HW8a: OpenDP

CS 208 Applied Privacy for Data Science, Spring 2022

Version 1.0: Due Fri, April 1, 5:00pm.

1. Use the OpenDP library (`pip install opendp`) to answer the following. [1] [2] [3]

   (a) Solve for the ideal noise scale of the Gaussian mechanism with an input sensitivity of 4 (measured in absolute distance) and privacy utilization ($\epsilon = 1, \delta = 1e-6$).

   (b) Find the radius of a 95% confidence interval for the noise addition in 1a.

   (c) A policy research organization wants to know the smallest sample size necessary to release an "accurate" ϵ=1 DP mean income. Determine the smallest dataset size such that, with 95% confidence, the DP release differs from the clipped dataset's mean by no more than 1000. Assume that neighboring datasets have a symmetric distance at most 2.[4] Also assume a clipping bound of 500,000. You may need to use trial and error to set the binary search bounds.

   (d) Find the largest clamping bound `b` for which clamping to `(-b, b)` admits an ϵ=1 DP sized bounded sum over integers.[5] Assume that two-sided geometric noise[6] with scale parameter of 100 is added, and neighboring datasets have a symmetric distance at most 2.

   (e) Briefly reflect on the usability of the library in a few sentences. Did you find it intuitive to solve for these missing unknowns? How could the documentation and/or interfaces be improved?

2. Consider a dataset consisting of 32-bit signed integers clamped to $[-2^{16}, 2^{16}]$. 32-bit integers can be used to represent any integer in the range $[-2^{31}, 2^{31} - 1]$. Give an example of a dataset and query that has unexpectedly high sensitivity in this practical setting, compared to a theoretical analysis. Propose a solution that preserves both utility and privacy.

   This question highlights how theoretical proofs oftentimes diverge from real implementations that use finite data types, and why it is recommended to use existing libraries that handle these concerns when making differentially private releases on sensitive datasets.

[1] Documentation: `https://docs.opendp.org/en/stable/user/application-structure.html`
[2] Aggregators: `https://docs.opendp.org/en/stable/user/transformation-constructors.html#aggregators`
[3] Example notebook: `https://github.com/opendp/cs208/blob/main/spring2022/examples/wk8_opendp.ipynb`
[4] Symmetric distances are the most broadly applicable, and are the default dataset metric throughout the library.
[5] Binary search utilities in the library default to floats. Set integer bounds to search over integers.
[6] Also known as discrete laplace noise. Favored over laplace when noising integers, to avoid float complications.