{U} \to \{0,1\}$ is a predicate. Want to approximate $(f_1(x), .., f_k(x))$, i.e. approximate the $F : \mathcal{U}^n \to \{0,1\}^n$. Suppose our private algorithm is $M_{\varepsilon,\delta}$, define the error to be

$$\text{err}(F, x, M_{\varepsilon,\delta}) := \frac{1}{k}\mathbb{E}\left[\|F(x) - M_{\varepsilon,\delta}(x)\|_2^2\right] = \frac{1}{k}\sum_{i=1}^n \mathbb{E}[(f_j(x) - (M_{\varepsilon,\delta})_j)^2]$$

Recall that the high dimensional global sensitivity of $f : \mathcal{U}^n \to \mathbb{R}^n$ is defined as

$$\Delta_p := \max_{x \sim x'} \|f(x) - f(x')\|_p$$

Now, let's consider a few specific scenarios.

Questions 1. Consider pure DP and Laplace noise, i.e. $M_{\varepsilon,\delta}(x) = F(x) + \text{Lap}(\Delta_1/\varepsilon)$. What is Δ_1? What is err?

Answer 1. Fix $x \sim x'$ and assume $x_l \neq x'_l$. Then for any j, we have

$$|f_j(x) - f_j(x')| = \frac{1}{n}|\sum_{i=1}^n (\varphi_j(x_i)) - \varphi_j(x'_i))| = \frac{1}{n}|\varphi_j(x_l) - \varphi_j(x'_l)| \leq \frac{1}{n}$$

Then by def, we can get

$$\Delta_1 = \max_{x,x'} \sum_{j=1}^k |f_j(x) - f_j(x')| \leq \frac{k}{n}$$

And with the choice of $\Delta_1 = \frac{k}{n}$, the error is also easy to compute:

$$\text{err} = \frac{1}{k}\sum_{j=1}^k \mathbb{E}[\text{Lap}(\frac{k}{\varepsilon n})^2] = \frac{1}{k}\sum_{j=1}^k \text{Var}(\text{Lap}(\frac{k}{\varepsilon n})) = 2\frac{k^2}{\varepsilon^2 n^2}$$

where the 2nd identity uses that our Laplacian distribution has mean zero.

Questions 2. Consider (ε, δ)-DP and Gaussian noise, i.e. $M_{\varepsilon,\delta}(x) = F(x) + N(0, \Delta_2^2\sigma_{\varepsilon,\delta}^2\mathbb{I}_k)$. What is Δ_2? What is err?

Answer 2. Similarly, we have

$$\Delta_2 = \max_{x\sim x'} \|F(x) - F(x')\|_2 = \max_{x\sim x'} (\sum_{k=1}^{j} (f_j(x) - f_j(x'))^2)^{1/2} \leq (\sum_{i=1}^{k} \frac{1}{n^2})^{1/2} = \frac{\sqrt{k}}{n}$$

Then we can choose $\delta_{\varepsilon,\delta}^2 = \frac{2\log(2/\delta)}{\varepsilon^2}$ and we have

$$\text{err} = \text{Var}(N(0, \Delta_2^2\sigma_{\varepsilon,\delta}^2\mathbb{I}_k)) = \Delta_2^2\sigma_{\varepsilon,\delta}^2 = \frac{k}{n^2} \cdot \frac{2\log(2/\delta)}{\varepsilon^2}$$

Questions 3. Write F for histogram with k bins. What is Δ_1 here?
Answer 3. Suppose we have bins $S_1, .., S_k$. Each predicate is then just

$$\varphi_j(x) = \begin{cases} 1 & \text{if } x \in S_j \\ 0 & \text{otherwise} \end{cases}$$

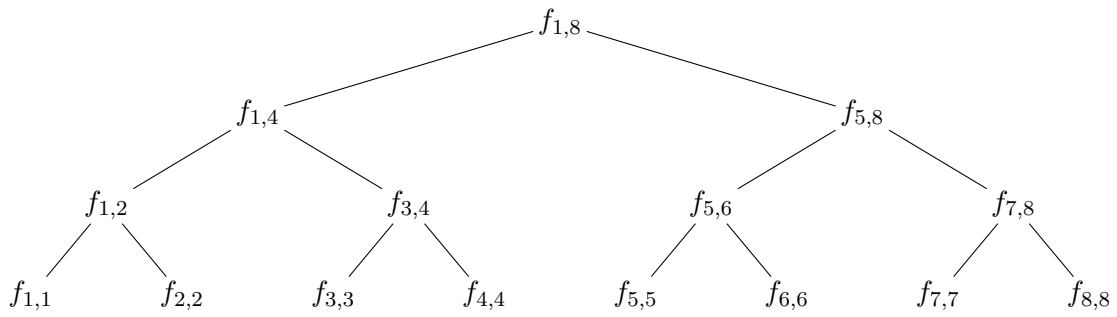
Note that for one changed data, only two bins are affected, so we have

$$\Delta_1(F_{\text{hist}}) = \max_{x\sim x'} \|F_{\text{hist}}(x) - F_{\text{hist}}(x')\|_1 = \frac{2}{n}$$

Questions 4. Interval Queries: let $\mathcal{U} = [d]$ and $\binom{d}{2}$ queries, $f_{s,t}(x) = \begin{cases} 1 & \text{if } s \leq x \leq t \\ 0 & \text{otherwise} \end{cases}$ where $1 \leq s \leq t \leq d$. What is Δ_1 How to add less noise?

Answer 4. A natural idea is to first answer $f_{1,1}, .., f_{d,d}$, which are just histograms, to get $\tilde{f}_{1,1}, .., \tilde{f}_{d,d}$; then we can write $\tilde{f}_{s,t} = \sum_{i=s}^{t} \tilde{f}_{i,i}$, which then implies average err $\approx O_{n,\varepsilon}(\text{poly}(d))$.

A better way to do it is to use the so-called "Binary Tree Mechanism". Below we have an example picture when $d = 8$.



Note that $\Delta_1(F_{\text{tree}}) \lesssim \frac{\log_2 d}{n}$, which then implies average err $\lesssim \frac{\text{poly}(\log_2 d)}{\varepsilon^2 n^2}$, i.e. an exponential improvement.

This leads to a general question: for a given "workload" of queries $F = (f_1, .., f_k)$, what questions should we answer? Stay tuned for an answer!