Mathematical Problems in Deep Learning: Homeworks #1 Solutions

1. **Probability Bound on Query Differences:** Consider a data $D = (d_1, d_2, ..., d_n)$ which is a n dimensional binary vector and another data $\tilde{D} = (\tilde{d}_1, \tilde{d}_2, ..., \tilde{d}_n)$. The difference between datasets is measured by

$$l(D, \tilde{D}) = \sum_{i=1}^{n} \mathbf{1}(d_i \neq \tilde{d}_i).$$

Define $q_{s_j}(D)$ as the response obtained by querying a subset of indices s_j. Let

$$B = \{\tilde{D} \mid l(D, \tilde{D}) \geq 256\alpha^2 n^2\}.$$

Prove that for $D_0 \in B$

$$\Pr(|q_{s_j}(D_0) - q_{s_j}(D)| \leq 4\alpha n) \leq \frac{9}{10}.$$

Solution:

Assume that two data $D = (d_1, \ldots, d_n)$ and $D_0 = (d_{0,1}, \ldots, d_{0,n})$ differ in at least

$$k \geq 256\,\alpha^2 n^2$$

positions. Let $F = (f_1, \ldots, f_n)$ be a n dimensional binary vector where $f_i = 1$ if and only if $i \in s_j$. Then,

$$\begin{aligned} X =& \left|q_{s_j}(D_0) - q_{s_j}(D)\right| \\ =& |F \cdot (D_0 - D)|. \end{aligned}$$

Note that $D_0 - D_1 \in \{-1, 0, 1\}^n$ is an n dimensional vector at least $k = 256\alpha^2 n^2$ of elements are nonzero. Then, for large enough n, the $F \cdot (D_0 - D)$ behaves like Gaussian with variance σ^2 at least $k/4 = 64\alpha^2 n^2$. This is because

$$\begin{aligned} \mathrm{Var}(F \cdot (D_0 - D)) =& \sum_{i=1}^{n} \mathrm{Var}(f_i(d_{0,i} - d_i)) \\ =& \sum_{d_{0,i} \neq d_i} \frac{1}{4} \\ \geq& k/4 = 64\alpha^2 n^2. \end{aligned}$$

Finally, for $Z \sim \mathcal{N}(0, (8\alpha n)^2)$

$$\begin{aligned} \Pr[|F \cdot (D_0 - D)| \leq 4\alpha n] \leq& \Pr[|Z| \leq 4\alpha n] \\ =& \Pr[|Z| \leq \sigma/2] \\ \leq& \frac{9}{10} \end{aligned}$$

2. **Differential Privacy Amplification via Subsampling:** Suppose we have an algorithm $M : X^m \to Y$ which is ϵ-DP. Consider the following algorithm $M' : X^n \to Y$, where $n > m$. When run on an input $X \in X^n$, it chooses a random subset of the input $X' \in X^m$ of size m, and outputs $M(X')$. Let $q = m/n$.

   Let $X, X' \in X^n$ be neighboring datasets. Without loss of generality, assume that:

   $$X = (x_1, x_2, \ldots, x_n), \quad X' = (x_1', x_2, \ldots, x_n).$$

   Define the sampling process: Let $I = (I_1, \ldots, I_m)$ be a random vector of size m, uniformly sampled (without replacement) from $\{1, \ldots, n\}$. Define the sampling function:

   $$s(X; I) = (x_{I_1}, \ldots, x_{I_m}).$$

   The new mechanism M' operates as: $M'(X) = M(s(X; I))$.

   Prove the following inequality:

   $$\Pr(M(s(X; I)) \in T \mid 1 \in I) \leq e^\epsilon \Pr(M(s(X; I)) \in T \mid 1 \notin I).$$

   Solution:
   For a pair of neighboring subsets $I_1, I_2 \subset \{1, \ldots, n\}$ of size m where $1 \in I_1$ and $1 \notin I_2$, we say $I_1 \sim_m^{(1)} I_2$. In other words, I_2 is replacing element '1' of I_1 with other index. Then, we have

   $$\Pr[M(s(X; I_1)) \in T] \leq e^\epsilon \Pr[M(s(X; I_2) \in T]. \tag{1}$$

   Consider all possible such pair of (I_1, I_2) and sum them up, then

   $$\sum_{I_1 \sim_m^{(1)} I_2} \Pr[M(s(X; I_1)) \in T] \leq e^\epsilon \sum_{I_1 \sim_m^{(1)} I_2} \Pr[M(s(X; I_2) \in T]. \tag{2}$$

   For each subset of size m set $1 \in I_1$, there are $n - m$ possible I_2's since the element '1' can be replaced by elements in I_1^c. For each subset of size m set $1 \notin I_2$, there are m possible I_1's since every element can be replaced by element '1'.

   $$\sum_{I_1 : |I_1| = m, 1 \in I_1} \Pr[M(s(X; I_1)) \in T] m \leq e^\epsilon \sum_{I_2 : |I_2| = m, 1 \notin I_2} \Pr[M(s(X; I_2) \in T](n - m). \tag{3}$$

   Thus, we have

   $$\Pr[M(s(X; I_1)) \in T, 1 \in I_1] m \leq e^\epsilon \Pr[M(s(X; I_2) \in T, 1 \notin I_2](n - m). \tag{4}$$

   This is equivalent to

   $$\Pr[M(s(X; I_1)) \in T | 1 \in I_1] \frac{1}{\binom{n-1}{m-1}} m \leq e^\epsilon \Pr[M(s(X; I_2) \in T | 1 \notin I_2] \frac{1}{\binom{n-1}{m}} (n - m), \tag{5}$$

   and finally we get

   $$\Pr[M(s(X; I_1)) \in T | 1 \in I_1] \leq e^\epsilon \Pr[M(s(X; I_2) \in T | 1 \notin I_2]. \tag{6}$$

3. **Sensitivity of K-Means Algorithm:** Consider a K-means clustering algorithm where d-dimensional data points $X_1, \ldots, X_n \in U$, where $U = \{x \in \mathbb{R}^d \mid \|x\|_1 \leq 1\}$. Given the centers $C_1, \ldots, C_k$, the mechanism f first group X_j's by

$$G_i = \{X_j : d(X_j, C_i) \leq d(X_j, C_m) \text{ for all } m \neq i\}$$

Then, the mechanism f outputs $f(C_1, \ldots, C_k) = (C'_1, \ldots, C'_k)$ where

$$C'_i = \frac{1}{|G_i|} \sum_{X_j \in G_i} X_j.$$

Determine the sensitivity of f.

Solution:
Let $x, x' \in \mathbb{R}^d$ be two data points such that $\|x\|_1 \leq 1$ and $\|x'\|_1 \leq 1$. Consider the worst-case scenario where the data set D and its neighboring data set D' differ in exactly one data point, and this data point belongs to a cluster that contains only that point.

In this case, the cluster mean in $f(D)$ is simply x, and the cluster mean in $f(D')$ is x'. Therefore, the change in output is:

$$\|f(D) - f(D')\|_1 = \|x - x'\|_1 \leq \|x\|_1 + \|x'\|_1 \leq 2.$$

Thus, the maximum possible change in the output (i.e., the global sensitivity) is 2, which occurs when a singleton cluster's data point is replaced by another point at L_1 distance 2.

Hence, the global sensitivity of f is $\boxed{2}$.

4. **Proof of Additive δ-DP:** We aim to show that M satisfies additive δ-Differential Privacy (δ-DP), which is defined as:

$$|\Pr(M(D) \in S) - \Pr(M(D') \in S)| \leq \delta,$$

for any neighboring datasets D and D' differing by at most one element and for any subset S of the output space.

   (a) Consider a mechanism $M(X)$ defined for a dataset $X = (X_1, \ldots, X_n)$ as follows:

$$M(X) = \begin{cases} X & \text{with probability } \delta, \\ \perp & \text{with probability } 1 - \delta. \end{cases}$$

   Show that M is additive δ-DP.

   (b) Consider a mechanism $M(x)$ defined for a dataset $X = (X_1, \ldots, X_n)$ as follows:

$$M(X) = (Y_1, \ldots, Y_n), \quad \text{where } Y_i = \begin{cases} X_i & \text{with probability } \delta, \\ \perp & \text{with probability } 1 - \delta. \end{cases}$$

   Show that M is additive δ-DP.

Solution:
(a) Consider two neighboring datasets X and X' that differ in exactly one element. Let T be any subset of the output space, and define the event E as "the mechanism outputs either X or X'". The complement event $\bar{E}$ corresponds to the mechanism outputting $\perp$, which occurs with probability $1-\delta$ regardless of the input.

Since the output is always $\perp$ under $\bar{E}$, the distributions of $M(X)$ and $M(X')$ are identical in this case. Therefore, we can write:

$$\Pr[M(X) \in T] = \Pr[M(X) \in T \mid \bar{E}] \cdot \Pr[\bar{E}] + \Pr[M(X) \in T \mid E] \cdot \Pr[E].$$

Because $\Pr[E] = \delta$, we have:

$$\Pr[M(X) \in T] = \Pr[M(X') \in T \mid \bar{E}] \cdot \Pr[\bar{E}] + \Pr[M(X) \in T \mid E] \cdot \delta.$$

Since $\Pr[M(X) \in T \mid E] \leq 1$, we get:

$$\Pr[M(X) \in T] \leq \Pr[M(X') \in T \mid \bar{E}] \cdot \Pr[\bar{E}] + \delta \leq \Pr[M(X') \in T] + \delta.$$

By symmetry, the same inequality holds with X and X' reversed. Hence:

$$|\Pr[M(X) \in T] - \Pr[M(X') \in T]| \leq \delta.$$

Therefore, the mechanism satisfies additive δ-differential privacy.

(b) Consider two neighboring datasets X and X', which differ only at the i-th position. Let T be any subset of the output space (i.e., a set of possible outputs), and define the event E as "the i-th coordinate is revealed (i.e., not replaced by $\perp$)". Then the complement $\bar{E}$ corresponds to the case where the i-th coordinate is replaced by $\perp$.

Conditioned on $\bar{E}$, the output distributions of $M(X)$ and $M(X')$ are identical, since all other coordinates are the same between X and X'. Therefore, we can write:

$$\Pr[M(X) \in T] = \Pr[M(X) \in T \mid \bar{E}] \cdot \Pr[\bar{E}] + \Pr[M(X) \in T \mid E] \cdot \Pr[E].$$

Note that $\Pr[E] = \delta$, and similarly, we get:

$$\Pr[M(X) \in T] \leq \Pr[M(X') \in T \mid \bar{E}] \cdot \Pr[\bar{E}] + 1 \cdot \delta.$$

But $\Pr[M(X') \in T \mid \bar{E}] \cdot \Pr[\bar{E}] = \Pr[M(X') \in T] - \Pr[M(X') \in T \mid E] \cdot \delta \leq \Pr[M(X') \in T]$. So we conclude:

$$\Pr[M(X) \in T] \leq \Pr[M(X') \in T] + \delta.$$

Thus, the mechanism satisfies additive δ-differential privacy.

5. **A different private algorithm:** Suppose that we wanted to answer a count query: $f(X) = \sum_{i=1}^{n} X_i$, where $X_i \in \{0, 1\}$. In class, we learned the Laplace mechanism: simply add Laplace noise with scale parameter $1/\epsilon$. But what if we do not have

access to Laplace noise? Suppose Z is a continuous uniform random variable, drawn uniformly from the interval $[-3/\epsilon, 3/\epsilon]$. Consider the statistic $\tilde{f}(X) = \sum_{i=1}^{n} X_i + Z$. Is $\tilde{f}$ $O(\epsilon)$-differentially private? If yes, prove it, with the best constant you can give in the privacy guarantee. If no, explain why not.

Solution:
No it is not, let $X = (0, \ldots, 0)$ and $X' = (1, 0, \ldots, 0)$. Then, the density around $1+3/\epsilon$ of $\tilde{f}(X)$ would be zero, while the density around $1 + 3/\epsilon$ of $\tilde{f}(X)$ is strictly positive.

6. **Mechanisms:** Consider the following mechanisms M that takes a dataset $x \in [0,1]^n$ and returns an estimate of the mean $\bar{x} = (\sum_{i=1}^{n} x_i)/n$.

(M1) $M_1(x) = [\bar{x} + Z]_0^1$, for $Z \sim \mathrm{Lap}(2/n)$, where for real numbers y and $a \leq b$, $[y]_a^b$ denotes the "clamping" function:

$$[y]_a^b = \begin{cases} a & \text{if } y < a \\ y & \text{if } a \leq y \leq b \\ b & \text{if } y > b \end{cases}.$$

(M2)

$$M_2(x) = \begin{cases} 1 & \text{w.p. } \overline{x} \\ 0 & \text{w.p. } 1 - \overline{x}. \end{cases}.$$

Which of the above mechanisms meet the definition of ϵ-differential privacy for a finite value of ϵ, and what is the smallest value of ϵ (possibly as a function of n) for which they do? As in class, here we are treating n as public knowledge (so it is not a privacy violation to reveal n), and working with the "change-one" definition of DP.

Solution:

(a) The maximum probability difference happens when $\bar{x} = 1$ and the minimum $\bar{x} = 1 - 1/n$. Thus, for M_1,

$$\begin{aligned} \frac{Pr(M_1(X) \in T)}{Pr(M_1(X') \in T)} &\leq \frac{f_X(1)}{f_X(1 - 1/n)} \\ &= e^{n/2(1/n)} = e^{1/2}. \end{aligned}$$

where f_X is a pdf of Laplace distribution with parameter $2/n$. Thus, the minimum ϵ for M_1 is $1/2$.

(b) Without loss of generality, it is enough to check the maximum ratio of probabilities at $M_2(X) = 1$.

$$\frac{Pr(M_1(X) = 1)}{Pr(M_1(X') = 1)} = \frac{\bar{X}}{\bar{X}'}.$$

This is unbounded since $\bar{X}'$ can be 0. Thus, it cannot be ϵ-DP for finite ϵ.

7. **Mean estimation with non-binary data:** In class, we saw how to estimate the mean of a dataset $\frac{1}{n}\sum_{i=1}^{n} X_i$ in the case when the X_i's are binary. Here, we will see how to estimate the mean of a dataset when this may not be the case.

   (a) Suppose we only knew the $X_i \in \mathbb{R}$ were real numbers. Prove that, for all $t \geq 0$, there is no ϵ-DP algorithm $M : \mathbb{R}^n \to \mathbb{R}$ such that

   $$Pr(|M(X) - f(X)| \leq t) \geq 0.9,$$

   where $\epsilon = 1$.

   (b) The previous problem showed that, in general, we can't privately estimate the mean of an unbounded dataset. Let's see how we can circumvent this issue. Give an algorithm $A_2 : \mathbb{R}^n \to \mathbb{R}$ with the following guarantees. The algorithm is ϵ-DP, for all possible datasets $(X_1, \ldots, X_n) \in \mathbb{R}^n$ If all $X_i \in [-R, R]$, then there exists some constant $C > 0$ such that

   $$Pr(|A_2(X) - f(X)| \leq \frac{CR}{n\epsilon}) \geq 0.9.$$

   The parameter R is known to the algorithm. Observe that this algorithm must always be private, but only needs to be accurate when the input dataset satisfies some additional properties

Solution:

(a) Without any bound, we can set $X = (0, \ldots, 0)$ and $X' = (2nt, 0, \ldots, 0)$. Note that $f(X) = 0$ and $f(X') = 2t$. First, due to ϵ-DP, we have

$$\frac{Pr(M(X) \in [-t, t])}{Pr(M(X') \in [-t, t])} \leq e^{\epsilon}$$

which implies $Pr(|M(X')| \leq t) \geq e^{-\epsilon} Pr(|M(X)| \leq t)$.
Also, the following should be true for both X and X',

$$\begin{aligned} Pr(|M(X)| \leq t) &\geq 0.9 \\ Pr(|M(X') - 2t| \leq t) &\geq 0.9 \end{aligned}$$

This cannot be true since

$$\begin{aligned} 0.1 &\geq Pr(|M(X') - 2t| > t) \\ &\geq Pr(|M(X')| \leq t) \\ &\geq e^{-\epsilon} Pr(|M(X')| \leq t) \\ &\geq e^{-\epsilon} 0. \end{aligned}$$

This is contradiction.

(b) The sensitivity of f is $2R/n$. Thus, if we apply Laplace mechanism with Laplacian parameter $2R/n\epsilon$, we can achieve ϵ-DP. The probability of error is

$$\begin{aligned}
Pr(|A_2(X) - f(X)| > CR/n\epsilon) =& Pr(|Y| > CR/n\epsilon) \\
=& e^{-C/2}.
\end{aligned}$$

where Y is Laplace random variable of parameter $2R/n\epsilon$ Thus, we can set $C = 2\log 10$ to get probaility of error equals to 0.1. In other words,

$$Pr(|A_2(X) - f(X)| \leq (2\log 10)R/n\epsilon) = 0.1$$

8. **Randomized Response, re-revisited:** We will see some generalizations of randomized response, beyond just binary alphabets. I will informally and vaguely describe an algorithm, you must rigorously define and specify the algorithm and prove that it is ϵ-differentially private.

   (a) Assume $X_i \in \{1, \ldots, k\}$ for the remainder of this problem. The vector $(Y_1, \ldots, Y_n) \in \{1, \ldots, k\}^n$ is output, where Y_i is equal to X_i with probability proportional to $g(\epsilon)$ (for some function g which you must specify), and equal to each $s \in \{1, ..., k\} \setminus X_i$ with probability proportional to 1. Specify the algorithm rigorously (define probability of randomized response) and prove that it is ϵ-DP.

   (b) Here is another way to generalize randomized response. The vector $(Y_1, \ldots, Y_n) \in \{0, 1\}^{kn}$ is output. $Y_i \in \{0, 1\}^k$ is a vector generated in the following manner: each X_i is first converted to a "one-hot" vector $\in \{0, 1\}^k$, where coordinate j is 1 if $j = X_i$ and 0 otherwise. Y_i generated from X_i by applying a bitwise randomized response (with appropriate parameter). Specify the algorithm rigorously (define probability of randomized response) and prove that it is ϵ-DP.

Solution:

(a) We can set $g(\epsilon) = e^\epsilon$, so that

$$\begin{aligned}
Pr(Y_i = X_i) =& \frac{e^\epsilon}{e^\epsilon + (k-1)} \\
Pr(Y_i = s) =& \frac{1}{e^\epsilon + (k-1)}
\end{aligned}$$

for $s \neq X_i$ The probability ratio is always bounded by e^ϵ.

(b) Without loss of generality, consider X and X' where $X_1 = (1, 0, \ldots, 0)$ and $X'_1 = (0, 1, 0, \ldots, 0)$ (and all others are the same $X_i = X'_i$).

$$\frac{Pr(Y_1 = (b_1, \ldots, b_k))}{Pr(Y'_1 = (b_1, \ldots, b_k))} = \frac{Pr(Y_{11}, Y_{12} = b_1, b_2)}{Pr(Y'_{11}, Y'_{12} = b_1, b_2)}.$$

When the bit-flipping probability is $q < 1/2$, the maximum ratio would be $(1-q)^2/q^2$ which should be bounded by ϵ. Thus, the flipping probability should be

$$q = \frac{e^{\epsilon/2}}{1+e^{\epsilon/2}}.$$

9. **Approximate DP:** Consider the following mechanisms M that takes a dataset $x \in [0,1]^n$ and returns an estimate of the mean $\bar{x} = (\sum_{i=1}^n x_i)/n$.

(M1) $M_1(x) = \bar{x} + [Z]_{-1}^1$, for $Z \sim \mathrm{Lap}(2/n)$.

(M2)

$$M_2(x) = \begin{cases} 1 & \text{w.p. } \overline{x} \\ 0 & \text{w.p. } 1-\overline{x}. \end{cases}.$$

The above mechanisms do not meet the definition of $(\epsilon, 0)$-differential privacy. For those mechanisms, calculate the smallest value of δ (again possibly as a function of n) for which they satisfy (ϵ, δ) differential privacy for a finite value of ϵ.

Solution.

(a) Suppose $\bar{x} = 1/n$ and $\bar{x'} = 0$. For any a, we have

$$Pr(a \leq 1/n + [Z]_{-1}^1 \leq 1 + 1/n) \leq e^\epsilon Pr(a \leq [Z]_{-1}^1 \leq 1 + 1/n) + \delta.$$

If $a > 1$, we have

$$Pr(a - 1/n \leq [Z]_{-1}^1 \leq 1) \leq \delta$$

which implies

$$\begin{aligned}\delta =& Pr(1 - 1/n \leq [Z]_{-1}^1 \leq 1) \\ =& Pr(1 - 1/n \leq Z) \\ =& \frac{1}{2} e^{-(n/2)(1-1/n)} \\ =& \frac{1}{2} e^{-n/2+1/2}.\end{aligned}$$

On the other hand, if $a < 1$, we have

$$\begin{aligned}& Pr(a - 1/n \leq [Z]_{-1}^1 \leq 1) \leq e^\epsilon Pr(a \leq [Z]_{-1}^1 \leq 1 + 1/n) + \delta \\ \Leftrightarrow & Pr(a - 1/n \leq [Z]_{-1}^1 \leq 1 - 1/n) \leq e^\epsilon Pr(a \leq [Z]_{-1}^1 \leq 1)\end{aligned}$$

which is true for $\epsilon = 1/2$. Thus, the above mechanism is $(1/2, 1/2e^{-n+1/2})$-DP. Note that you should specify ϵ as well since the mechanism is not (ϵ, δ)-DP for $\epsilon < 1/2$.

(b)

$$Pr(M_2(x) = 1) \le e^{\epsilon} Pr(M_2(x') = 1) + \delta \Leftrightarrow \bar{x} \le e^{\epsilon} \bar{x'} + \delta.$$

Thus, the δ should be

$$\delta = \max_{x,x'} \bar{x} - e^{\epsilon} \bar{x'} = 1/n.$$

Thus, the above mechanism is $(0, 1/n)$-DP. Note that you should clearly mention that the mechanism is (ϵ, δ)-DP for all ϵ.

10. **Regression:** Consider a dataset where each of its n rows is a pair of real numbers (x_i, y_i), each from an interval $[-b, b]$. Suppose we wish to find a best-fit linear relationship $y_i \approx \beta x_i$ between the y's and the x's. Non-privately, a standard way to estimate β is via the OLS regression formula

$$\hat{\beta} = \hat{\beta}(x, y) = \frac{S_{xy}}{S_{xx}} = \frac{\sum_i x_i y_i}{\sum_i x_i^2}.$$

This is called *ordinary least-squares (OLS)* regression, since $\hat{\beta}$ is the minimizer of the mean-squared residuals

$$\frac{1}{n} \sum_i (y_i - \hat{\beta} x_i)^2.$$

(a) Show that the function $\hat{\beta}(x, y)$ has infinite global sensitivity, and hence we cannot get a useful DP estimate of it via a direct application of the Laplace or Gaussian mechanisms.

(b) Show that S_{xy} and S_{xx} have global sensitivity that is bounded solely as a function of b, and hence each of these can be approximated in a DP manner using the Laplace mechanism.

(c) Using Part 10b with basic composition and post-processing, devise and implement an ϵ-DP algorithm for approximating $\hat{\beta}$ on an arbitrary dataset with $x_i, y_i \in \mathbb{R}$. In addition to the dataset $((x_1, y_1), \ldots, (x_n, y_n))$, your implementation should take as input parameters a clipping bound b and the privacy-loss parameter ϵ.

Solution.

(a) Let $X = \{(1, 0), \ldots, (0, 0)\}$ and $X' = \{(u, v), (0, 0), \ldots, (0, 0)\}$. Then, $\hat{\beta}(X) = 0$, and $\hat{\beta}(X') = v/u$. Since u can be arbitrarily close to 0 and the difference v/u is unbounded.

(b) Sensitivity of S_{xy} is $b \cdot b - b \cdot (-b) = 2b^2$ while sensitivity of S_{xx} is $b \cdot b - 0 \cdot 0 = b^2$.

(c) Based on composition theorem, you can mix two privatization with $\alpha\epsilon$ and $(1-\alpha)\epsilon$. In other words, compute $\tilde{S}_{xx}(X) = S_{xx}(X) + Z_{xx}$ where Z_{xx} be Laplace random variable with parameter $b^2/\alpha\epsilon$. Then, compute $\tilde{S}_{xy}(X) = S_{xy}(X) + Z_{xy}$ where Z_{xy} be a Laplace random variable with parameter $2b^2/(1-\alpha)$. Then, we compute $\tilde{\beta} = \tilde{S}_{xy}/\tilde{S}_{xx}$. The combining $\tilde{S}_{xx}$ and $\tilde{S}_{xy}$ is ϵ-DP due to composition theorem, and the final division does not affect the overall privacy due to post processing property.

Finally, you need to discuss the best α that minimizes the (any reasonable) loss between β and $\hat{\beta}$.