Privacy in Statistics and Machine Learning Spring 2021
In-class Exercises for Lecture 4 (Differential Privacy Foundations I)
February 4, 2021

Adam Smith and Jonathan Ullman

1. Let A be an ε-DP mechanism mapping $\mathcal{X}^n$ to the set $\mathcal{Y}$, let $E \subseteq \mathcal{Y}$ be an event, and let $\mathbf{x}, \mathbf{x}'$ be neighboring data sets.

   What is the shape of the region of possible pairs $(p, q) \in [0,1]^2$ such that $p = \mathbb{P}(A(\mathbf{x}) \in E)$ and $q = \mathbb{P}(A(\mathbf{x}') \in E)$? Can you describe it geometrically? As ε shrinks, does it get bigger or smaller? Are there points in $[0,1]^2$ that are not contained in this region for any finite $0 < \varepsilon < \infty$?

2. Prove or disprove: Let $A : \mathcal{X}^n \longrightarrow \mathcal{Y}$ be a *deterministic* algorithm where $\mathcal{X}$ has at least two different values. If A is ε-DP for some finite ε, then A ignores its input, that is, $A(\mathbf{x})$ is the same value regardless of $\mathbf{x}$.

3. Suppose we use the Laplace mechanism to estimate the number of individuals in a data set who reside in each of the 3,143 counties[1] in the US, using parameter $\varepsilon = 0.1$. What does Lemma 4.4 imply about the expected error of the count for Suffolk County, MA? What does it imply about the expectation of the largest error in the estimate of any county population?

4. Suppose we have a counting query $f(\mathbf{x}) = \sum_{i=1}^{n} \varphi(x_i)$ where $\varphi : \mathcal{X} \longrightarrow \{0,1\}$. The Laplace mechanism answers this query with noise parameter $1/\varepsilon$. Now consider the function $f^{(d)}(\mathbf{x})$ which outputs a vector of identical values

   $$f^{(d)}(\mathbf{x}) = \underbrace{(f(\mathbf{x}), f(\mathbf{x}), \ldots, f(\mathbf{x}))}_{d \text{ times}}.$$

   What is the global sensitivity of $f^{(d)}(\mathbf{x})$? Suppose you want to estimate $f(\mathbf{x})$ from the answer of the Laplace mechanism on query $f^{(d)}$. How would you estimate $f(\mathbf{x})$ and what would the variance of your estimate be? Does it increase, decrease, or stay roughly the same as d increases?

5. How does differential privacy interact with reconstruction attacks?

   Suppose A is an ε-differentially private algorithm that takes input $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \{0,1\}^n$. Consider an algorithm B that attempts to reconstruct the input from A's output: on input $A(\mathbf{x})$, it outputs a guess $\tilde{\mathbf{x}}$. Show that, for every algorithm B: if $\mathbf{x}$ is selected uniformly at random from $\{0,1\}^n$, and the algorithm B has access only to the output of A (nothing else), then

   $$\mathop{\mathbb{E}}_{\substack{\mathbf{x} \in_r \{0,1\}^n \\ \tilde{\mathbf{x}} = B(A(\mathbf{x}))}} (\#\,\text{errors}(\tilde{\mathbf{x}}, \mathbf{x})) \geq \frac{n}{e^{\varepsilon} + 1}$$

   Here, # errors(y, x) denotes the number of positions in which two vectors disagree (also called the Hamming distance). [2]

[1] This number includes county equivalents, and was drawn from the Wikipedia article “List of United States counties and county equivalents” in February 2021.

[2] In other words: when ε is small, differentially private algorithms do not allow for non-trivial reconstruction attacks. Even with no output at all, an attacker can always guess about $\frac{n}{2}$ of the bits of $\mathbf{x}$ in expectation (for example, by guessing the all-zeros string). The result above says that a attack based on differentially private output cannot do much better in expectation.

Hints: Use linearity of expectation. The number of errors can be written as a sum of randm variables E_i (for $i = 1$ to n), where E_i is 1 if $\tilde{\mathbf{x}}_i = x_i$ and 0 otherwise. What can you say about the conditional distribution of x_i given a particular output $A(\mathbf{x}) = a$? How big or small can $\Pr(x_i = 1|A(\mathbf{x}) = a)$ be? Given that, what is the largest possible probability that $E_i = 1$? What does that tell you about E_i's expected value? It might be helpful to think about what happens when A is the randomized response mechanism, though your final proof should apply to any ε-DP algorithm.