A Statistical Threshold for Adversarial Classification in Laplace Mechanisms

Ayşe Ünsal, Melek Önen

A Statistical Threshold for Adversarial Classification in Laplace Mechanisms

Ayşe Ünsal and Melek Önen
Digital Security Department, EURECOM, France
Email: firstname.lastname@eurecom.fr

***Abstract*—This paper studies the statistical characterization of detecting an adversary who wants to harm some computation such as machine learning models or aggregation by altering the output of a differentially private mechanism in addition to discovering some information about the underlying dataset. An adversary who is able to modify the published information from a differentially private mechanism aims to maximize the possible damage to the system while remaining undetected. We present a trade-off between the privacy parameter of the system, the sensitivity and the attacker's advantage (the bias) through determining the threshold for the best critical region of the hypothesis testing problem for deciding whether or not the adversary's attack is detected. Such trade-offs are provided for Laplace mechanisms using one-sided and two-sided hypothesis tests. Corresponding error probabilities are analytically derived and ROC curves are presented for various levels of the sensitivity, the absolute mean of the attack and the privacy parameter. Subsequently, we provide an interval for the bias induced by the adversary so that the defender detects the attack. Finally, we adapt the *Kullback-Leibler differential privacy* to adversarial classification.**

A full version of this paper is accessible at: https://arxiv.org/abs/2105.05610.

I. Introduction

The widespread use of Big Data technologies has opened the door for malicious attacks resulting in potentially devastating consequences in critical applications such as autonomous driving or healthcare. In particular, an adversary may look for means to modify models or their outputs and consequently wreak havoc on a system and its users. Furthermore, such techniques usually rely on large datasets to be efficient which increases the chance of fraudulent use of personal information. *Adversarial classification* (also called anomaly detection) is a statistical tool enabling the detection of modification/misclassification attacks whereas privacy-preservation commonly makes use of so-called differential privacy (DP) mechanisms. A mechanism, such as a randomized function of a dataset, is said to be *differentially private* if the level of privacy of individual participants and the output of the mechanism remain unaltered even when any one of the participants decides to submit or equivalently remove his/her input from the statistical dataset. This definition is also applicable to aggregate information of all participants. This paper studies the security of systems where DP mechanisms are also used by adversaries.

DP [1] is defined as a stochastic measure of privacy -that has a precise mathematical formulation- to ensure privacy of individual users when handling large datasets. DP mechanisms have furthermore been used to develop practical methods for protecting private user-data at the moment they provide information to the system. In these cases, the use of DP measure aims to maintain the accuracy of the underlying operation without incurring a cost of the privacy of individual participants. In some sense, DP is a notion of robustness against changes in the dataset. The degree of this change is measured/determined by an adjustable privacy parameter and the amount of the change, that any single argument to the system reflects on its output, is called the sensitivity.

Statistical classification is now widely used as a supervised machine learning approach and consists in placing or *classifying* an item into one of several categories based on a number of measurements of interest. In [2], classification is described as a hypothesis testing problem for choosing between two possible values that the parameter(s) of a probability distribution can take on to place this item into the right category. Adversarial classification is an application of this approach where an adversary tries to fool a classifier which detects outliers in order to remain undetected.

In this paper, we consider a scenario where privacy enhancing technologies, which were originally designed to support privacy protection of legitimate individuals, are used by adversaries to harm the security of systems. We assume that the adversary is aware of the underlying DP mechanism and its parameters and wants to benefit from it using it as an attack tool to avoid being detected [3], [4]. The adversary's goal is to maximize the possible damage while minimizing the probability of being detected. We study the statistical framework of adversarial classification under DP. Our goal is to evaluate the impact of privacy parameters on the actual power of the adversarial classification. In particular, we focus on the aggregation operation whereby parties contribute with some individual numerical data and the system collecting this data computes the sum of them. We establish a stochastic relation between the probability of the adversary's success and the privacy parameter in the specific case of Laplace mechanisms.

The addressed problem in this work differs from existing work on DP which considers an adversary model where the goal of the attacker is to solely discover some information about the dataset. For instance, the implicit strong adversary assumption in [5] is that the adversary has the knowledge of the entire dataset except for one entry. In this paper, our aim is to extend this model with a stronger adversary who also wants to harm the dataset and the output of the mechanism. We consider an adversary who is able to modify the published information from a differentially private mechanism which is a noisy version of the output. The adversary's goal in this model is to maximize the possible damage while remaining undetected. Thus, there are two sides of what the adversary wants to achieve: (i) s/he gives false data by modifying the released information with the biggest possible difference from the real data, (ii) all this modification has to be achieved without being detected. On the defender's end, the mechanism wants to preserve DP and detect the attack.

Related work: A simpler version of this problem is addressed by [3] from an adversarial perspective and the conflicting goals of the adversary are formulated as an optimization problem where

the bias induced by the adversary is the objective function to be maximized. Yet, the privacy parameter does not take part in the formulation of [3]. We seek a characterization of the trade-off between the attack (the change in the output) and the privacy parameter. On the other hand, in [6], the authors show that the sensitivity of a mechanism has also an impact on the differentially private output. The noise to be added on the output is calibrated accordingly. Such a characterization of the problem described in this paper introduces a third element as the impact of the attack to be included in this adjustment of the DP noise with respect to (w.r.t.) the sensitivity of the system. This would allow us to be able to determine a threshold for detecting the attacker, alternatively, for the attacker to remain undetected.

Our methodology is the framework of statistical hypothesis testing in a similar vein to [7] where the authors determine an appropriate value of the privacy parameter as a function of error probabilities in deciding on the presence or absence of a particular record in a dataset. Similarly, in [8], the author studies the differentially private hypothesis testing in the local setting where users locally add the DP noise on their personal data before submitting it to the dataset. In this paper, we tailor this approach for the problem described above as a first attempt for a solution for anomaly detection in Laplace mechanisms under global DP where the personal sensitive data is transmitted to a central server by the users and the server applies DP noise on the data before its release to the public.

Contributions: We consider a new attacker model where the adversary takes advantage of the underlying differentially private mechanism in order not to be detected. For this model, we derive a trade-off between the privacy protected adversary's advantage and the security of the system for the adversary to remain undetected while giving as much damage as possible to the system. Alternatively, such a trade-off can be used for the defender to preserve the privacy of the system and detect the attacker. This trade-off is defined in the framework of statistical hypothesis testing similarly to [7]. Thus, we establish statistical thresholds for detecting the attack as a function of the error probabilities for Laplace mechanisms through one-sided and two-sided hypothesis tests. Subsequently, these thresholds are used for deriving intervals for the impact of the attack (or the privacy budget) to remain undetected as a function of the error probabilities and the sensitivity. Additionally, we adopt the Kullback-Leibler (KL) DP definition of [5] to the addressed problem for adversarial classification and present numerical comparisons of different scenarios where the sensitivity of the system is less than, equal to and greater than the bias induced by the adversary on the published information.

II. Preliminaries and Model

In this part, we revisit certain notions from the existing literature on DP which will also be used in this paper. These preliminaries will be followed by a detailed definition of the addressed problem.

Definition 1. *Any two datasets that differ only in one record are called neighbors [9]. For two neighboring datasets, the equality* $d(x, \tilde{x}) = 1$ *holds, where* $d(.,.)$ *denotes the Hamming (or* l_1*) distance between two datasets.*

Definition 2. *Global sensitivity* s *[6] for a function (or a query)* $f : D \to \mathbb{R}^k$ *is the smallest possible upper bound on the distance between the images of* f *when applied to two neighboring datasets, that is* $\|f(x) - f(\tilde{x})\|_1 \leq s$.

Sensitivity has an opposite relationship with the privacy. Higher sensitivity of the query refers to a stronger requirement for privacy guarantee, consequently more noise is needed to achieve that guarantee.

Definition 3. $(\epsilon, \delta)-$ *DP [9]: A randomized algorithm* $\mathcal{Y}$ *is* $(\epsilon, \delta)-$ *differentially private if* $\forall S \subseteq Range(\mathcal{Y})$ *and for all neighboring datasets* x *and* $\tilde{x}$ *within the domain of* $\mathcal{Y}$*, the following inequality holds:* $\Pr[\mathcal{Y}(x) \in S] \leq \Pr[\mathcal{Y}(\tilde{x}) \in S] \exp\{\epsilon\} + \delta$.

A differentially private system is named after the probability distribution of the perturbation applied onto the query output in the global setting. The Laplace distribution is defined as $Lap(x; \mu, b) = \frac{1}{2b} \exp\left\{-\frac{|x-\mu|}{b}\right\}$ with the location parameter equal to its mean μ and variance $2b^2$ where b denotes the scale parameter.

Definition 4. *Laplace mechanism [6] is defined for a function (or a query)* $f : D \to \mathbb{R}^k$ *as follows*

$$\mathcal{Y}(x, f(.), \epsilon) = f(x) + (Z_1, \cdots, Z_k) \tag{1}$$

where $Z_i \sim Lap(b = s/\epsilon)$*,* $i = 1, \cdots, k$ *denote i.i.d. Laplace random variables.*

A. Problem Definition and Performance Criteria

In this part, we provide a detailed description of the addressed problem and define the quantitative components for establishing a statistical threshold for detecting the attack. A differentially private mechanism adds Laplace noise denoted by Z on the query output $f(x) = \sum_{i=1}^{n} X_i$ using the dataset in the following form $\mathbf{X} = \{X_1, \cdots, X_n\}$. The noisy output is denoted by Y_0 and defined as $\mathcal{Y}(x, f(.), \epsilon) = Y_0 = \sum_{i=1}^{n} X_i + Z$. An adversary modifies this public information -which has been released by the server- by adding one extra record that we denote by X_a. Here the addition is applied onto the existing dataset without any constraint on the value of X_a, i.e. it could take up on a positive as well as a negative value. The modified output becomes $(\sum_{i=1}^{n} X_i + X_a) + Z$.

We define the following simple hypothesis testing problem in order to determine the threshold for deciding whether or not the adversary's attack is detected.

$$\begin{aligned} H_0 &: \text{defender does not detect } X_a \\ H_1 &: \text{defender detects } X_a \end{aligned} \tag{2}$$

The hypothesis testing problem in (2) can be translated into deciding on the DP noise distribution with its parameters. Here H_0 and H_1 correspond to DP noise following the probability distributions p_0 with mean μ_0 and p_1 with mean μ_1, respectively. Therefore, the decision boils down to choosing between $Y_0 - \sum_{i=1}^{n} X_i$ and $Y_0 - [\sum_{i=1}^{n} X_i + X_a]$. Hence the shift in the location due to the addition of X_a to the dataset is $\Delta\mu = \mu_1 - \mu_0$. The corresponding likelihood ratio for this problem yields

$$\Lambda = \frac{\mathcal{L}(p_1)}{\mathcal{L}(p_0)} \underset{H_1}{\overset{H_0}{\lessgtr}} \kappa \tag{3}$$

where $\mathcal{L}(.)$ denotes the likelihood function for the corresponding hypothesis and κ is some positive number to be determined. Such a threshold is used to define the critical region in statistical hypothesis tests (the region where the null hypothesis is rejected).

This paper presents a precise trade-off between the attacker's advantage (or the bias induced by the adversary) $\Delta\mu$, the sensitivity s and the privacy parameter ϵ for Laplace mechanisms to characterize the threshold for detecting the attack, as a function of the error probabilities. α and β respectively denote type I and type II error probabilities which are defined for the hypothesis testing problem in (2) as follows:

$$\alpha = \Pr\left[H_0 \text{ reject} | H_0 \text{ is true}\right] \quad (4)$$

$$\beta = \Pr\left[H_1 \text{reject} | H_1 \text{is true}\right]. \quad (5)$$

Based on the definition of α, also called the *probability of false-alarm*, we denote its complement by $\bar{\alpha} = 1-\alpha$. Similarly, due to (5), the complement of type II error probability (or the *probability of mis-detection*) is denoted by $\bar{\beta} = 1 - \beta$. The probability of correct detection $\bar{\beta}$ (i.e. correctly deciding H_1) is also called the *power of the test* in statistics or the *recall* in machine learning terminology.

According to the Neyman-Pearson Theorem [10], the likelihood ratio compared against some positive threshold defines the best critical region of size α for testing a simple hypothesis against an alternative simple hypothesis with the largest (or equally largest) power. An extension of this result to testing against a composite alternative hypothesis is also possible. Such an extension is called *uniformly most powerful test* since for a test with the best critical region of size α is conducted for each possible value of the alternative hypothesis. Once we define the critical region for deciding between H_0 and H_1 in (2) as a function of $\Delta\mu$, the privacy parameter ϵ and the sensitivity s, we will derive the error probabilities and the power of the test analytically as well as compute and depict them numerically.

III. Main Results

We separate our results in two main groups for $(\epsilon, 0)$-DP in Laplace mechanisms for one-sided and two-sided hypothesis tests.

A. One-sided test

In this part, we will investigate two cases setting the alternative hypothesis H_1 as either $\mu_1 > \mu_0$ ($\Delta\mu > 0$) or $\mu_1 < \mu_0$ ($\Delta\mu < 0$). This corresponds to a one-sided hypothesis testing problem. The decision of choosing between the hypotheses in (2) boils down to choosing between $Y_0 - \sum_{i=1}^{n} X_i = Z \sim \text{Lap}(z; \mu_0, s/\epsilon)$ and $Y_0 - [\sum_{i=1}^{n} X_i + X_a] = Z \sim \text{Lap}(z; \mu_1, \theta(s/\epsilon))$ where $\theta \geq 1$ is the measure of the change in the privacy budget of the system with sensitivity s and privacy parameter ϵ. It should be noted that setting $\theta = 1$ translates into testing only the location parameter of the Laplacian DP noise. Our goal is to derive a relationship between the privacy parameter, type I and type II error probabilities as a function of the bias $\Delta\mu$ for the attacker to be successful, that is to fail to reject H_0.

The corresponding likelihood ratio to (2) becomes $\Lambda = \frac{\mathcal{L}(p_1(\mu_1,b_1);z)}{\mathcal{L}(p_0(\mu_0,b_0);z)} \underset{H_1}{\overset{H_0}{\lessgtr}} \kappa$, where κ is some positive number to be determined and (μ_i, b_i) for $i = 0, 1$ respectively denote the location and scale parameters of the Laplace distributions. The next theorem states our first main result.

Theorem 1. *The threshold of the best critical region of size α defined in (4) for deciding between the null hypothesis and its alternative of the one-sided hypothesis testing problem in (2) for a Laplace mechanism with the largest possible power $\bar{\beta}$ is given as a function of the probability of false alarm α, privacy parameter ϵ and global sensitivity s by*

$$k = \begin{cases} \mu_0 + \frac{s}{\epsilon}\ln(2(1-\alpha)) & \text{if } \alpha \in [0, .5] \\ \mu_0 - \frac{s}{\epsilon}\ln(2\alpha) & \text{if } \alpha \in [.5, 1] \end{cases} \quad (6)$$

Then according to the adversary's hypothesis testing problem, the defender detects the attack for $\Delta\mu > 0$ if the output of the Laplace mechanism Y_0 exceeds $(k + f(x))$ where $f(.)$ is the noiseless query output. Similarly, for $\Delta\mu < 0$, the attack is detected if $Y_0 < f(x) + k$.

Proof. We expand the likelihood ratio Λ as follows.

$$\Lambda = \frac{\frac{\epsilon}{2\theta s}\exp\left\{-\epsilon\frac{|z-\mu_1|}{\theta s}\right\}}{\frac{\epsilon}{2s}\exp\left\{-\epsilon\frac{|z-\mu_0|}{s}\right\}} \quad (7)$$

The likelihood ratio in (7) can be summarized by the following piecewise function based on the ordering of μ_0, μ_1 and z for $\mu_1 < \mu_0$.

$$\Lambda_I = \begin{cases} \frac{1}{\theta}\exp\left\{\frac{\epsilon}{\theta s}\left(z(1-\theta) + \theta\mu_0 - \mu_1\right)\right\} & \text{if } z < \mu_1 \\ \frac{1}{\theta}\exp\left\{-\frac{\epsilon}{\theta s}(z(1+\theta) - \theta\mu_0 - \mu_1)\right\} & \text{if } z \in [\mu_1, \mu_0] \\ \frac{1}{\theta}\exp\left\{-\frac{\epsilon}{\theta s}(z(1-\theta) + \theta\mu_0 - \mu_1\right\} & \text{if } z \geq \mu_0 \end{cases} \quad (8)$$

On the other hand, for $\mu_1 > \mu_0$, the corresponding likelihood ratio yields

$$\Lambda_{II} = \begin{cases} \frac{1}{\theta}\exp\left\{\frac{\epsilon}{\theta s}(z(1-\theta) + \theta\mu_0 - \mu_1)\right\} & \text{if } z < \mu_0 \\ \frac{1}{\theta}\exp\left\{\frac{\epsilon}{\theta s}(z(1+\theta) - \theta\mu_0 - \mu_1)\right\} & \text{if } z \in [\mu_0, \mu_1] \\ \frac{1}{\theta}\exp\left\{-\frac{\epsilon}{\theta s}(z(1-\theta) + \theta\mu_0 - \mu_1)\right\} & \text{if } z \geq \mu_1 \end{cases} \quad (9)$$

To be able to determine a threshold for deciding on either of the hypotheses in (2), we compute the false alarm rate α and the mis-detection error β applying the Neyman-Pearson lemma that guarantees maximizing the power of the hypothesis test for a given α.

a) Derivation of α: Using the definition in (4) and for $\Delta\mu > 0$, the probability of raising a false-alarm is derived with the following integration $\alpha = \int_k^\infty \frac{\epsilon}{2s}\exp\left\{-\frac{\epsilon|z-\mu_0|}{s}\right\} dz$, which is further expanded out in two possible ways.

$$\alpha = 1 - \frac{1}{2}\exp\left\{\frac{\epsilon}{s}(k-\mu_0)\right\}, \text{ for } k < \mu_0 \quad (10)$$

$$\alpha = \frac{1}{2}\exp\left\{-\frac{\epsilon}{s}(k-\mu_0)\right\}, \text{ for } k \geq \mu_0 \quad (11)$$

Rewriting (10) and (11) as an equality for k, we obtain the piecewise function (6) as a function of α. If the bias induced by the adversary is negative, i.e. $\Delta\mu < 0$, then the conditions to obtain (10) and (11) are swapped. For $\Delta\mu < 0$ and $k < \mu_0$, we obtain (11) as the probability of false-alarm.

b) How to determine κ?: According to the piecewise expansion of likelihood ratio function in (8) for $\Delta\mu > 0$, κ is confined in: $(1/\theta\exp\left\{\frac{\epsilon}{\theta s}(z(1-\theta) + \theta\mu_0 - \mu_1)\right\}$, $(1/\theta)\exp\left\{-\frac{\epsilon}{\theta s}(z(1-\theta) + \theta\mu_0 - \mu_1)\right\})$. Since $\Lambda \underset{H_1}{\overset{H_0}{\lessgtr}} \kappa$, H_0 is rejected for $\frac{1}{\theta}\exp\left\{\frac{\epsilon}{\theta s}(z(1+\theta) - \theta\mu_0 - \mu_1)\right\} > \kappa$. Due to the threshold k of the critical region defined in Theorem 1, we get $\kappa = \frac{1}{\theta}\exp\left\{\frac{\epsilon}{\theta s}(k(1+\theta) - \theta\mu_0 - \mu_1)\right\}$ for $\Delta\mu > 0$. By analogy, κ becomes $\frac{1}{\theta}\exp\left\{-\frac{\epsilon}{\theta s}(k(1+\theta) - \theta\mu_0 - \mu_1)\right\}$, for negative bias. □

c) Derivation of the power of the test: The power of the hypothesis test is the probability of rejecting the null hypothesis H_0 given that the alternative hypothesis, H_1, is true. Using the definition in (5) for $\Delta\mu > 0$ and $k < \mu_1$, we get

$$\bar{\beta} = \int_k^{\infty} \frac{\epsilon}{2\theta s} \exp\left\{\frac{\epsilon(\mu_1 - z)}{\theta s}\right\} dz = \frac{1}{2} \exp\left\{\frac{\epsilon(\mu_1 - k)}{\theta s}\right\} \tag{12}$$

As for $k > \mu_1$, the power function becomes

$$\bar{\beta} = 1 - \int_{-\infty}^{k} \frac{\epsilon}{2\theta s} \exp\left\{\frac{\epsilon(z - \mu_1)}{\theta s}\right\} dz = 1 - \frac{1}{2} \exp\left\{\frac{\epsilon(k - \mu_1)}{\theta s}\right\} \tag{13}$$

On the contrary for negative bias $\Delta\mu < 0$, the conditions based on k and μ_1 to obtain (12) and (13) are swapped. In Section V, we present receiving operating characteristic (ROC) curves - the probability of false-alarm α versus power of the test $\bar{\beta}$- for Theorem 1 as performance analysis.

Remark. *Special case of $\theta = 1$ and $|\Delta\mu_1| \leq s$: Setting $\theta = 1$ and $|\Delta\mu_1| \leq s$ in the likelihood ratio (7), we get $\exp\{-\epsilon\} \leq \Lambda \leq \exp\{\epsilon\}$, which is the $(\epsilon, 0)-$ DP.*

B. Two-sided test

As an alternative solution to the same problem a two-sided test could provide a more realistic solution where it is not possible to know the direction of the shift. Hence the hypothesis test in (2) can be conducted for choosing between $H_0 : Z \sim \text{Lap}(\mu_0, s/\epsilon)$ and $H_1 : Z \sim \text{Lap}(\mu_1, \theta s/\epsilon)$. This translates to choosing between

$$H_0 : \mu = \mu_0, b = s/\epsilon \tag{14}$$

$$H_1 : \text{at least one of the equalities does not hold} \tag{15}$$

where μ denotes the location parameter and b denoted the scale parameter of any Laplace distribution. The alternative can also be stated with the parameters $\mu = \mu_1, b = \theta s/\epsilon$ where $\theta \geq 1$. In this two-tailed test, there are two thresholds on each side of the origin to be determined for the critical region, each with a size of $\alpha/2$. Let k_1 and k_2 denote the thresholds greater and smaller than the origin, respectively. The next theorem presents our second main result.

Theorem 2. *The threshold of the best critical region of size α defined in (4) for choosing between the null hypothesis and its alternative of the two-sided hypothesis testing problem in (14)-(15) for a Laplace mechanism with the largest power $\bar{\beta}$ is*

$$k_1 = \mu_0 - (s/\epsilon)\log\alpha \tag{16}$$

$$k_2 = \mu_0 + (s/\epsilon)\log\alpha \tag{17}$$

The defender fails to detect the attack when the output of the Laplace mechanism Y_0 is confined in $(f(x) + k_2, f(x) + k_1)$ where $f(.)$ denotes the noiseless query output.

Proof. The probability of raising a false-alarm or having a type I error is derived by $\alpha = \int_{-\infty}^{k_2} \frac{\epsilon}{2s} \exp\left\{\frac{\epsilon(z-\mu_0)}{s}\right\} dz + \int_{k_1}^{\infty} \frac{\epsilon}{2s} \exp\left\{-\frac{\epsilon(z-\mu_0)}{s}\right\}$. Each addend of α corresponds to one half of the probability of false-alarm. Equating each integral to $\alpha/2$ and rewriting the equalities in terms of k_1 and k_2, we get the thresholds (16) and (17). □

A trade-off between μ_1, s and ϵ for detecting the attacker: Using the thresholds of Theorem 2, we determine an interval to confine the mean of the attacker's advantage to be detected by the DP mechanism. Alternatively, such an interval can be converted for the privacy parameter ϵ as a function of error probabilities, the attack and the sensitivity. The following result, Corollary 2.1, presents upper and lower bounds on the attacker's advantage so that the defender detects the attack. For the proof of Corollary 2.1, the reader is referred to [11].

Corollary 2.1. *The absolute bias $|\Delta\mu| = |\mu_1 - \mu_0|$ induced by the adversary is confined in the following interval so that the defender detects X_a and preserves $(\epsilon, 0)$- DP*

$$\frac{s}{\epsilon}\log\left(\alpha\bar{\beta}^{\theta}\right) < \Delta\mu < \frac{s}{\epsilon}\log\left(\alpha\bar{\beta}^{\theta}\right)^{(-1)} \tag{18}$$

for $\theta \geq 1$ where α and $\bar{\beta}$ are the significance level and the power of the test of (15), respectively.

IV. Relative Entropy

This section is reserved for the derivation of relative entropy or KL divergence between two Laplace distributions and its adaptation to adversarial classification through *KL-DP*.

Definition 5 (KL-DP, [5])**.** *A randomized mechanism $P_{Y|X}$ guarantees $\epsilon-$ KL-DP, if the following inequality holds for all its neighboring datasets x and $\tilde{x}$, $D(P_{Y|X=x}||P_{Y|X=\tilde{x}}) \leq \exp\{\epsilon\}$.*

In [5, Theorem 1], KL-DP is proven to satisfy the following chain of inequalities $(\epsilon, 0) - \text{DP} \geq \text{KL} - \text{DP} \geq (\epsilon, \delta) - \text{DP}$. For the described problem and the associated model given in Section II-A, the neighboring datasets could be imagined as those where the output of the query is $\sum_{i=1}^{n} X_i$ before the attack and $(\sum_{i=1}^{n} X_i + X_a)$ after the attack. The corresponding distributions are considered as the DP noise with and without the induced value of X_a by the attacker as in our original hypothesis testing problem in (2). To be consistent with the hypotheses in (2), we set $P_{Y|X=x} \sim Lap(\mu_0, s/\epsilon)$ and for its neighbor, we have $Lap(\mu_1, \theta s/\epsilon)$.

The relative entropy between $p_0 \sim Lap(\mu_0, b_0)$ and $p_1 \sim Lap(\mu_1, b_1)$ is obtained respectively for $\Delta\mu > 0$ and $\Delta\mu < 0$ as

$$D(p_0||p_1) = \log\left(\frac{b_1}{b_0}\right) - 1 + \frac{b_0}{b_1}\exp\left\{\frac{\mu_0 - \mu_1}{b_0}\right\} - \frac{\mu_0 - \mu_1}{b_1}, \tag{19}$$

$$D(p_0||p_1) = \log\left(\frac{b_1}{b_0}\right) - 1 + \frac{b_0}{b_1}\exp\left\{\frac{\mu_1 - \mu_0}{b_0}\right\} + \frac{b_0}{b_1} \tag{20}$$

Due to space limitations, the derivation of KL-DP is omitted and can be found in [11]. The case of positive bias is numerically evaluated in Section V setting $b_0 = s/\epsilon$, $b_1 = \theta(s/\epsilon)$ and $\mu_1 - \mu_0 = \Delta\mu$ for the hypothesis testing problem defined in (2).

Remark. *Authors of [3] also seek for the maximum bias induced by the adversary where the objective function is the minimum relative entropy between the probability distribution of the dataset before (p_0) and after the attack (p_1). Nevertheless, the choice of the objective function is set as $D(p_1||p_0) \leq \gamma$ for some γ. For the Laplace distribution, KL divergence is not symmetric, hence $D(p_0||p_1) \neq D(p_1||p_0)$. Therefore, due to Stein's lemma [12], $D(p_0||p_1)$ in (19) (or (20)) should be used instead.*

V. Numerical Evaluation and Conclusion

Numerical Evaluation: KL-DP (19) derived in Section IV is numerically evaluated in Figure 1 for different levels of attack in comparison to the sensitivity of the system for both $\theta = 1$ and $\theta = 1.5$. Accordingly, the effect of the attack is compared with the upper bound $\exp\{\epsilon\}$ from Definition 5. Figure 1 shows that increasing the impact of the attack w.r.t. the sensitivity, closes the gap with the upper bound and for the case of $|\Delta\mu| = 4 * s$ and under moderate privacy budget, KL-DP upper bound is violated. Figure 2 presents the ROC curves corresponding to the

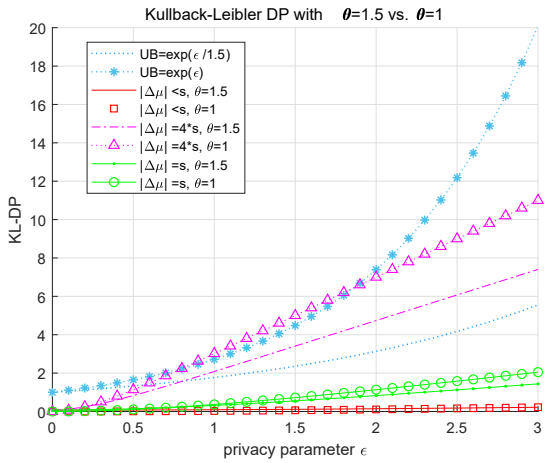

Figure 1. KL-DP for different values of ϵ and θ.

hypothesis test for the Laplace DP noise parameters. The plots depict different possible scenarios where the induced bias by the adversary is greater than, equal to and less than the sensitivity of the system. μ_0 is set to 0 hence $\Delta\mu = \mu_1$. We observe that when the privacy parameter ϵ is very small (e.g., $\epsilon = 0.015$), the test is no longer accurate and detecting the adversary can be considered similar to random guessing. On the other hand, when the privacy parameter is very large, the accuracy of the test becomes higher at the expense of the privacy guarantee. Furthermore, as opposed to [7, Theorem 5], we notice that ROC curves strongly depend on the sensitivity s, hence the mapping function (query) applied on the input. Particularly, when $\mu_1 > s$ the accuracy of the test becomes less important as the adversary is trying to harm the system. Figure 2 also shows that the choice of θ affects the power of the test. When $\theta = 1$, the test boils down to choosing between two location parameters. Figure 2 also shows that the power of the test on the y-axis decreases with θ. For each value of ϵ, ROC curves which correspond to $\theta = 1$ outperform those with a greater variance only after a certain level of α. As the privacy is decreased (equivalently ϵ is increased) this flip can be observed for smaller values of the probability of false alarm.

Conclusion: We characterized a statistical trade-off between the security of the Laplace DP mechanism and the privacy protected adversary's advantage in adversarial classification using one and two-tailed hypothesis testing. In both settings, we established trade-offs between the sensitivity of the system, privacy parameter and the induced bias by determining the threshold(s) of the critical region to decide whether or not the defender detects the attack. Such trade-offs are presented as functions of corresponding error probabilities. Numerical evaluation results show that increasing the privacy parameter also increases the

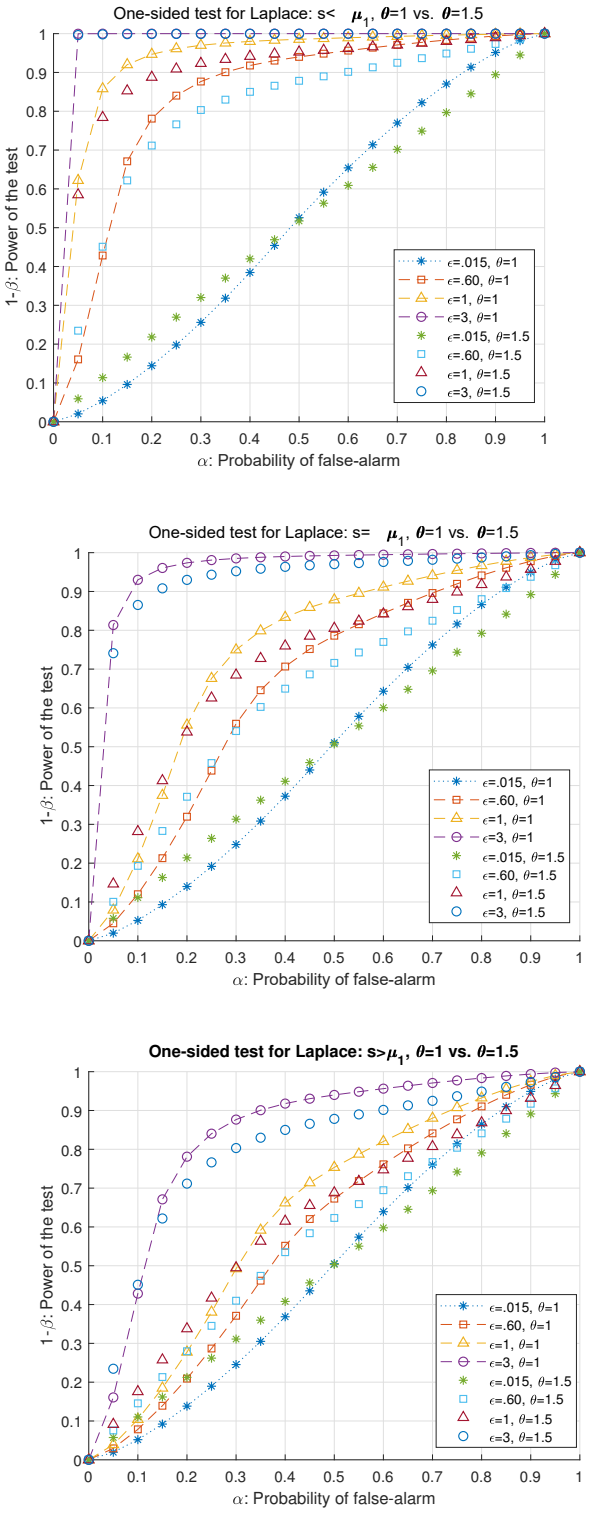

Figure 2. Eqs. (10)-(11) vs. (12)-(13) for different values of ϵ and θ.

accuracy of the hypothesis test. Additionally, we derived KL-DP for adversarial classification. Numerical evaluation shows that, the effect of increasing the impact of the attack closes the gap with the DP upper bound $\exp\{\epsilon\}$.

VI. Acknowledgements

This work has been supported by the 3IA Côte d'Azur project (reference number ANR-19-P3IA-0002).

REFERENCES

[1] C. Dwork, "Differential Privacy," in *Automata, Languages and Programming*, 2006, pp. 1–12.

[2] R. Hogg and A. Craig, *Introduction to Mathematical Statistics*. 4th Edition, Macmillan Publishing, New York, 1989.

[3] J. Giraldo, A. A. Cardenas, M. Kantarcioglu, and J. Katz, "Adversarial Classification Under Differential Privacy," in *NDSS 2020, Network and Distributed Systems Security Symposium, San Diego, CA, USA*, Feb. 2020.

[4] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana, "Certified robustness to adversarial examples with differential privacy," in *IEEE Symposium on Security and Privacy, San Francisco CA, USA*, May 2019, pp. 1054–1067.

[5] P. Cuff and Y. Langing, "Differential Privacy as a Mutual Information Constraint," in *CCS 2016, Vienna, Austria*, Oct. 2016.

[6] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis," in *Theory of Cryptography Conference*, 2006, pp. 265–284.

[7] C. Liu, X. He, T. Chanyaswad, S. Wang, and P. Mittal, "Investigating Statistical Privacy Frameworks from the Perspective of Hypothesis Testing," in *PETS 2019 Proceedings on Privacy Enhancing Technologies*, 2019, pp. 233–254.

[8] O. Sheffet, "Locally private hypothesis testing," in *Proceedings of Machine Learning Research*, 2018.

[9] C. Dwork and A. Roth, "The Algorithmic Foundations of Differential Privacy," *Foundations and Trends in Theoretical Computer Science 2014*, vol. 9, pp. 211–407, 2014.

[10] J. Neyman and E. Pearson, "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society A*, vol. 231, pp. 289–337, 1933.

[11] A. Ünsal and M. Önen, "A Statistical Threshold for Adversarial Classification in Laplace Mechanisms," May 2021. [Online]. Available: https://arxiv.org/abs/2105.05610

[12] T. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley Series in Telecommunications, 1991.