

Theory of Differential Privacy

Lecture 11:

Factorization

EN.601.438/638 Fall 2026

Lydia Zakynthinou

Johns Hopkins University

Oct 7, 2026

Dataset $x = (x_1, \dots, x_n) \in \mathcal{X}^n$, $\mathcal{X} = \{u_1, \dots, u_N\}$

Empirical distribution of x : $h_x \in \mathbb{R}^N$, $(h_x)_j = \frac{1}{n} |\{i : x_i = u_j\}|$.

Counting queries

Let $\phi : \mathcal{X} \rightarrow \{0, 1\}$

Write $v_\phi = (\phi(u_1), \dots, \phi(u_N)) \in \{0, 1\}^N$.

$$f_\phi(x) = \frac{1}{n} \sum_{i=1}^n \phi(x_i) = \sum_{j=1}^N \phi(u_j) \cdot (h_x)_j = \langle v_\phi, h_x \rangle.$$

Workload.

For queries $f_1, \dots, f_k$. I can write $F = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_i \\ \vdots \\ f_k \end{bmatrix} = \begin{bmatrix} v_{f_1} \\ \vdots \\ v_{f_k} \end{bmatrix}$

$$F \cdot h_x = \begin{pmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_k(x) \end{pmatrix} \in \mathbb{R}^k \quad F \in \mathbb{R}^{k \times N}$$

$\uparrow$
 $\in \mathbb{R}^N$

Approach 1: Gaussian mechanism.

$$A(x) = \underbrace{F \cdot h_x}_{\in \mathbb{R}^k} + N\left(0, \frac{\Delta 2 \log(2/\delta)}{\epsilon^2} \cdot \mathbb{I}_k\right).$$

- $\max_{x \sim x'} \|h_x - h_{x'}\|_1 \leq \frac{2}{n}$

- $\max_{x \sim x'} \|F h_x - F h_{x'}\|_2 = \max_{x \sim x'} \|F(h_x - h_{x'})\|_2 = \max_{\substack{x \sim x' \\ \|h_x - h_{x'}\|_1 \leq \frac{2}{n}}} \|F(h_x - h_{x'})\|_2$

$$= \max_{\|v\|_1 \leq \frac{2}{n}} \|F \cdot v\|_2 = \frac{2}{n} \max_{\|v\|_1 \leq 1} \|F v\|_2 = \frac{2}{n} \|F\|_{1 \rightarrow 2}$$

$$= \frac{2}{n} \|F\|_{\text{col}}$$

$$\Rightarrow \|A(x) - F h_x\|_2 \lesssim \frac{\sqrt{k} \cdot \|F\|_{1 \rightarrow 2}}{n} \cdot \frac{\sqrt{\log(1/\delta)}}{\epsilon}$$

$$\Rightarrow \overline{en} = \frac{1}{\sqrt{k}} \|A(x) - Fh_x\|_2 \lesssim \frac{\|F\|_{1 \rightarrow 2}}{n} \cdot C_{\varepsilon, \delta} \quad \text{"average error"}$$

Motivating example

- If $f_1 = f_2 = \dots = f \Rightarrow$ average error $\approx \frac{\sqrt{k}}{n}$ if Use Gaussian Mech.

$$D_2 = \frac{2}{n} \max_{\|v\|_1 \leq 1} \left\| \begin{pmatrix} y^T \\ \vdots \\ y^T \end{pmatrix} \cdot v \right\|_2 = \frac{2\sqrt{k}}{n}$$

- $\tilde{f}_i = f_1(x) + N(0, C_{\varepsilon, \delta}^2 \cdot \frac{2}{n}) \rightarrow \tilde{f}_i$ has error $\approx \frac{1}{n}$.

output $(\tilde{f}_1, \tilde{f}_1, \dots, \tilde{f}_1)$. Then average error (and worst-case) $\approx \frac{1}{n}$.

Idea: Choose a set of queries to answer and then reconstruct the answers to queries in F .

Factorization

Given $F \in \mathbb{R}^{k \times N}$, choose matrices

$$M \in \mathbb{R}^{r \times N}$$

$$R \in \mathbb{R}^{k \times r}$$

$\uparrow$
Strategy matrix

$\uparrow$
reconstruction

so that $F = R \cdot M$.

Factorization Mechanism

Privately release $M l_x + Z$, where $Z \sim N(0, \sigma^2 \mathbb{I}_r)$

then output $\tilde{F}_x = R(M l_x + Z)$.

$$\text{Since } F = R \cdot M, \tilde{F}_x = R(M l_x + Z) = R \cdot M l_x + R Z = \underbrace{F \cdot l_x}_{\text{non-private answers}} + \underbrace{R \cdot Z}_{\text{noise}}$$

Privacy

If $\sigma^2 = C_{\epsilon \delta}^2 \cdot \left(\frac{2}{n} \|M\|_{1 \rightarrow 2}^2 \right)^2$, then $M l_x + Z$ is $(\epsilon \delta)$ -DP.

By post-processing, the Factorization Mech is $(\epsilon \delta)$ -DP.

Accuracy

$$\|F h_x - \tilde{F}_x\|_2 = \|R \cdot Z\|_2$$

$$\mathbb{E}[\|R \cdot Z\|_2^2] = \sigma^2 \cdot \|R\|_F^2, \text{ since } Z \sim N(0, \sigma^2 \mathbb{I}_r). \text{ (fact)}$$

$$\text{From the privacy analysis } \sigma^2 \approx \left(\frac{\|M\|_{1 \rightarrow 2}}{n} \right)^2 C_{\epsilon, \delta}.$$

$$\overline{\text{err}} = \frac{1}{\sqrt{K}} \sqrt{\mathbb{E}[\|R Z\|_2^2]} \approx \frac{\|M\|_{1 \rightarrow 2} \cdot \|R\|_F}{\sqrt{K} \cdot n} \cdot C_{\epsilon, \delta}.$$

Factorization norm:

$$\gamma(F) = \min_{F=R \cdot M} (\|M\|_{1 \rightarrow 2} \cdot \|R\|_F) / \sqrt{K}$$

• Worst-case error?

$$\|F h_x - \tilde{F}_x\|_\infty \leq \|F h_x - \tilde{F}_x\|_2 \leq \frac{\gamma(F)}{n}.$$

$$\text{error for } j\text{-th query: } \leq \|M\|_{1 \rightarrow 2} \cdot \underset{\substack{\uparrow \\ \text{row}}}{\|R_j\|_2}$$