

Exercise 08

Differential Privacy

Reliable and Trustworthy Artificial Intelligence
ETH Zurich

Problem 1 (Singleton Sets). Let $M: \mathcal{A} \rightarrow \mathcal{B}$ 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

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

is equivalent to:

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

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

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 $\mathbf{c}^i \in \mathbb{R}_{\geq 0}^n$, where c_j^i 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 $\mathbb{R}_{\geq 0}^n$.

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 $\mathcal{A}_1, \dots, \mathcal{A}_k$ and let $Neigh$ be such that

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

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

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

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

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

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

where \min is applied element-wise.

How should the company select σ in order for the resulting collection of vectors $\mathbf{d}^1, \dots, \mathbf{d}^m$ 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.

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 :

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

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.

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.