Week 3 Exercises (ECE 598 DA)

Exercise (Understanding DP Definition). Let A be a mechanism that simply outputs the *entire dataset* $\mathbf{x}$ (an identity function with no randomness). Argue why A is *not* differentially private for any reasonable ε if the dataset has more than one possible value. Then contrast this with a trivial mechanism that outputs nothing (or just random noise independent of $\mathbf{x}$) and show that one *is* differentially private (with $\varepsilon = 0$). What does this say about the role of randomness and utility in DP?

Exercise (Global Sensitivity and Laplace Noise). A researcher wants to publish the *average income* of individuals in a database using $\varepsilon = 1$ differential privacy. Each individual's income x_i is in a known range $[0, \$100{,}000]$. The query function is $f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i$.

1. What is the global sensitivity $\Delta_1(f)$ of the average? *(Hint: consider two databases that differ in one person's income.)*

2. Describe the Laplace mechanism for releasing the average. What noise scale b (in dollars) should be used?

3. If $n = 1000$, roughly how large is the noise standard deviation? Would adding Laplace noise with that scale significantly distort the average for large n?

Exercise (Sequential vs Parallel Composition). A data analyst wants to publish two statistics about a dataset of 10,000 people: (A) the total number of individuals who have a certain disease, and (B) the total number of individuals who have a specific genetic marker. She uses the Laplace mechanism for each, with $\varepsilon_A = 0.5$ and $\varepsilon_B = 0.5$ (and $\delta = 0$ for both for simplicity). Consider two scenarios:

- **Scenario 1:** Both queries are on the *same population* of all 10,000 individuals.

- **Scenario 2:** Query A is asked on a subgroup of 5,000 individuals (cohort 1) and Query B on a *disjoint* subgroup of the other 5,000 individuals (cohort 2).

In each scenario, what is the overall privacy guarantee $(\varepsilon_{\text{overall}}, \delta_{\text{overall}})$ for releasing both A and B? Explain the difference.

Exercise (Advanced Composition Bound). Suppose a company wants to run $k = 100$ queries on a database with each query run under $(\varepsilon_0 = 0.1, \delta_0 = 10^{-6})$-DP. Using the basic composition theorem, the worst-case privacy after 100 queries would be $(100 \times 0.1, 100 \times 10^{-6}) = (10, 10^{-4})$-DP. Using the advanced composition theorem, we can achieve a tighter bound. **(a)** Compute ε_* for $k = 100$, $\varepsilon_0 = 0.1$, and choose $\delta' = 10^{-6}$ as an additional slack. Use the formula $\varepsilon_* = \sqrt{2k\ln(1/\delta')}\varepsilon_0 + k\varepsilon_0(e^{\varepsilon_0} - 1)$. **(b)** Compare ε_* with the basic bound of 10. **(c)** What is the overall δ in the advanced composition scenario?

Exercise (Moments Accountant / RDP Conceptual). You have a mechanism that at each query adds Gaussian noise with variance σ^2 (for simplicity, say each query is a counting query with $\Delta_2 = 1$). You run k such queries on the same data. Explain how you would use Rényi Differential Privacy to account for the overall privacy loss. Specifically: **(a)** If each query is $(\alpha, \bar{\varepsilon}_0)$-RDP, what is the RDP of k queries? **(b)** How do you convert the final RDP guarantee to an (ε, δ)? **(c)** Why might this approach yield a smaller ε than just using the basic (ε, δ) composition?