Exercise 08 - Solution

Differential Privacy

Reliable and Trustworthy Artificial Intelligence
ETH Zurich

Problem 1 (Singleton Sets)**.** Let $M\colon \mathcal{A} \to \mathcal{B}$ be a randomized mechanism with discrete outputs (i.e., the output set $\mathcal{B}$ is countable). Prove that in this case, the standard definition of ϵ-differential privacy, given by

$$\forall (a, a') \in \text{Neigh}.\ \forall S \subseteq \mathcal{B}. \qquad \Pr[M(a) \in S] \leq e^{\epsilon} \Pr[M(a') \in S] \tag{1}$$

is equivalent to:

$$\forall (a, a') \in \text{Neigh}.\ \forall b \in \mathcal{B}. \qquad \Pr[M(a) = b] \leq e^{\epsilon} \Pr[M(a') = b]. \tag{2}$$

Intuitively, this means that it is sufficient to only consider singleton attack sets $S := \{b\}$ when reasoning about differential privacy.

Solution 1.

First, note that Eq. (1) trivially implies Eq. (2), because if the inequality holds for *all* sets S, it also holds for all singleton sets $S := \{b\}$.

Second, we prove that Eq. (2) implies Eq. (1). Assume Eq. (2) holds. Then, for any neighboring a, a' and any discrete set $S \subseteq \mathcal{B}$, it is

$$\Pr[M(a) \in S] = \sum_{b \in S} \Pr[M(a) = b] \overset{(2)}{\leq} \Big(\sum_{b \in S} e^{\epsilon} \Pr[M(a') = b] \Big) = e^{\epsilon} \Pr[M(a') \in S],$$

which proves Eq. (1).

Problem 2 (Private Web Statistics)**.** A browser company wants to collect statistics from its m users as follows: for a given list of n websites, they want to determine how often, on average, a single user visits these websites during a specific time period.

In a federated setting, this can be achieved as follows: First, the browser of the i-th user locally collects a statistics vector $\mathbf{c}^i \in \mathbb{R}^n_{\geq 0}$, where c^i_j is the number of times this user visited the j-th website. Then, the company collects the vectors $\mathbf{c}^i$ for all users.

In order to protect the users' privacy, we want to hide the full statistics vector of an individual user. That is, we consider the DP neighborhood which allows exchanging, for a single user i, the vector $\mathbf{c}^i$ by an arbitrary vector in $\mathbb{R}^n_{\geq 0}$.

The following theorem may be useful to solve the subtasks below.

Theorem 1 (Parallel Composition)**.** *Assume the input database is partitioned into k subsets, where each subset contains the data of a distinct set of users. Further, let Neigh be a neighborhood which only allows changing the data in at most one of these user sets. Formally, let $\mathcal{A}$ be partitioned into $\mathcal{A}_1, \ldots, \mathcal{A}_k$ and let Neigh be such that*

$$\forall (a, a') \in \mathit{Neigh}.\ \exists i.\ \forall j \neq i.\ a_j = a'_j,$$

where a_j and a'_j represent the databases in the j-th partition $\mathcal{A}_j$.

Also, for $i \in \{1, \ldots, k\}$, let $M_i \colon \mathcal{A}_i \to \mathcal{B}_i$ be a (ϵ_i, δ_i)-DP mechanism. Then, their composition $M(a) := (M_1(a_1), \ldots, M_k(a_k))$ is $(\max_i \epsilon_i,\ \max_i \delta_i)$-DP.

1. To achieve DP, the company suggests introducing noise at the browsers. That is, the i-th user adds Laplace noise to $\mathbf{c}^i$ before sending $\mathbf{d}^i$ to the company:

   $$\mathbf{d}^i \leftarrow \mathbf{c}^i + (\mathrm{Lap}(0, \sigma), \ldots, \mathrm{Lap}(0, \sigma))$$

   Prove that irrespective of σ, this approach *cannot* satisfy ϵ-DP for *any* ϵ.

   Hint: The entries in $\mathbf{c}^i$ are unbounded.

2. To address the problem, the company suggests clipping the counts at an upper bound $b \in \mathbb{R}_{\geq 0}$ and instead compute

   $$\mathbf{d}^i \leftarrow \min(\mathbf{c}^i, b) + (\mathrm{Lap}(0, \sigma), \ldots, \mathrm{Lap}(0, \sigma)),$$

   where min is applied element-wise.

   How should the company select σ in order for the resulting collection of vectors $\mathbf{d}^1, \ldots, \mathbf{d}^m$ to be ϵ-DP for any ϵ? Provide a formula for σ and prove that the resulting mechanism is ϵ-DP.

3. The company computes the average visit counts as $\mathbf{a} = \frac{1}{m} \sum_{i=1}^m \mathbf{d}^i$.

   a) Assume $\mathbf{a}$ is published. Which level of DP does $\mathbf{a}$ achieve?

b) By the Central Limit Theorem, any component a_j converges to the true average $\frac{1}{m}\sum_{i=1}^{m} c_j^i$. Also, the variance of a_j can be approximated as $\frac{2\sigma^2}{m}$, which can be used to quantify the utility of the mechanism.

Using your choice of σ, how is the utility affected by the number of users m, the number of websites n, the bound b, and the level of privacy ϵ?

Note: The RAPPOR differential privacy mechanism [1], which is a sophisticated variant of the technique described here, has been reported to be used to collect statistics of Google Chrome users.

Solution 2.

1. Intuitively, this application of the Laplace mechanism does not work as $\mathbf{c}^i$ has unbounded sensitivity under the considered neighborhood.

   Formally, let σ and ϵ be arbitrary, and assume w.l.o.g. $n = 1$. Assume for the sake of contradiction that the mechanism satisfies ϵ-DP. Let c^i be arbitrary.

   By ϵ-DP, for p being the probability density function (PDF), it must be for all c'^i in the neighborhood of c^i:

   $$p(c^i + \mathrm{Lap}(0,\sigma) = 0) \leq e^\epsilon p(c'^i + \mathrm{Lap}(0,\sigma) = 0) \overset{\text{Lap PDF}}{\iff} \exp\left(\frac{-|c^i| + |c'^i|}{\sigma}\right) \leq e^\epsilon$$

   As this holds for any c^i, it must hold for $c^i = 0$, which implies

   $$\exp\left(\frac{|c'^i|}{\sigma}\right) \leq e^\epsilon \quad \iff \quad |c'^i| \leq \sigma\epsilon$$

   However, this inequality does not hold for any $c'^i > \sigma\epsilon$, even though such c'^i are in the neighborhood of $c^i = 0$. □

2. First, we compute the L1 sensitivity Δ_1 of $\min(\mathbf{c}^i, b)$. Any component may deviate by at most b within the neighborhood, because the components are bounded within the interval $[0, b]$. Hence, it is $\Delta_1 = \|(b, \ldots, b)\|_1 = \sum_{i=1}^{n} |b| = nb$.

   Now, we can select $\sigma = \Delta_1/\epsilon = nb/\epsilon$. By the theorem discussed in the lecture, the Laplace mechanism used to compute $\mathbf{d}^i$ is ϵ-DP for all i. Note that the considered neighborhood only changes the data of at most one user i. Hence, by Theorem 1 (parallel composition), the whole collection $\mathbf{d}^1, \ldots, \mathbf{d}^m$ is ϵ-DP.

   Note: The form of the neighborhood is critical here. If we allowed exchanging the data of multiple users, the collection $\mathbf{d}^1, \ldots, \mathbf{d}^m$ is not necessarily ϵ-DP any more.

By the regular composition theorem discussed in the lecture, the collection can be proven $m\epsilon$-DP in this case, which is however a very weak guarantee for large m.

3. a) As DP is maintained under post-processing, the result is still ϵ-DP.

   b) The variance for $\sigma = nb/\epsilon$ is $\frac{2\sigma^2}{m} = \frac{2n^2b^2}{\epsilon^2 m}$. As expected, more privacy (lower ϵ) increases the variance (lower utility). Similarly, more websites or a higher bound decrease the utility, because they require introducing more noise. In contrast, the more users contribute to the statistics, the better the utility.

Problem 3 (DP for Robustness)**.** Let f be a classifier and define the randomized classifier $\tilde{f}$ as $\tilde{f}(a) := f(a + \eta)$, where η is some random noise.

1. Assume $\tilde{f}$ is ϵ-DP for some $\epsilon \in \mathbb{R}$ and some symmetric neighborhood Neigh. Further, assume that $\tilde{f}$ satisfies the following separation condition for some a and c:

$$\forall c' \neq c. \quad \Pr[\tilde{f}(a) = c] > e^{2\epsilon} \Pr[\tilde{f}(a) = c'] \tag{3}$$

   Next, let the classifier g be constructed by applying randomized smoothing to f (i.e., $g(a) := \arg\max_j(\Pr[\tilde{f}(a) = j])$). Prove that g is robust to perturbations in the neighborhood Neigh. Formally, prove:

$$\forall a' \text{ s.t. } (a, a') \in \text{Neigh}. \quad g(a') = c$$

2. Assume $\tilde{f}$ is (ϵ, δ)-DP for some $\epsilon, \delta \in \mathbb{R}$ and some symmetric neighborhood Neigh. How does the separation condition in Eq. (3) need to be adapted such that we can again prove robustness, analogously as in the previous subtask?

   Hint: Extend Eq. (3) by an additive term.

Solution 3.

1. To demonstrate robustness, we need to show that for any a' in the neighborhood of a, it is $\Pr[\tilde{f}(a') = c] > \Pr[\tilde{f}(a') = c']$ for every class $c' \neq c$.

   Let a' be any input in the neighborhood and $c' \neq c$ be arbitrary. It is:

$$\begin{aligned}
\Pr[\tilde{f}(a') = c] &\geq e^{-\epsilon} \Pr[\tilde{f}(a) = c] && (\epsilon\text{-DP, symm. Neigh}) \\
&> e^{-\epsilon} e^{2\epsilon} \Pr[\tilde{f}(a) = c'] && (3) \\
&\geq e^{-\epsilon} e^{2\epsilon} e^{-\epsilon} \Pr[\tilde{f}(a') = c'] && (\epsilon\text{-DP, symm. Neigh}) \\
&= \Pr[\tilde{f}(a') = c']
\end{aligned}$$

□

2. We introduce an additive term Z (whose form is to be determined) in Eq. (3) and perform the following analogous derivation:

$$\begin{aligned}
\Pr[\tilde{f}(a') = c] &\geq e^{-\epsilon}(\Pr[\tilde{f}(a) = c] - \delta) && ((\epsilon, \delta)\text{-DP, symm. Neigh}) \\
&> e^{-\epsilon}(e^{2\epsilon} \Pr[\tilde{f}(a) = c'] + Z - \delta) && ((3) \text{ with } Z) \\
&= e^{\epsilon} \Pr[\tilde{f}(a) = c'] + e^{-\epsilon} Z - e^{-\epsilon}\delta \\
&\geq e^{\epsilon} e^{-\epsilon}(\Pr[\tilde{f}(a') = c'] - \delta) + e^{-\epsilon} Z - e^{-\epsilon}\delta && ((\epsilon, \delta)\text{-DP, symm. Neigh}) \\
&= \Pr[\tilde{f}(a') = c'] + e^{-\epsilon} Z - e^{-\epsilon}\delta - \delta
\end{aligned}$$

The last term is equal to the desired probability $\Pr[\tilde{f}(a') = c']$ if we set Z s.t.

$$\begin{aligned}
e^{-\epsilon} Z - e^{-\epsilon}\delta - \delta &= 0 \\
e^{-\epsilon} Z &= e^{-\epsilon}\delta + \delta \\
Z &= (e^{\epsilon} + 1)\delta
\end{aligned}$$

References

[1] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response". In: *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security - CCS '14*. 2014. DOI: 10.1145/2660267.2660348.