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

Instructor: Jonathan Ullman

Problem 1 (Random Subsampling)**.**

Given a dataset $x \in \mathcal{X}^n$, and $m \in \{0, 1, \dots, n\}$, 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 $n \in \mathbb{N}$, $|\mathcal{X}| \geq 2$, $m \in \{1, \dots, n\}$, $\varepsilon > 0$, 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 $\mathbf{q}(x) = (q_1(x), \dots, q_k(x))$ 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

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

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

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

Solution 1.

(a) Let $\mathcal{X} = \{0, 1\}$ and consider the two datasets $x = 0^n$ and $x' = 10^{n-1}$. Now define $S = \{z \in \{0, 1\}^m \mid z \neq 0^m\}$. Then for every ϵ and every $\delta < m/n$

$$e^\varepsilon \Pr[A(x) \in S] + \delta = \delta < \frac{m}{n} = \Pr[A(x') \in S],$$

contradicting (ε, δ)-dp of M.

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

(b) We'll use $T \subseteq \{1, \dots, n\}$ to denote the identities of the m-subsampled rows (i.e. their row number, not their actual contents). Note that T is a random variable, and that the randomness of A' includes both the randomness of the sample T and the random coins of A. Let $x \sim x'$ be adjacent databases and assume that x and x' differ only on some row t. Let x_T (or x'_T) be a subsample from x (or x') containing the rows in T. Let S be an arbitrary subset of the range of A'. For convenience, define $p = m/n$

To show $(p(e^\varepsilon - 1), p\delta)$-dp, we have to bound the ratio

$$\frac{\Pr[A'(x) \in S] - p\delta}{\Pr[A'(x') \in S]} = \frac{p\Pr[A(x_T) \in S \mid i \in T] + (1-p)\Pr[A(x_T) \in S \mid i \notin T] - p\delta}{p\Pr[A(x'_T) \in S \mid i \in T] + (1-p)\Pr[A(x'_T) \in S \mid i \notin T]}$$

by $e^{p(e^\varepsilon - 1)}$. For convenience, define the quantities

$$\begin{aligned}
C &= \Pr[A(x_T) \in S \mid i \in T] \\
C' &= \Pr[A(x'_T) \in S \mid i \in T] \\
D &= \Pr[A(x_T) \in S \mid i \notin T] = \Pr[A(x'_T) \in S \mid i \notin T]
\end{aligned}$$

We can rewrite the ratio as

$$\frac{\Pr[A'(x) \in S]}{\Pr[A'(x') \in S]} = \frac{pC + (1-p)D - p\delta}{pC' + (1-p)D}$$

Now we use the fact that, by (ε, δ)-dp, $A \le e^\varepsilon \min\{C', D\} + \delta$. The rest is a calculation:

$$\begin{aligned}
&pC + (1-p)D - p\delta \\
&\le p(e^\varepsilon \min\{C', D\} + \delta) + (1-p)D - p\delta \\
&\le p(\min\{C', D\} + (e^\varepsilon - 1)\min\{C', D\}) + \delta) + (1-p)D - p\delta \\
&\le p(\min\{C', D\} + (e^\varepsilon - 1)(pC' + (1-p)D) + \delta) + (1-p)D - p\delta \\
&\qquad\qquad\qquad\qquad \text{(Because } \min\{x, y\} \le \alpha x + (1-\alpha)y \text{ for every } 0 \le \alpha \le 1\text{)} \\
&\le p(C' + (e^\varepsilon - 1)(pC' + (1-p)D) + \delta) + (1-p)D - p\delta \qquad \text{(Because } \min\{x, y\} \le x\text{)} \\
&\le p(C' + (e^\varepsilon - 1)(pC' + (1-p)D)) + (1-p)D \\
&\le (pC' + (1-p)D) + (p(e^\varepsilon - 1))(pC' + (1-p)D) \\
&\le (1 + p(e^\varepsilon - 1))(pC' + (1-p)D) \\
&\le e^{p(e^\varepsilon - 1)}(pC' + (1-p)D)
\end{aligned}$$

So we've succeeded in bounding the necessary ratio of probabilities. Note, if you are willing to settle for $(O(\varepsilon m/n), O(\delta m/n))$-dp the calculation is much simpler. All this algebra is mostly just to get the tight bound.

(c) Assume for the sake of contradiction that there is an (ε, δ)-dp algorithm $A' : \mathcal{X}^{n'} \to \mathbb{R}^k$ such that

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

where $n' \approx n/\varepsilon$ will be chosen later. We will construct a $(1, e\delta/\varepsilon)$-dp algorithm $A : \mathcal{X}^n \to \mathbb{R}^k$ that satisfies

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

which violates the assumption.

Let $n = n'/m$ for $m = 1/\varepsilon$. We will simply assume that n'/m is an integer. Given a dataset $x \in \mathcal{X}^n$, we construct the dataset $x_{\otimes m} \in \mathcal{X}^{n'}$ by making m identical copies of each row of x. Now, two observations:

- If x, y are any two datasets in $\mathcal{X}^n$ that differ on at most one row, then the resulting datasets $x_{\otimes m}, y_{\otimes m}$ are datasets in $\mathcal{X}^{n'}$ that differ on at most m rows. Therefore, if we define the algorithm $A : \mathcal{X}^m \to \mathbb{R}^k$ to be $A(x) = A'(x_{\otimes m})$, then the resulting algorithm A satisfies (ε', δ')-differential privacy for

  $$\varepsilon' = m\varepsilon = 1 \qquad \delta' = me^{\varepsilon m}\delta = e\delta/\varepsilon$$

  by the "group privacy" property of differential privacy.
- Since statistical queries are linear, for every $\mathbf{q}$, we have $\mathbf{q}(x) = \mathbf{q}(x_{\otimes m})$. Therefore, by assumption

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

However, combining these two facts contradicts our assumption that no such $(1, e\delta/\varepsilon)$-differentially private algorithm $A : \mathcal{X}^n \to \mathbb{R}^k$ exists.