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 \((\varepsilon, \delta)\)-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 \( A : \mathcal{X}^m \rightarrow \mathcal{R} \) be any algorithm. We define the algorithm \( A' : \mathcal{X}^n \rightarrow \mathcal{R} \) as follows: choose \( x' \) to be a random \( m \)-subsample of \( x \), then output \( A(x') \).

Prove that if \( A \) is \((\varepsilon, \delta)\)-differentially private, then \( A' \) is \((\varepsilon/(e^\varepsilon - 1), \delta m/n)\)-differentially private. Thus, if we have an algorithm with the relatively weak guarantee of 1-differential privacy, we can get an algorithm with \( \varepsilon \)-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 \((\varepsilon, \delta)\)-differential privacy requires \( \Omega(1/\varepsilon) \) more samples than achieving \((1, \delta)\)-differential privacy. Let \( q(x) = (q_1(x), \ldots, q_k(x)) \) be a collection of statistical queries.\(^1\)

Assume that there is no \((1, \delta)\)-differentially private algorithm \( A : \mathcal{X}^n \rightarrow \mathbb{R}^k \), such that
\[
\forall x \in \mathcal{X}^n \quad \mathbb{E}\|A(x) - q(x)\|_\infty \leq 1/100.
\]

Show that for some \( n' = \Omega(n/\varepsilon) \), there is no \((\varepsilon, \varepsilon \delta/100)\)-differentially private algorithm \( A : \mathcal{X}^{n'} \rightarrow \mathbb{R}^k \) such that
\[
\forall x' \in \mathcal{X}^{n'} \quad \mathbb{E}\|A(x') - q(x')\|_\infty \leq 1/100.
\]

Solution 1.

(a) Let \( \mathcal{X} = \{0, 1\} \) and consider the two datasets \( x = 0^n \) and \( x' = 10^{n-1} \). Now define \( S = \{z \in \{0, 1\}^m | z \neq 0^n\} \). Then for every \( \varepsilon \) and every \( \delta < m/n \)
\[
e^\varepsilon \Pr[A(x) \in S] + \delta = \delta < \frac{m}{n} = \Pr[A(x') \in S],
\]
contradicting \((\varepsilon, \delta)\)-dp of \( M \).

\(^1\)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 \( \phi : \mathcal{X} \rightarrow [0, 1] \).
(b) We’ll use \( T \subseteq [1, \ldots, n] \) 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 \( (p(e^\epsilon - 1), p\delta) \)-dp, we have to bound the ratio

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

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

\[
C = \Pr[A(x_T) \in S | i \in T] \\
C' = \Pr[A(x'_T) \in S | i \in T] \\
D = \Pr[A(x_T) \in S | i \not\in T] = \Pr[A(x'_T) \in S | i \not\in T]
\]

We can rewrite the ratio as

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

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

\[
pC + (1 - p)D - p\delta \\
\leq p(e^\epsilon \min\{C', D\} + \delta) + (1 - p)D - p\delta \\
\leq p(\min\{C', D\} + (e^\epsilon - 1)\min\{C', D\} + \delta) + (1 - p)D - p\delta \\
\leq p(\min\{C', D\} + (e^\epsilon - 1)(pC' + (1 - p)D) + \delta) + (1 - p)D - p\delta \\
\leq (1 + p(e^\epsilon - 1))(pC' + (1 - p)D) \\
\leq e^{p(e^\epsilon - 1)}(pC' + (1 - p)D)
\]

So we’ve succeeded in bounding the necessary ratio of probabilities. Note, if you are willing to settle for \((O(\epsilon m/n), O(\delta m/n))\)-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 \((\epsilon, \delta)\)-dp algorithm \( A' : \mathcal{X}^n' \to \mathbb{R}^k \) such that

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

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

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

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 \( A : \mathcal{X}^m \rightarrow \mathbb{R}^k \) to be \( A(x) = A'(x_{\otimes m}) \), then the resulting algorithm \( A \) satisfies \((\varepsilon', \delta')\)-differential privacy for
  \[
  \varepsilon' = m\varepsilon = 1 \quad \delta' = me^{m}\delta = e\delta/\varepsilon
  \]
  by the “group privacy” property of differential privacy.

- Since statistical queries are linear, for every \( q, \) we have \( q(x) = q(x_{\otimes m}). \) Therefore, by assumption
  \[
  \forall x \in \mathcal{X}^n \quad E[\|A(x) - q(x)\|_\infty] \leq 1/100.
  \]

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