601.438 / 601.638 Algorithmic Foundations of Differential Privacy Spring 2025
Homework #2 **Due:** March 13, 2025, 9:00am

Reminder: you may work in groups of up to three people, but must write up solutions entirely on your own. Collaboration is limited to discussing the problems – you may not look at, compare, reuse, etc. any text from anyone else in the class. Please include your list of collaborators on the first page of your submission. Many of these problems have solutions which can be found on the internet – please don't look. You can of course use the internet (including the links provided on the course webpage) as a learning tool, but don't go looking for solutions.

Please include proofs with all of your answers, unless stated otherwise.

---

1 Synthetic Databases (50 points)

In class we saw a few ways of releasing a synthetic database (or data structure) to answer linear queries a set of queries Q. A natural question is whether this cost us something, or was without loss of generality. That is, by insisting on releasing a synthetic database rather than just releasing answers to the queries in Q, do we actually incur more loss than is necessary? Let's show that this is not the case.

More formally, as in class, let X be the domain of rows and let $\mathcal{D}$ be the set of probability distributions over X (so $\sum_{x \in X} D_x = 1$ and $D_x \geq 0$ for all $x \in X$). Let Q be a set of (normalized) linear queries. Let $\vec{a} = (a_f)_{f \in Q}$ be a set of values so that $|a_f - f(D)| \leq \alpha$ for all $f \in Q$. Think of these as (noisy) answers to the queries in Q. Prove that we can *post-process* $\vec{a}$ to produce a synthetic database $\hat{D} \in \mathcal{D}$ so that $\max_{f \in Q} |f(\hat{D}) - f(D)| \leq 2\alpha$ (do not worry about the running time of your algorithm, just that it is post-processing, i.e., it does not use knowledge of D or even of α).

Note that this implies that outputting a synthetic database is essentially (up to a factor 2) without loss of generality: if we had some other way of outputting answers to the queries in Q, we could just turn that into a synthetic database.

2 Rectangle Queries (50 points)

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 M, and each datapoint is a pair $(x_i, y_i) \in [M]^2$. A rectangle query $f_{s,t}^{u,v}$ is defined by $1 \leq s \leq t \leq M$ and $1 \leq u \leq v \leq M$, and is equal to the number of datapoints in D that are in the associated rectangle, i.e.,

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

Let $\mathcal{R}$ denote the set of all rectangle queries.

(a) (10 points) What is the global sensitivity of $\mathcal{R}$? Give your answer in Θ-notation, and prove it.

(b) (10 points) Let $\mathcal{C}$ denote the set of *corner-aligned* rectangle queries. This is the subset of rectangle queries that include the lower-left corner $(1, 1)$, i.e., $\mathcal{C}$ has the rectangle queries of

the form $f_{1,t}^{1,v}$. What is the global sensitivity of $\mathcal{C}$? As before, give your answer in Θ-notation and prove it.

(c) (30 points) Generalize the binary tree mechanism to answer all rectangle queries with error $O(\frac{1}{\epsilon}\log^a M$ for some constant exponent a.