Back to linear queries: factorization and projection

- For $D \in \mathcal{D}$, let h_D be $\overset{\text{normalized}}{\text{histogram}}$ of D:

$$(h_D)_x = D_x = \frac{\#\text{ copies of } x \text{ in } D}{n} \qquad (n = |D|)$$

- $\overset{\text{normalized}}{\text{Linear}}$ query $f = \sum_{x \in D} f(x) \frac{D_x}{n} = \sum_{x \in D} f(x) (h_D)_x$

$$= \langle f, h_D \rangle$$

- Given a collection of linear queries $f_1, \dots f_k$ $(= Q)$,

want to get $f_1(D) = \langle f_1, h_D \rangle$, $f_2(D) = \langle f_2, h_D \rangle, \dots$

$m = |X|$

$$F = \begin{bmatrix} f_1(x_1) & f_1(x_2) & \cdots & f_1(x_m) \\ f_2(x_1) & f_2(x_2) & \cdots & f_2(x_m) \\ & \vdots & & \\ f_k(x_1) & f_k(x_2) & \cdots & f_k(x_m) \end{bmatrix} \begin{bmatrix} (h_D)_{x_1} \\ \vdots \\ (h_D)_{x_m} \end{bmatrix}$$

want to return $F h_D$!

Laplace/Gaussian mechanism: return $F h_D + \underset{\text{Lap/Gaussian noise}}{Z}$

Gaussian mechanism: std deviation $\approx_{\varepsilon,\delta} \Delta_2 F$

$$\Delta_2 F = \max_{D \sim D'} \| F h_D - F h_{D'} \|_2 = \max_{D \sim D'} \| F(h_D - h_{D'}) \|_2$$

Think of D, D' same size n (swap model of neighboring)

$\Rightarrow D \sim D' \Rightarrow \|h_D - h_{D'}\|_1 \le \frac{2}{n}$

$$\le \max_{v \in \mathbb{R}^n : \|v\|_1 \le \frac{2}{n}} \| F v \|_2$$

$$= \frac{2}{n} \max_{v \in \mathbb{R}^n : \|v\|_1 \le 1} \| F v \|_2$$

$$= \frac{2}{n} \cdot (\text{largest } \ell_2\text{-norm of any col of } F)$$

$$= \frac{2}{n} \cdot \| F \|_{1 \to 2}$$

In other words: add Gaussian noise w/ std dev $\quad c_{\varepsilon,\delta} \cdot \frac{\|F\|_{1\to 2}}{n}$

Error: measure ℓ_2-norm instead of ℓ_∞

error of answers a isn't $\|a - F h_D\|_\infty = \max_{i \in (k)} (a_i - f_i(D))$,

but $\frac{1}{k^{1/2}} \|a - F h_D\|_2$

rescaling: if $\|a - F h_D\|_\infty = \alpha \Rightarrow \frac{1}{k^{1/2}} \|a - F h_D\|_2 \le \frac{1}{\sqrt{k}} \cdot (k \alpha^2)^{1/2} = \alpha$

"ℓ_2 average error"

Error of Gaussian mechanism M_G:

$$E\left[\frac{1}{k^{1/2}} \| F h_D - M_G(D) \| \right] = O(c_{\varepsilon,\delta} \Delta_2 F) = O\left(c_{\varepsilon,\delta} \frac{\|F\|_{1\to 2}}{n}\right)$$

$= O\left(c_{\varepsilon,\delta} \cdot \frac{\sqrt{k}}{n}\right)$ (each entry of $F \in [-1,1]$)

Can we improve this?

<u>Factorization</u>: Factor F into "measurement" and "reconstruction".

Motivating example: Say just repeat the same query f many times

$$F = \begin{bmatrix} \text{—— } f \text{ ——} \\ \text{—— } f \text{ ——} \\ \text{—— } f \text{ ——} \end{bmatrix} \quad \text{e.g.} \quad \begin{bmatrix} 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ & & \vdots & & \end{bmatrix}$$

Gaussian mechanism: $\Delta_2 F = \Theta\left(\frac{\sqrt{k}}{n}\right)$, but clearly only need to answer once with noise $\approx \Theta\left(\frac{1}{n}\right)$, then repeat!

One measurement: $M = [\text{—— } f \text{ ——}] = [1\ 1\ 0\ 0\ 1]$

$$R = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \qquad F = RM$$

Mechanism : $M_{R,M}(D) = \underset{\substack{\uparrow \\ \text{post-processing!}}}{R}(Mh_D + Z) = RMh_D + RZ$

$= Fh_D + RZ$

Z just needs to be $\overset{\text{Gaussian}}{\text{noise}}$ making queries M private, not queries F! std dev. $\sigma = c_{\varepsilon,\delta} D_2 M = O\left(c_{\varepsilon,\delta} \frac{\|M\|_{1\to 2}}{n}\right)$

Gaussian mechanism: $M = F$, $R = I$

(Note: binary tree mechanism special case!)

Error :

$$E\left[\frac{1}{k^{1/2}} \|Fh_D - M_{R,M}(D)\|_2\right] = \frac{1}{k^{1/2}} E\left[\|RZ\|_2\right]$$

$$E[\|RZ\|_2] \le \left(E[\|RZ\|_2^2]\right)^{1/2}$$

$$= \left(E\left[\sum_{i=1}^{k} (RZ)_i^2\right]\right)^{1/2}$$

$$= \left(\sum_{i=1}^{k} E[(RZ)_i^2]\right)^{1/2}$$

$$= \left(\sum_{i=1}^{k} E[\langle r_i, \underset{\substack{\uparrow \\ \text{Gaussian w/ std dev. } \sigma}}{Z}\rangle^2]\right)^{1/2}$$

$$= \left(\sum_{i=1}^{k} \|r_i\|_2^2 \sigma^2\right)^{1/2}$$

$$= \sigma\left(\sum_{i=1}^{k} \|r_i\|_2^2\right)^{1/2} = \sigma\left(\sum_{i=1}^{k} \sum_{j=1}^{\overset{\#\text{ queries in } M}{\ell}} R_{ij}^2\right)^{1/2}$$

$= \sigma \cdot \|R\|_F$ ← Frobenius Norm

$$\Rightarrow E(\text{error}) = \frac{1}{k^{1/2}} \sigma \|R\|_F = \frac{1}{k^{1/2}} c_{\varepsilon,\delta} \frac{\|M\|_{1\to 2}}{n} \cdot \|R\|_F$$

$$= \frac{c_{\varepsilon,\delta} \|M\|_{1\to 2} \|R\|_F}{k^{1/2} n}$$

Gaussian Mech: $M = F$, $R = I$ $\Rightarrow$ $\|M\|_{1\to 2} = \|F_{1\to 2}\|$

$\|R\|_F = k^{1/2}$

$$\Rightarrow c_{\varepsilon,\delta} \frac{\|F_{1\to 2}\|}{n} \quad \checkmark$$

Repeat same query many times: $\|M\|_{1\to 2} = 1$

$\|R\|_F = k^{1/2}$

$$\Rightarrow c_{\varepsilon,\delta} \cdot \frac{1}{n}$$

Given F, many choices of M, R!

Def: Given $F \in \mathbb{R}^{k \times m}$, factorization norm is

$$\gamma(F) = \min_{\substack{R,M: \\ F = RM}} \left(\frac{\|R\|_F \|M\|_{1\to 2}}{k^{1/2}} \right)$$

Thm: Given k linear queries represented by $F \in \mathbb{R}^{k \times m}$, $\exists$ (ε,δ)-DP mechanism with expected ℓ_2-error

$$O\left(\frac{c_{\varepsilon,\delta} \cdot \gamma(F)}{n}\right)$$

Note: Can also have approximate factorizations, reason about how error propagates

Projection:

Want answers that "make sense".

$\mathcal{C}$ = "answers that make sense":

$$= \{a \in \mathbb{R}^k : \exists h \in \mathbb{R}^m_{\geq 0} \text{ s.t. } \|h\|=1 \text{ and } a = Fh\}$$

$$\approx \{a \in \mathbb{R}^k : \exists D \in \mathcal{D} \text{ s.t. } a = Fh_D\}$$

$$\Pi_{\mathcal{C}}(a) = \underset{a' \in \mathcal{C}}{\operatorname{argmin}} \|a - a'\|_2$$

Facts:

- $\Pi_{\mathcal{C}}(a)$ just post-processing! If a computed by (ε,δ)-DP mechanism, still (ε,δ)-DP.

- Doesn't increase error since C closed, convex: a^* is true answers.

$$\| \Pi_C(a) - a^* \|_2 \leq \| a - a^* \|_2 \quad \forall a$$

Projection Mechanism:

- Use Gaussian (or factorization) mechanism on F to get a
- Output $\Pi_C(a)$.

Since Π_C is post-processing, expected l_2-error $\leq c_{\varepsilon,\delta} \frac{k^{1/2}}{n}$ (k = # queries in F)

Analyze carefully:

Thm: $E[l_2$ error of projection mechanism$] \leq$

$$O\left(\left(\frac{c_{\varepsilon,\delta} \log^{1/2} m}{n} \right)^{1/2} \right) \qquad m = |X|$$

Interpretation: $\frac{1}{\sqrt{n}}$ instead of $\frac{1}{n}$: worse.

if $n < k^{1/2}$, gaussian has error > 1, even though values in $[0,1]$: meaningless!

but if $n > \log^{1/2} m$, projection still meaningful!

so interesting if $\log^{1/2} m < n$

- database has to be somewhat large in universe size, but not too bad.