(Linear) Query Release & Synthetic Data

- Recap query release problem
- Why "linear"?
- Factorization Framework.
  - ↳ "Matrix Mechanism"

Linear Query Release

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

$\mathcal{X}$ "data universe"

Statistics $f_1, \dots, f_k$

$$f_i(x) = \frac{1}{n} \sum_{j=1}^{n} \varphi_i(x_j) \quad , \quad \varphi_i : \mathcal{X} \mapsto \{0,1\}$$

"predicate"

Asian $\wedge$ "age ≥ 30" $\to 1$

(Extends to $\varphi_i : \mathcal{X} \mapsto \mathbb{R}$)

Goal: output $\vec{a} = (a_1, \dots, a_k)$

$$\underbrace{\left(\frac{1}{k} \sum_{i=1}^{k} \left(f_i(x) - a_i \right)^2 \right)^{\frac{1}{2}}}_{\text{"}\ell_2\text{ error"}} \leq \alpha$$

$$\vec{F}(x) = \left(f_1(x), \dots, f_k(x) \right)$$

$$\frac{1}{\sqrt{k}} \left\| \vec{F}(x) - a \right\|_2 \leq \alpha.$$

Mechanisms: Laplace

Gaussian

Binary Tree Mechanism

Gaussian Mechanism

$$M(x) = \vec{F}(x) + Z \,, \qquad Z \sim N(0, \sigma^2 I_{k \times k}).$$

$$\sigma^2 = C_{\varepsilon,\delta} \cdot \Delta_2^2$$

$C_{\varepsilon,\delta} \approx \frac{\log(1/\delta)}{\varepsilon^2}$; Δ_2: ℓ_2-sensitivity

$$\Delta_2 = \max_{x \sim x'} \|F(x) - F(x')\|_2$$

$$= \max_{u, u' \in \mathcal{X}} \frac{1}{n} \|F(u) - F(u')\|_2$$

$$\leq \max_{u \in \mathcal{X}} \frac{2}{n} \|F(u)\|_2$$

$$\leq 2 \frac{\sqrt{k}}{n}$$

ℓ_2 error

$$\mathbb{E}\left[\frac{\|Z\|_2}{\sqrt{k}}\right] \leq C_{\varepsilon,\delta} \cdot \sigma$$

$$= \frac{2 \cdot C_{\varepsilon,\delta} \, k^{\frac{1}{2}}}{n}$$

Why "Linear"?

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

data universe $\mathcal{X} = \{1, \dots \dots, m\}$, $|\mathcal{X}| = m$

Histogram $h_x \in \mathbb{R}^m$

$$\forall u \in \mathcal{X} : (h_x)_u = \frac{1}{n} \left| \{ j : x_j = u \} \right|$$

univ: $\mathcal{X} = \{1, 2, 3\}$

dataset: $x = (1, 2, 3, 3, 1)$ $\longrightarrow$ $h_x = \left(\frac{2}{5}, \frac{1}{5}, \frac{2}{5}\right)$.

$h_{x'} = \left(\frac{1}{5}, \frac{2}{5}, \frac{2}{5}\right)$

① $\forall$ dataset x, $\|h_x\|_1 = 1$

② $\forall$ neighbors x & x', $\|h_x - h_{x'}\|_1 \leq \frac{2}{n}$

③ Queries $f_1, \dots, f_k$
$\varphi_1, \dots, \varphi_k$

$$F = \begin{array}{cc} & m \\ k & \begin{pmatrix} \varphi_1(u_1) & \cdots & \varphi_1(u_m) \\ \vdots & \ddots & \vdots \\ \varphi_k(u_1) & \cdots & \varphi_k(u_m) \end{pmatrix} \end{array}$$

answer vector $\vec{F}(x) = F h_x$

"Only for thought experiment" $m \approx 2^d$

Revisiting Gaussian Mechanism.

$$M(x) = F h_x + Z \quad , \quad Z \sim N(0, \sigma^2 I_{k\times k})$$

$$\sigma^2 = C_{\varepsilon,\delta}^2 \cdot \Delta_2^2$$

$$\Delta_2 = \max_{x\sim x'} \| F \underbrace{(h_x - h_{x'})}_{v} \|_2$$

$$\le \max_{\substack{v\in\mathbb{R}^m \\ \|v\|_1 = \frac{2}{n}}} \|Fv\|_2$$

$$= \frac{2}{n} \max_{\substack{v\in\mathbb{R}^m \\ \|v\|_1 = 1}} \|Fv\|_2 \longrightarrow$$ "largest ℓ_2-norm of a column in F"

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

$$\mathbb{E}\left[\frac{\|Z\|_2}{\sqrt{k}}\right] = 2\,\frac{C_{\varepsilon,\delta} \cdot \boxed{\|F\|_{1\to 2}}}{n}$$

Characterize Sensitivity.

worst-case $\sqrt{k}$

Can we do better when F has some structure?

General Factorization Framework

Histogram $h_x \in \mathbb{R}^m$ $\Big\{$ dataset of size n

linear queries $F \in \mathbb{R}^{k \times m}$

Want to release $F h_x$

① Approximate $\tilde{F} \approx F$

Focus of today

② Factorize $\tilde{F} = \underbrace{R}_{\text{"Reconstruction"}} \quad \underbrace{M}_{\text{"measurement"}}$

$$\hat{a} = R\,(M h_x + \underbrace{Z}_{\text{noise}})$$

$$= R M h_x + R Z$$

$$= \tilde{F} h_x + R Z$$

③ Post-processing to $\tilde{a}$ to satisfy some "consistency" properities

Factorization

Given linear queries $F \in \mathbb{R}^{k \times m}$

(Trivial Example) $f_1 = \cdots = f_k$

Gaussian Mechanism: error $\approx \frac{k^{\frac{1}{2}}}{n} \cdot C_{\varepsilon,\delta}$.

$a_1 = f_1(x) + z$

$\vec{a} = (a_1, \dots\dots, a_1)$ error $\approx \frac{1}{n} \, C_{\varepsilon,\delta}$.

$$F = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ & & \cdots & & \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{pmatrix}}_{R \in \mathbb{R}^{k \times \ell}} \underbrace{\begin{pmatrix} 1 & 0 & 1 & 0 & 1 \end{pmatrix}}_{M \in \mathbb{R}^{\ell \times m}}$$

More generally:

Gaussian Mech v.s. $R(Mh_x + \underbrace{z}_{\propto \|M\|_{1\to 2}})$

$F h_x + \boxed{z}$ i.i.d noise. $= Fh_x + \underbrace{\boxed{Rz}}_{\text{Correlated noise.}}$

proportional to $\|F\|_{1\to 2}$ (could be $\sqrt{k}$)

Factorization

For R, M s.t. $F = RM$

$$\mathcal{M}_{R,M}(x) = R\left(M h_x + Z\right)$$

error

$$= F h_x + \underbrace{\boxed{RZ}}_{\text{Correlated noise.}}$$

$$Z \sim N(0, \sigma^2 I_{\ell \times \ell})$$

$$\sigma^2 = C_{\varepsilon,\delta}^2 \|M\|_{1\to 2}^2$$

Factorization

For fixed $RM = F$, analyze error $\mathbb{E}\left[\frac{\|Rz\|_2}{\sqrt{k}}\right]$

$$\underbrace{\begin{pmatrix} r_1 \cdot z \\ \vdots \\ r_k \cdot z \end{pmatrix}}_{Rz} = \underbrace{\begin{pmatrix} \text{—} r_1 \text{—} \\ \vdots \\ \text{—} r_k \text{—} \end{pmatrix}}_{R} \underbrace{\begin{pmatrix} z_1 \\ \vdots \\ z_\ell \end{pmatrix}}_{z}, \quad z \sim N(0, \sigma^2 I_{\ell \times \ell})$$

$$\boxed{\begin{aligned} &\text{Fact.} \quad r_i \cdot z \sim N(0, \sigma^2 \|r_i\|_2^2) \\ &\mathbb{E}[(r_i \cdot z)^2] = \sigma^2 \|r_i\|_2^2 . \end{aligned}}$$

$$\begin{aligned}
\mathbb{E}[\|Rz\|_2] &\underset{\text{Jensen}}{\leq} \left(\mathbb{E}[\|Rz\|_2^2] \right)^{\frac{1}{2}} \\
&= \left(\mathbb{E}\left[\sum_{i=1}^{k} (r_i \cdot z)^2 \right] \right)^{\frac{1}{2}} \\
&= \left(\sum_{i=1}^{k} \mathbb{E}[(r_i \cdot z)^2] \right)^{\frac{1}{2}} \\
&= \left(\sum_{i=1}^{k} \sigma^2 \|r_i\|_2^2 \right)^{\frac{1}{2}} \\
&= \sigma \underbrace{\left(\sum_{i=1}^{k} \|r_i\|_2^2 \right)^{\frac{1}{2}}} \\
&= \sigma \, \|R\|_F
\end{aligned}$$

$$R = \begin{pmatrix} \text{—} r_1 \text{—} \\ \vdots \\ \text{—} r_k \text{—} \end{pmatrix}$$

σ: Scales with $\|M\|_{1\to 2}$.

$\|R\|_F$: Frobenius norm.

Putting together: Expected error

$$\frac{1}{\sqrt{k}} \mathbb{E}[\|Rz\|_2] \leq O\left(\frac{C_{\varepsilon,\delta} \, \|R\|_F \, \|M\|_{1\to 2}}{\sqrt{k} \quad n} \right)$$

Example / Exercise :

Binary Tree Mechanism.

Factorization Framework.

Error $O\left(\frac{C_{\varepsilon,\delta}}{n} \cdot \boxed{\frac{\|R\|_F \cdot \|M\|_{1\to 2}}{\sqrt{k}}}\right)$

Factorization norm of F

$$\gamma(F) = \min\left\{ \frac{\|R\|_F \cdot \|M\|_{1\to 2}}{\sqrt{k}} : RM = F \right\}$$

Theorem. For every $F \in \mathbb{R}^{k\times m}$, there is (ε,δ)-DP mechanism with ℓ_2-error $\leq O\left(\frac{C_{\varepsilon,\delta}}{n} \cdot \gamma(F)\right)$.