(Linear) Query Release

- Recap: Factorization Mechanism
- Projection Mechanism
- Preview: Synthetic Data
- Online Learning

Update: lecture notes / Slides

Linear Query Release

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

$\mathcal{X}$ "data universe"

Statistics $f_1, \ldots, f_k$

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

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

$$\left(\frac{1}{k} \sum_{i=1}^{k} \left(f_i(x) - a_i \right)^2 \right)^{\frac{1}{2}} \leq \alpha$$

"ℓ_2 error"

$$\boxed{\begin{gathered} \vec{F}(x) = (f_1(x), \ldots, f_k(x)) \\ \frac{1}{\sqrt{k}} \| \vec{F}(x) - a \|_2 \leq \alpha. \end{gathered}}$$

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

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

General Factorization Framework

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

$\begin{cases} \text{dataset of size } n \\ \text{linear queries } F \in \mathbb{R}^{k \times m} \end{cases}$

Want to release $F h_x$

① Approximate $\tilde{F} \approx F$

② Factorize $\tilde{F} = R \quad M$

"Reconstruction" "measurement"

$$\hat{a} = R\,(M h_x + z)$$

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

For R, M s.t. $F = \overset{\text{Reconstruction}}{R}\overset{\text{Measurement}}{M}$

$$M_{R,M}(x) = R\Big(\underbrace{Mh_x + Z}\Big)$$

$$= Fh_x + \boxed{RZ} \leftarrow \text{Correlated noise}$$

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

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

Factorization Framework.

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

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

Standard Gaussian $$O\left(\frac{C_{\varepsilon,\delta}}{n} \cdot \sqrt{k}\right)$$

$\to$ Factorization $\approx$ pre-processing. Gaussian

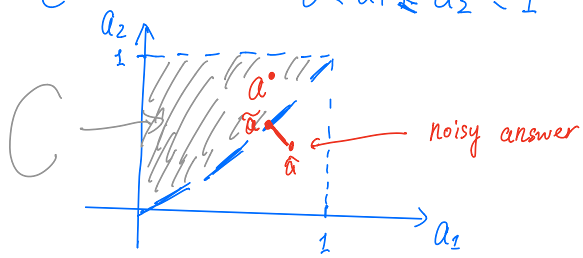
$\to$ post-processing : Projection mechanism

Consistency

$\mathcal{X} = \{1,2,3\}$ $\qquad F = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix}$ $\qquad \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = F \cdot h_x$

Gaussian : $\begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \end{pmatrix} = F h_x + Z$

"Constraint" $\quad 0 \leq a_1 \leq a_2 \leq 1$



Projection:

$$\tilde{a} = \Pi_C(\hat{a}) = \arg\min_{a' \in C} \|a' - \hat{a}\|_2$$

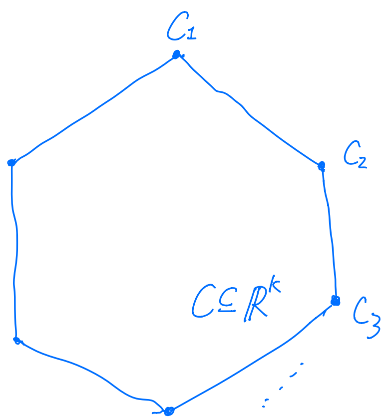
1. $\tilde{a}$ is consistent
2. $\|\tilde{a} - a\|_2 \leq \|\hat{a} - a\|_2$
3. $\tilde{a}$ is (ε, δ)-DP.

Projection may help accuracy!

$$F = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix}$$

$$a_1 + a_2 = a_3$$

Consistency



"Convex hull"

Queries matrix

$$F = \begin{array}{cc} & m \\ k & \begin{pmatrix} | & | & & | \\ c_1 & c_2 & \cdots & c_m \\ | & | & & | \end{pmatrix} \end{array}$$

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

$$C = \{ a \in \mathbb{R}^k : \exists\, h \in \mathbb{R}_+^m,\ \|h\|_1 = 1,\ a = Fh \}$$

The Projection Mechanism

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

Gaussian Mechanism :

$$a = Fh$$

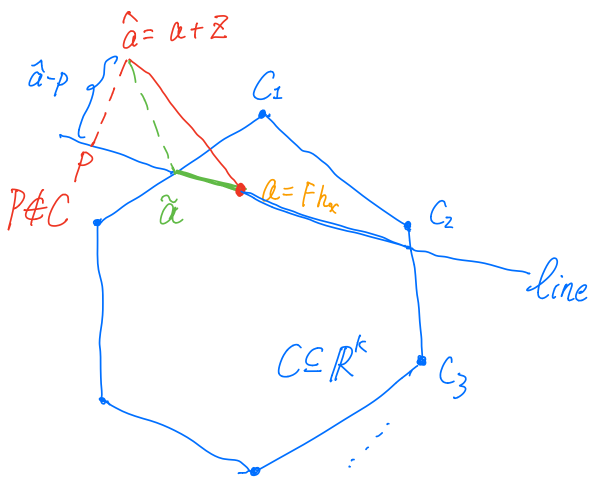
$$\hat{a} = a + Z \quad , \qquad Z \sim N(0, \sigma^2 I_{k\times k})$$

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

Projection

return $\tilde{a} = \arg\min_{a' \in C} \| a - a' \|_2$

Analysis



"ℓ_2 error"

$$\|\tilde{a}-a\|_2^2$$

$$= \langle \hat{a}-a,\ \tilde{a}-a \rangle$$

$$\leq \langle \hat{a}-a,\ p-a \rangle$$

$$= \langle \hat{a}-a,\ p-a + \underbrace{\hat{a}-p}_{\perp\ \hat{a}-a} \rangle$$

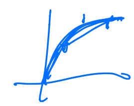
$$= \langle \hat{a}-a,\ \hat{a}-a \rangle$$

$$= \langle \hat{a}-a,\ Z \rangle$$

$$= \langle \hat{a}, z \rangle - \langle a, Z \rangle$$

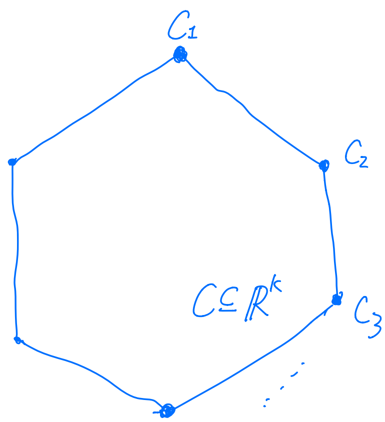
$$\leq |\langle \hat{a}, z \rangle| + |\langle a, z \rangle|$$

$$\leq \underbrace{2 \max_{v \in C} |\langle v, Z \rangle|}_{\text{Data independent.}}$$



Analysis. $$\mathbb{E}\left[\frac{\|\tilde{a}-a\|_2}{\sqrt{k}}\right] \underset{\text{Jensen Ineq}}{\le} \mathbb{E}\left[\frac{\|\tilde{a}-a\|_2^2}{k}\right]^{\frac{1}{2}}$$

(Plug in) $$= \left(\frac{2\,\mathbb{E}_z\left[\max_{v\in C}|\langle v,z\rangle|\right]}{k}\right)^{\frac{1}{2}}$$



Fact. (Suffices to think about vertices)

$$\max_{v\in C}|\langle v,z\rangle| = \max_{j\in[m]}|\langle c_j,z\rangle|$$

Columns of F

$$\langle c_j,z\rangle \sim N(0,\|c_j\|_2^2\,\sigma^2).$$

Fact. If $W_1,\dots,W_m$ are Gaussian with variance $\le (\sigma')^2$, & zero-mean then $\mathbb{E}[\max\{|W_1|,\dots,|W_m|\}] \le \sigma'\cdot\sqrt{\log m}$

$$\Longrightarrow \mathbb{E}_z\left[\max_{v\in C}|\langle v,z\rangle|\right] = \mathbb{E}_z\left[\max_{c_j,\,j\in[m]}|\langle c_j,z\rangle|\right]$$

$$= \sqrt{k}\cdot\sigma\cdot\sqrt{\log m}.$$

$$\mathbb{E}\left[\frac{\|\tilde{a}-a\|_2}{\sqrt{k}}\right] \underset{\substack{\text{Jensen}\\\text{Ineq}}}{\leq} \mathbb{E}\left[\frac{\|\hat{a}-a\|_2^2}{k}\right]^{\frac{1}{2}}$$

$$= \left(\frac{2\ \mathbb{E}_z\left[\max_{v\in C}|\langle v, z\rangle|\right]}{k}\right)^{\frac{1}{2}}$$

$$\leq O\left(\left(\frac{C_{\varepsilon,\delta}\cdot\sqrt{\log m}}{n}\right)^{\frac{1}{2}}\right).$$

Projection Mechanism Bound.

Just Gaussian Mechanism

$$\min\left\{ O\left(\left(\frac{C_{\varepsilon,\delta}\cdot\sqrt{\log m}}{n}\right)^{\frac{1}{2}}\right),\ O\left(\frac{C_{\varepsilon,\delta}\cdot\sqrt{k}}{n\varepsilon}\right)\right\}$$

Dependence on n

$\frac{1}{\sqrt{n}}$ v.s. $\frac{1}{n}$

Dependence on k

No dep. v.s. $\sqrt{k}$

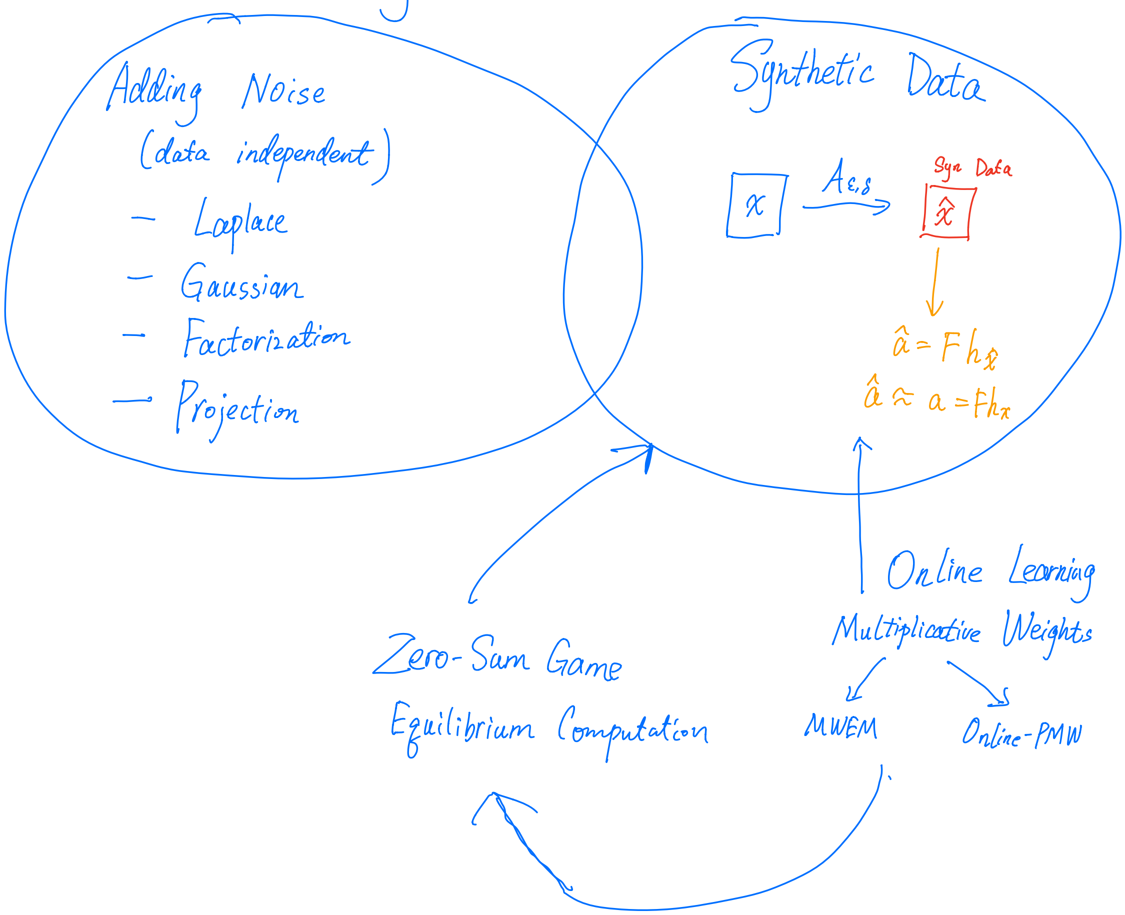
Dependence on m

$\sqrt{\log m}$ v.s. No dep.

Example: $\mathcal{X}=\{0,1\}^d$, $m=2^d$, $k=2^d$

$\left(\frac{d^{\frac{1}{2}}}{n}\right)^{\frac{1}{2}}$ v.s. $\frac{2^{d/2}}{n}$

Query Release



Online Learning

"Sequential decision making"

Setting:
- a set of actions $\{1, \dots, k\} = [k]$.
- "Game" between decision-maker D & Adversary A

"may know D's alg
not D's randomness"

For $t = 1, \dots T$:

D chooses distribution $p^t \in \Delta(A)$

A chooses cost vector $c^t \in [0,1]^k$

$a^t \sim p^t$ sampled action

D pays cost $c^t_{a^t}$ and observes c^t.

Focus on:

$$\mathbb{E}_{a \sim p^t}[c^t_a] = \langle p^t, c^t \rangle$$

Total cost: $\sum_{t=1}^{T} c^t_{a^t}$

How to measure D?

- Compare w/ the sequence $a^1 \dots a^T$
  Hopeless....
- Compare w/ the best action in hindsight

$$\text{Regret} = \frac{1}{T} \sum_{t=1}^{T} c^t_{a^t} - \min_{a \in [k]} \frac{1}{T} \sum_{t=1}^{T} c^t_a$$

Online Learning

Give an algorithm Mutiplicative Weights.

$$\mathbb{E}[\text{Regret}] \leq O\left(\sqrt{\frac{\ln(k)}{T}}\right).$$