HW 4: Differential Privacy Foundations

CS 208 Applied Privacy for Data Science, Spring 2025

Version 1.1: Due Fri, Feb. 28, 5:00pm.

Instructions: Submit a PDF file that contains both your written responses as well as your code to the assignment on Gradescope. Read the section "Collaboration & AI Policy" in the syllabus for our guidelines regarding the use of LLMs and other AI assistance on the assignments.

1. **User-level vs. Event-level Privacy.** [SV: replaced the previous problem on privacy units with this one. It seems that the some of the answers to the previous problem are already in HoDP, which calculates sensitivities in terms of L and U? You could instead incorporate some of that into the section notes, but be sure to stay consistent with the formalism we're using here in this new problem.][SV: lots of background here that is summarizing stuff that is in, or implicit in, HoDP and the lectures. Shall we move it to the section notes and just point the students there? (But then section notes need to be published asap, and we might need more time for getting those ready.)] Recall that differential privacy is defined with respect to a dataset space $\mathcal{X}$ and an adjacency relation $\sim$ on $\mathcal{X}$ that determines our *privacy unit*: DP protects information that can differ between adjacent datasets.

   A more general formulation uses dataset *metrics* $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ as covered in HoDP Chapter 3. Given a dataset metric, we can obtain an adjacency relation by defining $x \sim_d x'$ if $d(x, x') \leq 1$. We have seen two examples for the dataset space $\mathcal{X} = \mathcal{R}^* \overset{\text{def}}{=\!=} \bigcup_{n=0}^{\infty} \mathcal{R}^n$, where $\mathcal{R}$ is the space from which individual records (or rows) come:[1]

   - *Hamming distance*: if we have two datasets $x, x' \in \mathbb{R}^*$ such that $|x| = |x'| = n$ (where $|x|$ is the number of records in x), then $d_{Ham}(x, x') \overset{\text{def}}{=\!=} |\{i : x_i \neq x'_i\}|$. If $|x| \neq |x'|$, then we define $d_{Ham}(x, x') = \infty$. The induced adjacency relation $\sim_{Ham}$ is the "change one row" notion of adjacency, which we used to formalize *bounded DP* where the number n of records is treated as publicly known.
   - *Symmetric distance*: if we have two datasets $x, x' \in \mathbb{R}^*$, then $d_{Sym}(x, x')$ is the size of the symmetric difference between x and x' as multisets, i.e. the number of records that need to be added or removed from x for it to have the same number of occurrences of each element as x'. $\sim_{Sym}$ is the "add or remove a row" notion of adjacency, which we used to formalize *unbounded DP*, where the number n of records needs also to be protected with DP.

   Another difference between d_{Ham} and d_{Sym} is that the former depends on the order of elements in x and x' and the latter does not. It is possible to define other metrics to bridge this distinction, but we do not do so here.

   Recall that a transformation $T : \mathcal{X} \to \mathcal{Y}$ between two metric spaces (with input metric d_{IM} and output metric d_{OM}) is called *c-stable* if for all $x, x' \in \mathcal{X}$, we have $d_{OM}(T(x), T(x')) \leq c \cdot d_{IM}(x, x')$. Stable transformations are useful because they compose well with differentially

[1] In class, we often used $\mathcal{X}$ for the row space, and $\mathcal{X}^n$ for the dataset space.

private mechanisms: If T as above is c-stable with respect to d_{IM} and d_{OM} and $M : \mathcal{Y} \to \mathcal{Z}$ is ε-DP with respect to $\sim_{OM}$, then their chaining $(M \circ T)(x) \stackrel{\text{def}}{=} M(T(x))$ is $c\varepsilon$-DP. We can understand the privacy analysis of the Laplace mechanism using this fact, by taking $\mathcal{Y} = \mathbb{R}$, $d_{OM} = d_{Abs}$, and $M(y) = y + \text{Lap}(1/\varepsilon)$. Then T being c-stable is the same as saying its global sensitivity is bounded as $\Delta T \leq c$, and M being ε-DP with respect to $\sim_{Abs}$, amounts to saying that the pdfs of $y + \text{Lap}(1/\varepsilon)$ and $y' + \text{Lap}(1/\varepsilon)$ are within an e^{ε} factor if $|y - y'| \leq 1$.

We can use this formalism to distinguish "user-level" and "event-level" privacy and reason about "contribution bounding" as often comes up in DP deployments (including the Wikimedia release you read about). In the rest of this problem, take $\mathcal{X} = \mathcal{R}^*$ for a row space of the form $\mathcal{R} = \mathcal{I} \times \mathcal{D}$, where $\mathcal{I}$ is a space of userids and $\mathcal{E}$ is space of events associated with a user (e.g. a user edited a particular Wikipedia page at a particular time). So each record is of the form $(\texttt{userid}, \texttt{event})$ and the same `userid` can appear arbitrarily many times in the dataset. [ZR: should we be more explicit here that the number of times that a userid can appear in the dataset is not bounded? Will this make problem (a) below too easy?] [SV: changed "multiple times" to "arbitrarily many" times] For two datasets $x, x' \in \mathcal{X}$, we can define $d_{User}(x, x')$ to be the number of additions or removals of *users* to transform x into x'. Here removing a user means removing all records with a given `userid`, and adding a user means adding all records with a given `userid`. *User-level privacy* means DP with respect to $\sim_{User}$. In contrast, *event-level privacy* simply means DP with respect to the usual $\sim_{Sym}$ or $\sim_{Ham}$ in the case of bounded DP.

(a) Consider the a query $q : \mathcal{X} \to \mathbb{R}$ defined by $q(x) = |x|$. What is the global sensitivity Δq with respect to $\sim_{User}$ and with respect to $\sim_{Sym}$?

(b) Consider the transformation $T : \mathcal{X} \to \mathcal{X}$ where $T(x)$ keeps only the first $\leq B$ records associated with each `userid` (and just drops the remaining ones). For each pair of metrics $d_{IM}, d_{OM} \in \{d_{Sym}, d_{Ham}, d_{User}\}$, calculate the smallest c such that T is a c-stable transformation from d_{IM} to d_{OM}. [SV: is all 6 pairs overkill?] [ZR: I think keeping the 6 pairs is fine. You want them to pick IM and OM distinct?] [SV: Oh right, we have 9 pairs. (I didn't mean for them to be distinct.) However, many of them should be easy once they've done a few of them. But please test-solve it to check.][JT: I think 9 is reasonable and many are quick.]

(c) Combining Part 1a and Part 1b with $IM = User$ and an appropriate choice of OM, deduce a bound on global sensitivity of $(q \circ T) : \mathcal{X} \to \mathbb{R}$ with respect to $\sim_{User}$.

2. **Regression:** [ZR: keep/adjust this problem?] [SV: I really like it. Made some comments below. Also would like to hear feedback on the experimental part - is it clear enough? not too tedious? and gives interesting results?] Consider a dataset where each of its n rows is a pair of real numbers (x_i, y_i), each from an interval $[-b, b]$. Suppose we wish to find a best-fit linear relationship $y_i \approx \beta x_i$ between the y's and the x's. Non-privately, a standard way to estimate β is via the OLS regression formula

$$\hat{\beta} = \hat{\beta}(x, y) = \frac{S_{xy}}{S_{xx}} = \frac{\sum_i x_i y_i}{\sum_i x_i^2}.$$

This is called *ordinary least-squares (OLS)* regression, since $\hat{\beta}$ is the minimizer of the mean-squared residuals

$$\frac{1}{n}\sum_i (y_i - \hat{\beta}x_i)^2. \tag{1}$$

(a) Show that the function $\hat{\beta}(x, y)$ has infinite global sensitivity with respect to $\sim_{Ham}$, [SV: added "wrt $\sim_{Ham}$"] and hence we cannot get a useful DP estimate of it via a direct application of the Laplace or Gaussian mechanisms.

(b) Show that S_{xy} and S_{xx} have global sensitivity that is bounded solely as a function of b, and hence each of these can be approximated in a DP manner using the Laplace mechanism.

(c) Using Part 2b together with a stable clipping transformation, chaining, [SV: added mention of a stable clipping transformation and chaining] basic composition, and post-processing, devise and implement an ε-DP algorithm for approximating $\hat{\beta}$ on an arbitrary dataset with $x_i, y_i \in \mathbb{R}$. [SV: do we want them to start from scratch, or use one of James' notebooks or some OpenDP code as a starting point?] [ZR: This is a fairly simple function, so I think it's ok for them to start from scratch] In addition to the dataset $((x_1, y_1), \ldots, (x_n, y_n))$, your implementation should take as input parameters a clipping bound b and the privacy-loss parameter ε.

(d) Evaluate the performance of your algorithm using a Monte Carlo simulation with synthetic data. Set $\varepsilon = .1$, $b = 1$, generate the x_i's uniformly at random from $[-1/2, 1/2]$, and generate the y_i's according to a linear model with slope 1 and Gaussian noise, but clipped to $[-1, 1]$ to satisfy the range requirements: [SV: if they are supposed to include clipping as part of their DP mechanism, then we shouldn't clip the y_i's for them, right?] [JT: In the initial description before part (a), we say we're only looking at datasets where (x_i, y_i) are each generated from an interval $[-b, b]$. However, in part (c), we ask them to generate a function where (x_i, y_i) can come from all of $\mathbb{R}^2$. So I think the "clipping" in the previous sentence is to remain faithful to the initially-described data generation, but maybe we should make the two descriptions consistent]

$$y_i = [x_i + \mathcal{N}(0, .02)]_{-1}^{1}.$$

For each $n = 100, 200, 300, \ldots, 5000$, run many Monte Carlo trials to estimate and plot the bias and standard deviation of both the OLS estimate $\hat{\beta}$ and the DP estimate $\tilde{\beta}$. Your plot should have n on the x-axis, and bias and standard deviation on the y-axis on a scale from -1.0 to 1.0, with the endpoints also representing values of magnitude larger than 1. Try to run enough trials to obtain smooth curves.

(If $\hat{\theta} = \hat{\theta}(z)$ is an estimator of a population parameter θ based on a dataset z, then the *bias* of $\hat{\theta}$ is $\mathrm{E}[\hat{\theta} - \theta]$, where the expectation is taken over both the dataset z and any randomization used by estimator $\hat{\theta}$; note that the bias can be positive or negative—do keep track of the sign. The "bias-variance tradeoff" says that the MSE of an estimator is the sum of its squared bias and its variance; in previous homeworks, we evaluated the (R)MSE of DP estimators, now we are doing a finer analysis by separating the MSE into the bias and variance.)

(e) Try to give an intuitive explanation of the source of the bias you see in Part 2d and on what kinds of dataset distributions this might be largest. How might bias in particular (not just MSE) have an impact on downstream applications?

[ZR: we will want to update this with the new final project guidelines. I think James posted a new one in the slack channel]

3. **Final Project Ideas** The final projects are an important focus of this course, and we want you to start thinking about yours as soon as possible. Please read the "Final Project Guidelines" (to be posted shortly on the course webpage: `https://github.com/opendp/cs208/blob/main/spring2025/final%20projects/Final%20Project%20Guidelines%202025.pdf`) and submit about a paragraph as described in the "Topic Ideas" bullet.

Collaborators

Please list all collaborators for this problem set. ChatGPT and other AI tools should be treated similarly to collaboration with your peers in the class. You may use these tools to help you understand the material and as part of your brainstorming process, but you should not be asking the tools to solve the homework problems for you. If you do use such tools, you must cite them and list the prompts you entered and responses obtained below.