Privacy in Statistics and Machine Learning — Spring 2025
In-class Exercises for Lecture 7 (The Binary Tree Mechanism)
February 11, 2025

Adam Smith (based on materials developed with Jonathan Ullman)

Problems with marked with an asterisk (*) are more challenging or open-ended.

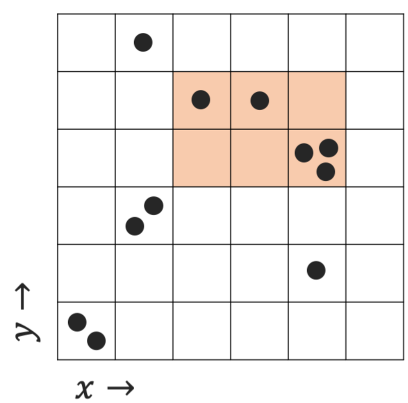
1. (Exercise 2.1 in the notes) Let $F : [D]^n \to \mathbb{R}^{\binom{D}{2}}$ that takes a dataset $\mathbf{x}$ and outputs a vector of answers containing $f_{s,t}(\mathbf{x})$ for every $1 \le s \le t \le D$. Prove that the global sensitivity of F is $\Theta(D^2)$.

2. (Exercise 2.4 in the notes) Let $\mathcal{T}$ be the set of intervals in the binary tree mechanism with domain size D that is a power of 2. Prove that for every t, we can express the interval $\{1, \ldots, t\}$ as the union of at set of at most $\log_2 D$ intervals in $\mathcal{T}$. For example, $\{1, \ldots 11\} = \{1, \ldots, 8\} \cup \{9, 10\} \cup \{11, 11\}$.

3. In this question we'll generalize the ideas of the binary tree mechanism to answer **rectangle queries**. Here the data universe is the two-dimensional grid with side length D, and each datapoint is a pair $(x_i, y_i) \in [D]^2$. A rectangle query $f_{s,t}^{u,v}$ is defined by ranges $1 \le s \le t \le D$ and $1 \le u \le v \le D$ and

$$f_{s,t}^{u,v}(\mathbf{x}) = \#\{i : s \le x_i \le t \text{ and } u \le y_i \le v\} \tag{1}$$

Here's an example depicting data in the domain $[6]^2$. Black dots are the data points (the position within the grid cells is irrelevant) and the orange shaded area represents the rectangle query $f_{3,5}^{4,5}$, whose answer on this dataset is 5.



   (a) How many rectangle queries are there? What is the global sensitivity of the set of all rectangle queries on the domain $[D]^2$? If we use the Laplace mechanism to answer all such queries, how much noise do we add to each query?

   (b) Suppose we only want to answer the subset of queries called *corner-aligned rectangle queries.* This is the subset of rectangle queries that include the lower-left corner $(1, 1)$, and have the form $f_{1,t}^{1,v}$.

i. Using the Laplace mechanism, how much noise would we add to answer all corner-aligned rectangle queries?

ii. How can you express any rectangle query $f_{s,t}^{u,v}$ as a linear combination of a small number of corner-aligned rectangle queries?

iii. How much noise would we incur if we use the Laplace mechanism to answer all corner-aligned rectangle queries and then recover the answer to the other rectangle queries?

(c) **(*)** Generalize the binary-tree mechanism to answer all rectangle queries with error $O(\frac{1}{\varepsilon}\log^a D)$ for some constant exponent a.

4. **(Privacy Under Continual Observation)** Suppose we have data records $x_1, \ldots, x_n$ in $[0, 1]$ that arrive online, as an algorithm is running. Specifically, suppose there are T release times that are decided a priori. Each record has an arrival time $t_i \in \{1, \ldots, T\}$ that indicates the release time soonest after its arrival. At each of these times, we wish to release an ε-differentially private approximation to the sum of data records that have arrived so far. That is, we want to release

$$f_t(\mathbf{x}) = \sum_{i:\ t_i \leq t} x_i \quad \text{for } t = 1, \ldots, T\,.$$

The algorithm does not know the records ahead of time. That is, it must release an approximation to $f_t(\mathbf{x})$ without knowing the records which arrive after time t.

We could release T independent estimates $a_1, \ldots, a_T$, where a_t consist of the sum of records with $t_i = t$. However, using these to estimate $f_t(\mathbf{x})$ would incur error $\Omega(\sqrt{t})$. (Why?)

(a) Show how the binary tree scheme can be used to approximate all the values $f_t(\mathbf{x})$ with error $O\big(\frac{1}{\varepsilon}\log^3 T\big)$ by an ε-differentially private algorithm that learns the data records only as they arrive.

(b) (*) Show that the algorithm above can be implemented using only $O(\log T)$ space (that is, we don't need to remember the entire input, just $O(\log T)$ numbers in $[0, n]$).