# HW 7: The Exponential Mechanism and Attacks on Utility

CS 208 Applied Privacy for Data Science, Spring 2025

**Version 1.1: Due Fri, Mar. 28, 11:59pm.**

**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. **Continuous Exponential Mechanism and Data-Dependent Clipping Bounds:** In all of the parts below, the dataset is $x \in [0, B]^n$ and our adjacency notion is $\sim_{Ham}$. In all of the implementation parts, you should write code that takes as input $B \geq 0$, $n \in \mathbb{N}$, $x \in [0, B]^n$, and $\varepsilon > 0$.

   (a) For a dataset $x \in [0, B]^n$ and $\alpha \in [0, 1]$, the $\alpha$'th *quantile* of $x$ is $q_\alpha = (\mathrm{sort}(x)_{\lfloor \alpha n \rfloor} + \mathrm{sort}(x)_{\lceil \alpha n \rceil})/2$. HoDP Chapter 4 suggests the following score function for using the exponential mechanism to estimate $q_\alpha$:

   $$s(x, y) = -|(1-\alpha) \cdot \#\{i : x_i < y\} - \alpha \cdot \#\{i : x_i > y\}| .$$

   Later in HoDP Chapter 4, it is described how to implement the exponential mechanism when the output domain is a continuous interval, like $\mathcal{Y} = [0, B]$, provided the score function $s(x, \cdot)$ is piecewise constant on $\mathcal{Y}$ (as is the above score function). Put the above together to implement a differentially private algorithm that outputs a differentially private estimate of the $\alpha$'th quantile, for any desired $\alpha \in [0, 1]$ specified as an additional parameter.

   (b) In section, it was shown that the following algorithm for estimating the *Winsorized* mean is *not* $\varepsilon$-DP:

   $$M(x) = \frac{1}{n} \cdot \sum_{i=1}^{n} [x_i]_{q_{.05}}^{q_{.95}} + \mathrm{Lap}\left(\frac{B}{\varepsilon n}\right),$$

   where $[x]_a^b$ is defined as in HW3.

   In contrast, the following algorithm *is* $\varepsilon$-DP: use your algorithm from Part 1a to get $\varepsilon/3$-DP estimates $\hat{q}_{.05}$ and $\hat{q}_{.95}$ of the .05 and .95 quantiles, and output

   $$M(x) = \frac{1}{n} \cdot \left(\sum_{i=1}^{n} [x_i]_{\hat{q}_{.05}}^{\hat{q}_{.95}}\right) + \mathrm{Lap}\left(\frac{3(\hat{q}_{.95} - \hat{q}_{.05})}{\varepsilon n}\right).$$

   What DP properties does $M$ use that makes it $\varepsilon$-DP, even though the algorithm from section is not?

   (c) The dataset `FultonPUMS5full.csv` provides the 5% PUMS Census file for Fulton. For $\varepsilon = 1$ and each $B \in \{5 \times 10^5, 5 \times 10^6, 5 \times 10^7\}$, estimate the RMSE of DP mean income for each PUMA in Fulton.[1] Run this analysis to compare (i) the ordinary Laplace mechanism

[1] You can assume that the size of each PUMA dataset is public information.

for a mean and (ii) the algorithm from Part 1b. Show box-and-whisker plots of the DP mean incomes for each PUMA and algorithm, noting the true means. (In the GitHub repo, we have given you `hw6_starter.py` for producing such plots comparing the your algorithm from Part 1b to the ordinary Laplace mechanism). Order PUMA by mean income, or perhaps skew of income, or anything you think reveals an interesting pattern. Give an intuitive explanation of the cases (datasets and parameter settings) in which algorithm (i) performs better than algorithm (ii) and vice-versa.

2. **Attacking Utility:** Recall the Wikimedia Foundation's deployment of DP for pageviews-by-country statistics. The Wikimedia Foundation also currently publishes the total number of pageviews (globally, i.e., from users across the world) on Wikipedia articles per day, but without any privacy protection.

   Imagine that Wikimedia is now considering using DP to protect global pageview counts under $\epsilon$-DP. Also imagine that a global news publication uses these counts to determine which culture and media topics to devote extra coverage toward, presumably based on people's interests. The news publication is concerned that the use of DP will make it harder for them to identify the top $k$ pages that fall under the culture and media category.[2] Your goal is to help the news publication and Wikimedia understand the utility of the data protected under DP for this task. Assume that, for your evaluation, you have access to the currently published global pageviews, which are not protected under DP.

   (a) Describe in detail your approach to evaluating utility of the DP-noised data for the news organization's task. In your response, specify (1) what metric(s) you would use and why, (2) the simulations you would run, and (3) comparisons you would make. Would you test different privacy-loss parameters? If so, how and why? If not, why not?

   (b) Suppose a different data user of Wikipedia pageviews depends not only on information about the top $k$ pages in a category, but also the specific numbers of pageviews per page. Briefly describe an example of a such a use case.

   (c) If you were now to evaluate utility of the DP-noised data for the use case you defined above, how would you change your evaluation approach? Would your metric(s) change, and why?

3. **Form Project Groups and Revise Topic Ideas:**

   (a) By Monday 3/24 (right after spring break), use the spreadsheet to self-organize into groups of size 3–5 students. You should start by emailing the topic proposer (Column D), who should take first responsibility for coordinating those interested in the topic. (The rest of you can share your email addresses in Columns E+ as well). Once you've settled on a group, put all other occurrences of your name in the sheet in strikethrough format (but don't remove it, in case group shuffling is needed). If there are 6+ people interested in a topic, you should split into smaller groups. Write to the course staff on Ed if you need help in group formation.

[2] Wikimedia uses a machine learning model to predict an article's topic area, and publishes the article's predicted topic area along with its pageviews. For the purposes of this question, you can treat these predicted categories as ground truth. However, if you are curious about the model Wikimedia uses, see here: `https://meta.wikimedia.org/wiki/Machine_learning_models/Production/Language_agnostic_link-based_article_topic`.

(b) By Friday 3/28, your group should submit a paragraph with a revised set of ideas for your project (ideally all coherent with the general topic described in the spreadsheet, but with some alternatives in the specific questions to be asked or the methodology to be used). See the Final Project Guidelines. This will be a separate Gradescope submission from the rest of your homework, with one submission per group. See our Ed post for how late days and extensions work for the group deadlines.