

# Theory of Differential Privacy

EN.601.438/638, Fall 2026

## Problem Set 2: Linear Queries

100 points • 15% of course grade • Due: October 21, 23:59

**Submission.** Submit one PDF on Gradescope. Include your name and list, for each problem, any collaborators and outside resources or tools you used. Unless a problem explicitly asks only for discussion, justify every answer with a proof.

**Collaboration.** You may discuss definitions, concepts, and high-level ideas with classmates and course staff. You may not view another student's written solution, jointly compose a proof, or reuse text. After a discussion, put away your notes and write your solution independently from scratch.

**Generative AI.** You may use an AI tool as a peer to clarify definitions, request a new example, or quiz yourself. Do not paste the assignment questions into a tool, request a solution or solution outline, keep the assignment open while using the tool, or submit AI-generated reasoning. Close the tool before writing your solution, and disclose how you used it. Some problems or closely related exercises may appear in the online materials for the BU/NEU or Waterloo courses, as well as previous iterations of this course; you may not consult or copy solutions, if any, from those courses.

**Conventions.** Unless stated otherwise, datasets have fixed size and neighboring datasets differ in one row (replace-one adjacency). All logarithms are natural. Privacy and failure parameters such as  $\delta$  and  $\beta$  lie in  $(0, 1)$  unless stated otherwise. Problems 3 and 4 use raw counts and sums, whereas in lectures we usually use normalized quantities. If you consult lecture notes, make sure to account for the  $\frac{1}{n}$  factor.

### Facts you may use

- If  $Z \sim \text{Lap}(b)$ , then for every  $t \geq 0$ ,

$$\Pr(|Z| \geq t) = e^{-t/b}, \quad \Pr(Z \geq t) = \Pr(Z \leq -t) = \frac{1}{2}e^{-t/b}.$$

- If a vector-valued function  $G$  has  $\ell_1$ -sensitivity at most  $\Delta$ , then releasing

$$G(x) + (Z_1, \dots, Z_k), \quad Z_i \stackrel{\text{iid}}{\sim} \text{Lap}(\Delta/\varepsilon),$$

is  $\varepsilon$ -differentially private.

- If  $Z_1, \dots, Z_k \stackrel{\text{iid}}{\sim} \text{Lap}(b)$ , then with probability at least  $1 - \beta$ ,

$$\max_{i \leq k} |Z_i| \leq b \log(k/\beta).$$

- You may use the shorthand

$$c_{\varepsilon, \delta} = \frac{\sqrt{2 \log(2/\delta)}}{\varepsilon}$$

for the privacy-dependent constant of the Gaussian mechanism.

- For a matrix  $A$ ,  $\|A\|_{1 \rightarrow 2}$  is the largest Euclidean norm of a column of  $A$ , and  $\|A\|_F$  is its Frobenius norm.

**Problem 1. Synthetic databases are essentially without loss** [25 points]

Let  $\mathcal{X}$  be a finite data universe, and let  $\Delta(\mathcal{X})$  be the set of all probability distributions over  $\mathcal{X}$ . A linear query is specified by a function  $f : \mathcal{X} \rightarrow [0, 1]$  and is evaluated on a distribution  $p \in \Delta(\mathcal{X})$  by

$$f(p) = \sum_{x \in \mathcal{X}} p(x)f(x).$$

Let  $\mathcal{Q}$  be a finite set of linear queries. Suppose we are given a vector of answers

$$a = (a_f)_{f \in \mathcal{Q}}$$

such that, for some unknown distribution  $p^* \in \Delta(\mathcal{X})$  and some  $\alpha \geq 0$ ,

$$|a_f - f(p^*)| \leq \alpha \quad \text{for every } f \in \mathcal{Q}.$$

Think of  $a$  as a vector of noisy answers to the queries in  $\mathcal{Q}$ . Let  $\hat{p}$  be any minimizer of

$$\max_{f \in \mathcal{Q}} |f(p) - a_f|$$

over  $p \in \Delta(\mathcal{X})$ . Such a minimizer exists because  $\Delta(\mathcal{X})$  is compact and the objective is continuous.

(a) **Accuracy.** [14 points] Prove that

$$\max_{f \in \mathcal{Q}} |f(\hat{p}) - f(p^*)| \leq 2\alpha.$$

(b) **Privacy interpretation.** [11 points] Suppose the answer vector  $a$  was produced by an  $\varepsilon$ -DP mechanism. Explain why releasing  $\hat{p}$  is also  $\varepsilon$ -DP. Then explain, in a few sentences, why this shows that outputting a synthetic distribution is information-theoretically without loss up to a factor of 2 in error for a fixed query class.

**Problem 2. Factorization mechanisms** [25 points]

In this problem, compare mechanisms by average root mean squared  $\ell_2$  error. Ignore the common factor  $1/n$ . For a workload  $F \in \mathbb{R}^{k \times m}$  and a factorization  $F = RM$ , the Gaussian factorization mechanism has error term

$$\frac{c_{\varepsilon, \delta} \|M\|_{1 \rightarrow 2} \|R\|_F}{\sqrt{k}}.$$

Directly answering  $F$  corresponds to the trivial factorization  $F = I_k F$ , and has error term  $c_{\varepsilon, \delta} \|F\|_{1 \rightarrow 2}$ .

(a) **A concrete workload.** [10 points] Consider the workload matrix

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

Compute  $\|F\|_{1 \rightarrow 2}$ . What is the average root mean squared  $\ell_2$  error term for the direct Gaussian mechanism?

Find matrices  $R$  and  $M$  with  $F = RM$  such that

$$\|M\|_{1 \rightarrow 2} \|R\|_F < 2\|F\|_{1 \rightarrow 2}.$$

Compute the two sides to verify the improvement.

- (b) **Composing factorizations. [7 points]** Suppose  $F = RM$  is a factorization of a workload  $F$ , and suppose another workload  $F'$  can be written as

$$F' = R'F.$$

Consider the mechanism that privately answers the strategy workload  $M$ , reconstructs answers to  $F$  using  $R$ , and then reconstructs answers to  $F'$  using  $R'$ . Write the final output in the form

$$F'h_x + \text{noise}.$$

Express the average root mean squared  $\ell_2$  error term in terms of  $M$ ,  $R$ , and  $R'$ .

- (c) **Binary tree as a factorization for  $D = 8$ . [8 points]** Let  $F \in \{0, 1\}^{8 \times 8}$  be the prefix workload

$$F_{t,a} = \mathbf{1}\{a \leq t\}, \quad t, a \in \{1, \dots, 8\}.$$

Let  $\mathcal{T}$  be the set of binary-tree intervals over  $\{1, \dots, 8\}$ :

$$\begin{aligned} &[1, 8], \\ &[1, 4], [5, 8], \\ &[1, 2], [3, 4], [5, 6], [7, 8], \\ &[1, 1], [2, 2], [3, 3], [4, 4], [5, 5], [6, 6], [7, 7], [8, 8]. \end{aligned}$$

Define  $M \in \{0, 1\}^{15 \times 8}$  by

$$M_{I,a} = \mathbf{1}\{a \in I\}.$$

For each  $t \in \{1, \dots, 8\}$ , decompose  $[1, t]$  into disjoint intervals from  $\mathcal{T}$ . Use these decompositions to define a reconstruction matrix  $R \in \{0, 1\}^{8 \times 15}$  and prove that

$$F = RM.$$

### Problem 3. Rectangle queries

[30 points]

This problem uses raw counts. The data universe is the two-dimensional grid  $[D]^2$ , where  $[D] = \{1, \dots, D\}$ . A dataset is

$$x = ((x_1, y_1), \dots, (x_n, y_n)) \in ([D]^2)^n.$$

For  $1 \leq s \leq t \leq D$  and  $1 \leq u \leq v \leq D$ , define the rectangle query

$$f_{s,t}^{u,v}(x) = |\{i : s \leq x_i \leq t \text{ and } u \leq y_i \leq v\}|.$$

- (a) **All rectangles. [8 points]** How many rectangle queries are there? Let  $R(x)$  be the vector containing the answers to all rectangle queries. Prove that the  $\ell_1$ -sensitivity of  $R$  is  $\Theta(D^4)$ . If we use the vector Laplace mechanism to answer all rectangle queries directly with privacy parameter  $\varepsilon$ , what noise scale is sufficient for each query?

- (b) **Corner-aligned rectangles.** [8 points] A corner-aligned rectangle is a rectangle of the form

$$[1, t] \times [1, v].$$

Let  $C(x)$  be the vector of all corner-aligned rectangle-query answers. Prove that the  $\ell_1$ -sensitivity of  $C$  is  $\Theta(D^2)$ . Show how to express an arbitrary rectangle query  $f_{s,t}^{u,v}$  as a linear combination of four corner-aligned rectangle queries.

- (c) **A two-dimensional binary tree mechanism.** [14 points] Assume  $D$  is a power of 2. Generalize the one-dimensional binary tree mechanism to answer all rectangle queries. One natural strategy workload consists of dyadic rectangles  $I \times J$ , where  $I$  and  $J$  are intervals from the one-dimensional binary tree over  $[D]$ .

Describe the mechanism and prove that it is  $\varepsilon$ -DP. Then prove that, with probability at least  $1 - \beta$ , all rectangle queries are answered with error

$$O\left(\frac{\log^a D + \log^{a-1} D \cdot \log(1/\beta)}{\varepsilon}\right)$$

for some universal constant exponent  $a$ . You do not need to optimize  $a$ , but you should state the value of  $a$  obtained by your analysis.

#### Problem 4. Private continual observation

[20 points]

In this problem, records arrive online. There are  $T$  public release times, numbered  $1, \dots, T$ . Record  $i$  has an arrival time  $t_i \in [T]$  and value  $x_i \in [0, 1]$ . At every time  $t$ , the goal is to release a private approximation to the cumulative sum

$$f_t(x) = \sum_{i:t_i \leq t} x_i.$$

The algorithm does not know records whose arrival time is in the future. Neighboring streams differ in one record  $(t_i, x_i)$ .

- (a) **A naive approach.** [4 points] Suppose we privately release each time increment

$$g_t(x) = \sum_{i:t_i=t} x_i$$

with independent Laplace noise and estimate  $f_t(x)$  by summing the first  $t$  noisy increments. Explain why the error grows like  $\Omega(\sqrt{t}/\varepsilon)$  for some  $t$ .

- (b) **Tree aggregation over time.** [9 points] Assume  $T$  is a power of 2. Show how to use a binary tree over the time indices  $1, \dots, T$  to privately release all cumulative sums  $f_t(x)$  while processing records online. State precisely which noisy tree interval sums are released, and explain why the algorithm does not need to know future records when it releases the answer at time  $t$ .

- (c) **Privacy and error.** [7 points] Prove that your mechanism is  $\varepsilon$ -DP. Then prove that, with probability at least  $1 - \beta$ , all  $T$  released cumulative sums have error

$$O\left(\frac{\log^3 T + \log^2 T \cdot \log(1/\beta)}{\varepsilon}\right).$$

- (d) **Space complexity.** [5 bonus points] Show that the algorithm can be implemented using only  $O(\log T)$  stored partial sums at any time.