CS7880: Rigorous Approaches to Data Privacy, Spring 2017
POTW #1

Instructor: Jonathan Ullman

Due Sun, Jan 22th, 11:59pm
(Email to jullman+PrivacyS17@gmail.com)

- **You may work on this homework in pairs if you like. If you do, you must write your own solution and state who you worked with.**
- Solutions must be typed in LaTeX.
- Aim for clarity and brevity over low-level details.

Problem 1 (Random Subsampling). Given a dataset $x \in \mathcal{X}^n$, and $m \in \{0, 1, \ldots, n\}$, a *random m-subsample of x* is a new (random) dataset $x' \in \mathcal{X}^m$ formed by keeping a random subset of m rows from x and throwing out the remaining $n - m$ rows.

(a) Show that for every $n \in \mathbb{N}$, $|\mathcal{X}| \geq 2$, $m \in \{1, \ldots, n\}$, $\varepsilon > 0$, and $\delta < m/n$, the algorithm $A(x)$ that outputs a random m-subsample of $x \in \mathcal{X}^n$ is *not* (ε, δ)-differentially private.

(b) Although random subsamples do not ensure differential privacy on their own, a random subsample does have the effect of "amplifying" differential privacy. Let $A : \mathcal{X}^m \to \mathcal{R}$ be any algorithm. We define the algorithm $A'(x) : \mathcal{X}^n \to \mathcal{R}$ as follows: choose x' to be a random m-subsample of x, then output $A(x')$.

Prove that if A is (ε, δ)-differentially private, then A' is $(\frac{(e^\varepsilon - 1)m}{n}, \frac{\delta m}{n})$-differentially private. Thus, if we have an algorithm with the relatively weak guarantee of 1-differential privacy, we can get an algorithm with ε-differential privacy by using a random subsample of a dataset that is larger by a factor of $1/(e^\varepsilon - 1) = O(1/\varepsilon)$.

(c) **(Optional.)** We can also show that some sort of converse is true—for many tasks achieving (ε, δ)-differential privacy *requires* $\Omega(1/\varepsilon)$ more samples than achieving $(1, \delta)$-differential privacy. Let $\mathbf{q}(x) = (q_1(x), \ldots, q_k(x))$ be a collection of statistical queries.[1] Assume that there is *no* $(1, \delta)$-differentially private algorithm $A : \mathcal{X}^n \to \mathbb{R}^k$, such that

$$\forall x \in \mathcal{X}^n \quad \mathbb{E}\left[\|A(x) - \mathbf{q}(x)\|_\infty\right] \leq 1/100.$$

Show that for some $n' = \Omega(n/\varepsilon)$, there is *no* $(\varepsilon, \varepsilon\delta/100)$-differentially private algorithm $A : \mathcal{X}^{n'} \to \mathbb{R}^k$ such that

$$\forall x' \in \mathcal{X}^{n'} \quad \mathbb{E}\left[\|A(x') - \mathbf{q}(x')\|_\infty\right] \leq 1/100.$$

[1] Recall that a statistical query $q(x)$ takes a dataset $x = (x_1, x_2, \ldots) \in \mathcal{X}^*$ of arbitrary size, and outputs $\mathbb{E}_{x_i \sim x}[\phi(x_i)]$ for some function $\phi : \mathcal{X} \to [0, 1]$.