Today:

- Answering Linear Queries
- Laplace & Gaussian Noise
- Special Cases
  - Histograms
  - Interval Queries
- Factorization

Toolkit so far:

→ bound global sensitivity, add noise

→ optimization, use DP-GD

Important class of problems with well-developed theory, useful all over in practice

Linear Queries ⟺ Answering Many Counts

Data $x = (x_1, \ldots, x_n) \in \mathcal{U}^n$

Queries $f_j(x) = \frac{1}{n} \sum_{i=1}^{n} \phi_j(x_i)$

$f_1, \ldots, f_k$ $\qquad \phi_j : \mathcal{X} \to \{0, 1\}$

"false" "true"

$F(x) : \mathcal{U}^n \to [0,1]^k$

Avg squared ℓ_2 error, alg A, answers $(a_1, \ldots, a_k)$

$$\frac{1}{k} \sum_{j=1}^{k} (f_j(x) - a_j)^2 = \frac{1}{k} \|F(x) - A(x)\|_2^2$$

Ask:

① For ε-DP, add $\mathrm{Lap}(\Delta_1/\varepsilon)$ to each coordinate. What is Δ_1? What is avg error?

(2) (ε, δ)-DP, $\mathcal{N}(0, \Delta_2^2 \cdot \sigma_{\varepsilon,\delta}^2 \cdot \mathbb{I}_d)$.
What is Δ_2? Avg error?

(3) Apply this problem to histograms with k bins. What is Δ_1?

(4) Let $x_i \in \mathcal{X} = [d] = \{1, 2, \ldots, d\}$.
Set $\binom{d}{2}$ "threshold queries" for $1 \le s \le t \le d$, $q_{s,t}(x_i) = \begin{cases} 1 & \text{if } s \le x_i \le t \\ 0 & \text{o.w.} \end{cases}$

What is Δ_1?
How can we do better?

(1) $\max_{x \sim x'} \|F(x) - F(x')\|_1 = \frac{k}{n}$

$$\mathbb{E} \frac{1}{k} \|F(x) - A(x)\|_2^2 = \frac{1}{k} \mathbb{E} \|Lap(\tfrac{k}{\varepsilon n})\|^2$$
$$= \frac{1}{k} \mathbb{E} \sum_{j=1}^{k} (Lap(\tfrac{k}{\varepsilon n}))^2$$
$$= Var\left(Lap(\tfrac{k}{\varepsilon n})\right) = \frac{2k^2}{\varepsilon^2 n^2}.$$

(2) $\max_{x \sim x_i} \|F(x) - F(x')\|_2 \simeq \frac{\sqrt{k}}{n}$

$cov \Rightarrow \sigma^2_{\varepsilon\delta} \frac{k}{n^2} \approx \frac{\log 1/\delta}{\varepsilon^2} \frac{k}{n^2}$ ☺

(3) ok

(4) $\Delta \leq d^2$

Simple improvement:

answer subset queries, use answers to those to build up.

$f_{11}, f_{12}, f_{13}, \ldots, f_{1d}$

$\Delta_1 \leq d$

tree, take $d = 8$

f_{18}

f_{14} f_{58}

f_{12} f_{34} f_{56} f_{67}

f_{11} f_{22} f_{33} f_{44} f_{55} f_{66} ...

Claim Δ_1 for above is $\Delta_1 \leq 2 \log_2 d$.

General Question:

Given target "workload" of queries, what's the best subset to answer?