

Theory of Differential Privacy

Lecture 8:

Linear Queries, Synthetic data, SmallDB.

EN.601.438/638 Fall 2026

Lydia Zakynthinou

Johns Hopkins University

Sept 28, 2026

Recap

- $x = (x_1, \dots, x_n) \in \mathcal{X}^n$. $M: \mathcal{X}^n \rightarrow \mathcal{Y}$ is (ϵ, δ) -DP if:
For all $x \sim x'$ (one coordinate replaced) and output event S ,
$$\Pr[M(x) \in S] \leq e^\epsilon \Pr[M(x') \in S] + \delta$$

 $\delta = 0$: "pure" DP
 $\delta > 0$: "approximate" DP.
- Sensitivity : $\Delta g = \max_{x \sim x'} |g(x) - g(x')|$
- Laplace Mech.: $g(x) + \text{Lap}\left(\frac{\Delta g}{\epsilon}\right)$
- Composition : $\epsilon \rightarrow \epsilon \cdot k$ // $\epsilon \rightarrow \sqrt{k} \epsilon$, $\delta \rightarrow k \delta$
- post processing : $\epsilon \rightarrow \epsilon$.
- Exp. Mech.: finite range R , score $u(x, r)$ $\Pr[M(x) = r] \propto \exp\left(\frac{\epsilon \cdot u(x, r)}{2 \Delta u}\right)$.

Linear Queries

For public functions $\phi_i: \mathcal{X} \rightarrow [0, 1]$. Define:

$$f(x) = \frac{1}{n} \sum_{i=1}^n \phi_i(x_i)$$

e.g.

$$\phi_i = \phi = \mathbf{1}_{\{z \in A\}}$$

people older than 40.

$$f(x) = \frac{1}{n} \sum_{i: x_i \in A} \mathbf{1} = \frac{\#\{i: x_i \in A\}}{n}$$

for any neighboring $x \sim x'$: $|f(x) - f(x')| \leq \frac{1}{n} |\phi_i(x_i) - \phi_i(x'_i)| \leq \frac{1}{n}$.

Public workload $\mathcal{Q} = \{f_1, f_2, \dots, f_k\}$. Goal is to output $a_f, \forall f \in \mathcal{Q}$:

$$\Pr \left[\max_{f \in \mathcal{Q}} |a_f - f(x)| \leq \alpha \right] \geq 1 - \epsilon$$



"Answers a_{f_j} should be close to $f_j(x)$ $\forall j \in [k]$, with high probability."

Naïve

Output:

$$= \text{Lap}\left(\frac{1}{n\epsilon'}\right)$$

$$a_f = f(x) + \text{Lap}\left(\frac{k}{n\epsilon}\right), \quad \forall f \in \mathcal{Q}.$$

Privacy

By composition: each of the $i \in [k]$ is $\frac{\epsilon}{k}$ -DP.
 $\Rightarrow$ overall algorithm is ϵ -DP.

Accuracy

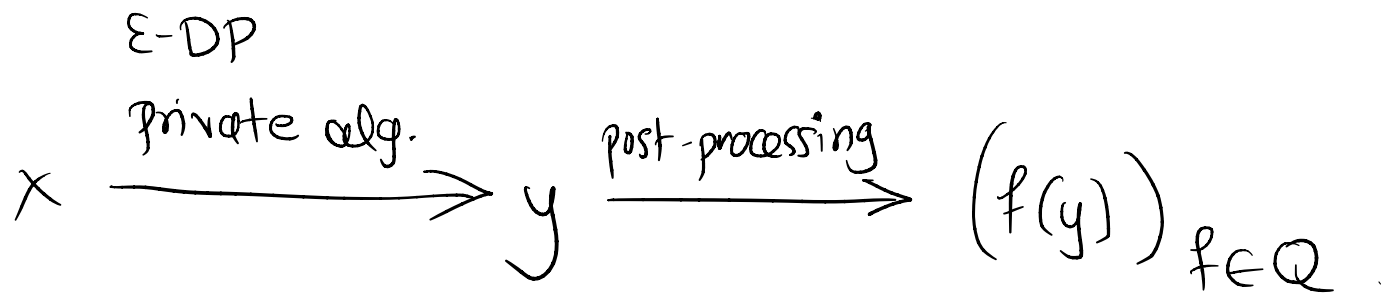
$$\Pr \left[\max_{f_i \in \mathcal{Q}} |f_i(x) - a_{f_i}| = \max_{f_i \in \mathcal{Q}} |Z_{f_i}| > \frac{k}{n\epsilon} \log \frac{k}{\theta} \right] \leq \theta.$$

↑
error grows ^{super}linearly with k .

New idea

Release $y = (y_1, \dots, y_m) \in \mathcal{X}^m$. Then answer:

$$a_{f_i} = f_i(y). \quad \text{err}_Q(x, y) = \max_{f \in Q} |f(x) - f(y)|.$$



Existence of good synthetic database y^* let $\Phi_i = \Phi = f$

Fix $x \in \mathcal{X}^n$, $\mathcal{Q}$, $\alpha \in (0, 1)$. let $m = \left\lceil \frac{\log(4k)}{2\alpha^2} \right\rceil$. Then there exists

$$y^* \in \mathcal{X}^m \text{ s.t. } \text{err}_{\mathcal{Q}}(x, y^*) \leq \alpha.$$

Proof sketch: Independently draw:

$$I_1, \dots, I_m \sim \text{Unif}\{1, \dots, n\}, \quad Y_j = x_{I_j}$$

For fixed $f \in \mathcal{Q}$: $0 \leq f(Y_j) \leq 1$ and $\mathbb{E}[f(Y_j)] = \frac{1}{n} \sum_{i=1}^n f(x_i) = f(x)$.

$$\Pr \left[\left| \frac{1}{m} \sum_{i=1}^m \overbrace{f(Y_j)}^{=f(Y)} - f(x) \right| > \alpha \right] \leq 2e^{-2m\alpha^2} \quad (\text{Hoeffding's inequality})$$

$$\Pr_Y [\text{err}_{\mathcal{Q}}(x, y) > \alpha] \leq \sum_{f \in \mathcal{Q}} \Pr [|f(y) - f(x)| > \alpha] \leq k \cdot 2 \cdot e^{-2m\alpha^2} \leq \frac{1}{2} < 1.$$

$\Rightarrow \forall$ fixed $x \in \mathcal{X}^n \exists y^* \in \mathcal{X}^m$: $\text{err}_{\mathcal{Q}}(x, y^*) < \alpha$ as long as $m = \left\lceil \frac{\log(4k)}{2\alpha^2} \right\rceil$.

SmallDB

Public parameters $\mathcal{X}, n, \varepsilon, \alpha; K=|\mathcal{Q}|$.

$$m = \left\lceil \frac{\log(4K)}{2\alpha^2} \right\rceil, \quad \mathcal{R} = \mathcal{X}^m, \quad |\mathcal{R}| = |\mathcal{X}|^m.$$

For every $y \in \mathcal{R}$: $u(x, y) = -\text{err}_{\mathcal{Q}}(x, y) = -\max_{f \in \mathcal{Q}} |f(y) - f(x)|$.

(Show that:) $\Delta u = \max_{y \in \mathcal{R}} \max_{x \sim x'} |u(x, y) - u(x', y)| \dots \leq \frac{1}{n}$.

$\alpha/2$

SmallDB($x, \mathcal{Q}, \varepsilon, \alpha$):

1) Sample and release $\hat{y} \in \mathcal{R}$: $\mathcal{X}^m, m = \left\lceil \frac{\log(4K)}{2(\alpha/2)^2} \right\rceil$.

$$\Pr[\hat{y} = y] = \frac{\exp\left(-\frac{n \cdot \varepsilon \cdot u(x, y)}{2}\right)}{\sum_{z \in \mathcal{R}} \exp\left(-\frac{n \cdot \varepsilon \cdot u(x, z)}{2}\right)}$$

2) Release $d_f = f(\hat{y}) \quad \forall f \in \mathcal{Q}$.

- SmallDB is ϵ -DP. (by the guarantees of the Exp. Mech.) ✓
- Accuracy

let $\text{OPT} = \min_{y \in R} \text{err}_Q(x, y)$. By the guarantees of the Exp. Mech.:

$$\Pr \left[\text{err}_Q(x, \hat{y}) > \text{OPT} + \frac{2}{\epsilon n} (\log |R| + \log(1/\delta)) \right] \leq \delta. \quad (1)$$

- $\text{OPT} < \alpha$. (by existence of y^* achieving OPT for $m = \lceil \frac{\log(4k)}{2\alpha^2} \rceil$).

- $\log |R| = m \cdot \log |X|$, $m \leq \frac{\log(4k)}{2\alpha^2} + 1$

substitute in (1)

$$\Rightarrow \text{w.p. } 1 - \delta: \text{err}_Q(x, \hat{y}) \leq \alpha + \frac{\log |X| \cdot \log(4k)}{\epsilon n \alpha^2} + \frac{2(\log |X| + \log(1/\delta))}{\epsilon n}$$

To achieve $\star$: set $\alpha = \alpha/2$ in the beginning of the algorithm, (m will change accordingly). Then we say that $\star$ is satisfied as

long as
$$n \geq \frac{8 \cdot \log |X| \cdot \log(4k)}{\epsilon \alpha^2} + \frac{8(\log |X| + \log(1/\delta))}{\epsilon \alpha} \quad (\text{double-check!})$$