Chernoff Information as a Privacy Constraint for Adversarial Classification

Ayşe Ünsal
Digital Security Dept.,
EURECOM
Biot, France
ayse.unsal@eurecom.fr

Melek Önen
Digital Security Dept.,
EURECOM
Biot, France
melek.onen@eurecom.fr

***Abstract*—This work inspects a privacy metric based on Chernoff information, *Chernoff differential privacy*, due to its significance in characterization of the optimal classifier's performance. Adversarial classification, as any other classification problem is built around minimization of the (average or correct detection) probability of error in deciding on either of the classes in the case of binary classification. Unlike the classical hypothesis testing problem, where the false alarm and mis-detection probabilities are handled separately resulting in an asymmetric behavior of the best error exponent, in this work, we focus on the Bayesian setting and characterize the relationship between the best error exponent of the average error probability and ε-differential privacy [1]. Accordingly, we re-derive Chernoff differential privacy in terms of ε-differential privacy using the Radon-Nikodym derivative and show that it satisfies the composition property for sequential composition. Subsequently, we present numerical evaluation results, which demonstrates that Chernoff information outperforms Kullback-Leibler divergence as a function of the privacy parameter ε, the impact of the adversary's attack and global sensitivity for the problem of adversarial classification in Laplace mechanisms.**

***Index Terms*—Chernoff information, ε- differential privacy, Kullback-Leibler divergence, adversarial classification, composition**

I. Introduction

Classification theory covers the problem of optimally placing observations into different categories which are called the *classes*. Each class is defined due to an optimal rule according to some probabilistic description which may or may not be subjected to some unknown parameter(s). The major challenge here is to determine the optimal classifier's performance. This performance is commonly characterized in connection with error probabilities in deciding between one of the classes. In this paper, we employ the best average error exponent, namely *Chernoff information/divergence* as a data privacy metric in classifying adversarial examples targeting machine learning (ML) algorithms.

ML applications have gained significant traction, particularly over the past decade. In order to produce accurate results in an efficient manner, ML techniques heavily depend on large datasets, which jeopardizes privacy and security of innocent internet users who are contributing (knowingly or not) to online statistical datasets. The rising popularity of such applications that quickly found their place in our day-to-day lives in critical areas, consequently creates personal data privacy concerns while making data owners become prone to privacy breaches. Adversarial ML [2] studies privacy and security attacks and develop defense strategies to counter these attacks. Introducing adversarial examples to ML systems is a specific type of sophisticated and powerful attack, where additional -sometimes specially crafted- or modified inputs are provided to the system with the intent of being misclassified by the model as legitimate. In order to counter adversarial attacks, which aims to alter the data, as a possible defense strategy, *adversarial classification* targets to correctly detect such adversarial examples. Here the adversary's goal is to deceive the classifier, that is designed to detect outliers. In addition to the security angle of these type of misclassification attacks, user-data privacy is also subject to violations which creates an interplay between the security (adversary's aspect) and privacy (classifier/defender's aspect) in studying adversarial classification.

Differential privacy (DP), originally defined and studied in [3], is the mathematical foundation of user data privacy in statistical datasets. A randomized algorithm, also called a mechanism, guarantees the data to be analyzed without revealing personal information of any of the participants by employing DP. Essentially, a *differentially private* mechanism ensures the level of the privacy of its individual participants and the output of the analysis to remain unaltered, even when *any* user decides to leave or join the dataset with their personal information.

In this paper, we address the problem of adversarial classification under DP using a new definition based on *Chernoff information* [4]. Here, we consider a strong adversary who targets a differentially private mechanism not only to discover its information but also to alter it in order to benefit from it. The classification problem here is to choose between correctly detecting the modified data and failing to do so. The main contributions of the work can be summarized as follows.

- We study the so-called Chernoff DP as a function of Radon-Nikodym derivative which is related to Dwork's ε-DP. We present their comparison as a function of the privacy budget and show that ε-DP implies Chernoff DP for a range of ε values as a function of the prior probabilities assigned to each hypothesis.
- We show that the symmetry property of Chernoff infor-

mation and composition property of DP is preserved in this adaptation for both Kullback-Leibler and Chernoff DP.
- We present a numerical comparison among different variations of divergence based DP metrics.
- Ultimately, we numerically compare the performances of these two privacy metrics as a function of the impact of the attack against the global sensitivity of the query and the privacy budget.

Outline: Section II starts off with properties and different definitions of DP and continues with basic preliminaries on binary hypothesis testing problem. Section III offers divergence based DP definitions and their corresponding properties and existing comparisons. Our main result on the relation between Chernoff-DP and ε-DP is presented in Section IV. We present numerical evaluation of KL and Chernoff divergence functions for the problem of adversarial classification and finally, we draw conclusions of our findings and briefly discuss the future work in Section VI.

Related work: One could question the necessity of *yet* another information-theoretic DP metric, given the thorough studies on Kullback-Leibler [5], [6] and Rényi divergences [7], various definitions of mutual information [5], [8]–[10] and min-entropy [9], [11] based privacy metrics. Kullback-Leibler divergence and Chernoff information carry great importance in classification problems since they correspond to the best error exponent for mis-detection and average error probabilities, respectively. In [12], the authors study a similar problem to the current paper by employing the bias induced by the attacker as the objective function in a multi-criteria optimization problem subject to the Kullback-Leibler divergence between the probability distributions of the dataset before and after the attack. The optimization step is performed without taking the privacy budget into account. Kullback-Leibler divergence as well as other statistical divergence functions are used in many works for classification tasks [13]–[17]. To the best of our knowledge, Chernoff information has not been used for adversarial classification in the literature despite being thoroughly studied as in [18]–[20] and so on.

II. Preliminaries and Background

This part is reserved for providing some preliminaries and for building a background on some notions from measure theory, statistics and information theory that will be used throughout the paper.

A. Various properties and definitions of DP

In probability and measure theory, a function μ on a field $\mathcal{F}$ in a space Ω is called a measure given that the following conditions are met:

- $\mu(A) \in [0, \infty]$ for $A \in \mathcal{F}$;
- $\mu(\emptyset) = 0$;

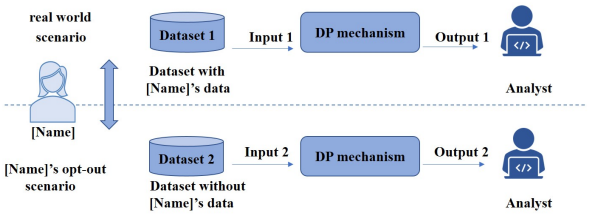

Fig. 1. Differential privacy

- if the sequence $\mathcal{F}$-sets $A_1, A_2, \cdots$ are a disjoint sequence of $\mathcal{F}$-sets where $\cup_{k=1}^{\infty} A_k \in \mathcal{F}$, then the following equality holds.

$$\mu\left(\cup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mu(A_k) \tag{1}$$

The pair $(\Omega, \mathcal{F})$ is called a *measurable space* if $\mathcal{F}$ is a σ-field in the sample space Ω [21]. If a measure P for $P(A)$ equals 0 whenever another measure $Q(A)$ equals 0, then P is said to be dominated by another measure Q, denoted as $P << Q$. For $P << Q$, the Radon-Nikodym derivative of P with respect to (w.r.t.) Q is denoted by $\frac{dP}{dQ}$ [22].

Definition 1. *[(ε, δ)-closeness [5]] Probability distributions P and Q defined over the same measurable space $(\Omega, \mathcal{F})$ are called (ε, δ)-close denoted by $P \overset{(\varepsilon,\delta)}{\approx} Q$ if the following couple of inequalities hold for any $A \in \mathcal{F}$.*

$$P(A) \leq e^{\varepsilon} Q(A) + \delta$$
$$Q(A) \leq e^{\varepsilon} P(A) + \delta$$

When $\delta = 0$, $(\varepsilon, 0)$- closeness between P and Q can be represented by the Radon-Nikodym derivative as follows:

$$e^{-\varepsilon} \leq \frac{dP}{dQ}(a) \leq e^{\varepsilon}, \ \forall a \in \Omega. \tag{2}$$

Equivalently, we have $\left|\log \frac{dP}{dQ}(a)\right| \leq \varepsilon$. In this paper, $\log(.)$ denotes the natural logarithm function.

Definition 2. *[23] Any two datasets $x, \tilde{x}$ that differ only in one record are called neighbors. 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 anticipates symmetry among neighbors in terms of the size of the datasets. This could be further relaxed to include the datasets of different sizes, where neighborhood is due to addition or removal of a record as depicted in Figure 1. Both definitions ensure that the neighbors differ in a single row (in one user's data).

Definition 3 ($(\varepsilon, \delta)-$ differential privacy)**.** *A randomized algorithm $\mathcal{M}$ is $(\varepsilon, \delta)-$ differentially private if $\forall S \subseteq Range(\mathcal{M})$ and $\forall x, \tilde{x}$ that are neighbors within the domain of $\mathcal{M}$, the following inequality holds.*

$$\Pr\left[\mathcal{M}(x) \in S\right] \leq \Pr\left[\mathcal{M}(\tilde{x}) \in S\right] e^{\varepsilon} + \delta. \tag{3}$$

The randomized mechanism $\mathcal{M}$ can also be represented by the conditional distribution of the dataset $X^n = (X_1, X_2, \cdots, X^n)$ with the corresponding output Y as $P_{Y|X^n}$. In this case, $P_{Y|X^n}$ satisfies $(\varepsilon, \delta)-$DP for all neighboring x^n and $\tilde{x}^n$ if the following holds:

$$P_{Y|X^n} \overset{(\varepsilon,\delta)}{\approx} P_{Y|\tilde{X}^n} \tag{4}$$

Although, we remind the reader of the original definition of $(\varepsilon, \delta)-$DP of [1] in Definition 3, we will stick to the expression based on closeness given by equation (4) and Definition 1 throughout the manuscript. Next, we remind the reader of the Kullback-Leibler DP (KL-DP).

Definition 4 (KL-DP, [5])**.** *A randomized mechanism $P_{Y|X}$ is said to guarantee ε- KL-DP, if the following inequality holds for all its neighboring datasets x and $\tilde{x}$,*

$$D(P_{Y|X^n}||P_{Y|\tilde{X}^n}) \leq \varepsilon. \tag{5}$$

Definition 5 (MI-DP [5])**.** *ε-mutual information differential privacy[1] (MI-DP) holds for the randomized mechanism $P_{Y|X^n}$ if the following inequality is satisfied*

$$\sup_{i, P_{X^n}} I(X_i; Y|X^{-i}) \leq \varepsilon \tag{6}$$

where X^{-i} denotes the sequence of n random variables excluding X_i.

B. Order of DP metrics

The main contribution of [5] is the comparison of $(\varepsilon, \delta)-$DP with KL-DP and MI-DP and their ordering in terms of their strength as a privacy metric. Namely, for two privacy metrics a-DP and b-DP with $a, b > 0$, a-DP is said to be stronger than b-DP, denoted by

$$a\text{-DP} \succeq b\text{-DP} \tag{7}$$

if for any positive b', there exists a positive a', such that

$$a'\text{-DP} \Longrightarrow b'\text{-DP}. \tag{8}$$

In other words, a-DP is stronger than b-DP since a-DP implies b-DP for any non-negative and non-zero b'. Accordingly, [5, Theorem 1] and its proof show that the following chain of inequalities hold

$$\varepsilon\text{-DP} \succeq \text{KL-DP} \succeq \text{MI-DP} \succeq (\varepsilon, \delta)\text{-DP} \tag{9}$$

where ε-DP and δ-DP are (ε, δ)-DP when δ and ε are zero, respectively. Here $\succeq$ denotes the privacy guarantee on its left hand side (l.h.s.) is *stronger than* the metric on its right hand side (r.h.s.), i.e. its existence implies the one on the r.h.s.

[1]The unit is set to be in nats instead of bits due to the use of natural logarithm.

C. Hypothesis Testing

As one of the most commonly used divergence definition, Kullback-Leibler distance [24], [25] between two probability measures P and Q with the dominating measure μ is defined as

$$D(P||Q) = \int p \log \frac{p}{q} d\mu \tag{10}$$

Chernoff information, also called Chernoff divergence is officially defined as follows.

Definition 6 (Chernoff Information [4])**.** *Chernoff information between any two probability measures P and Q with the dominating measure μ and the prior probability α is defined as follows:*

$$C(P, Q) = \max_{\alpha \in (0,1)} - \log \int p^{\alpha} q^{1-\alpha} d\mu. \tag{11}$$

(11) is equally written via the logarithm of α-skewed Bhattacharya coefficient $C_{\alpha}(P, Q)$ as $C(P, Q) = \max_{\alpha \in (0,1)} - \log C_{\alpha}(P, Q)$. An important property of Chernoff information, which is not captured by Kullback-Leibler divergence, is symmetry, i.e. $C(P, Q) = C(Q, P)$.

The importance of the divergences defined above by (10) and (11) is proven by Stein's Lemma [4], [26] in the settings of classical and Bayesian binary hypothesis testing. Accordingly, for two opposing hypothesis H_0 and H_1 that are set to choose between two probability distributions based on the observations, probabilities of false alarm P_{fa} and mis-detection P_{miss} are respectively defined by probabilities $P_{fa} = \Pr[\text{Choose } H_1|H_0 \text{ correct}]$ and $P_{miss} = \Pr[\text{Choose } H_0|H_1 \text{ correct}]$. In case of the existence of prior probabilities that are weights of the opposing hypothesis assigned by the analyst or existing due to prior analyses, the average error probability is the average of P_{fa} and P_{miss} and is equal to

$$P_e = \alpha P_{fa} + (1 - \alpha) P_{miss} \tag{12}$$

for $\alpha \in (0, 1)$. The optimal classifier obeys the following asymptotics for an $M-$dimensional random vector of observations.

$$\lim_{M \to \infty} \frac{P_{fa}}{M} = -D(Q||P), \text{ for fixed } P_{miss} \tag{13}$$

$$\lim_{M \to \infty} \frac{P_{miss}}{M} = -D(P||Q), \text{ for fixed } P_{fa} \tag{14}$$

$$\lim_{M \to \infty} \frac{P_e}{M} = -C(P, Q) \tag{15}$$

In addition to the relations between error exponents and divergence functions, Chernoff showed in [4] that $C(P, Q)$ can be used to obtain $D(P||Q)$ as

$$\left[\frac{dC_{\alpha}(P, Q)}{d\alpha}\right]_{\alpha=0} = D(Q||P) \tag{16}$$

$$\left[\frac{dC_{\alpha}(P, Q)}{d\alpha}\right]_{\alpha=1} = -D(P||Q) \tag{17}$$

where α denotes the prior probability. In this paper, we re-derive DP based Chernoff information tailored for the problem of adversarial classification.

III. Chernoff DP, KL-DP and Composability

We first restate Chernoff DP by the following definition.

Definition 7 (Chernoff-DP)**.** *A randomized mechanism $P_{Y|X}$ for input X and corresponding output Y is said to guarantee ε-Chernoff DP, if the following inequality holds for all its neighboring datasets x and $\tilde{x}$, $C(P_{Y|X^n}, P_{Y|\tilde{X}^n}) \leq \varepsilon$, where $\varepsilon > 0$.*

The goal of this paper is to discuss how Chernoff-DP fits into the chain of inequalities given by (9). In [27], we used a relaxed version of the above definition based on logarithm of α-skewed Bhattacharya affinity coefficient, since the maximum based on the choice of parameter α is upper bounded by ϵ, Chernoff information is upper bounded by ϵ for any value of α.

Remark 1. *A natural question can be asked regarding the choice of divergence function. Why do we need a new metric based on Chernoff information while we already have Rényi DP [7]? In particular case of adversarial classification as well as in general for any classification problem, Rényi divergence provides no information regarding the probability of correctly detecting the adversary's attack by altering the data. On the other hand, as reminded to the reader in Section II-C, due to Stein's lemma [4] Chernoff information corresponds to the highest achievable exponent for the average error probability of the optimal classifier's performance.*

One of the most important properties of DP is known as *composition* which has a role of preventing accumulated privacy leakage over several independent analyses. In other words, composition is how DP protects against an attacker who could combine several chunks of information from various sources. Composability property, in particular *sequential composition*, asserts that a set of mechanisms represented by different queries each individually satisfying DP, also collectively satisfies DP. Resulting privacy budget is proven to be proportional to the number of queries. Namely, sequential composition [28], [29], provides an upper bound on the privacy budget of releasing a series of query outputs of several DP mechanisms applied on the same data. Officially, sequential composition is defined by the following theorem.

Theorem 1. *For ε_j-differentially private m mechanisms $\mathcal{M}_1, \cdots, \mathcal{M}_m$ defined over $\mathcal{X}^n \to \mathcal{Y}$. A collection of m randomized algorithms $\mathcal{M}(x) = (\mathcal{M}_1(x), \cdots, \mathcal{M}_m(x))$ defined over $\mathcal{X}^n \to \mathcal{Y}^m$ and is run (over the same input data) independently, composability of DP ensures that $\mathcal{M}$ satisfies $\sum_j^m \varepsilon_j$-DP.*

The proof simply follows the multiplicative behavior of independent events composing a tuple even though the reference fails to mention the assumption of independence. The same logic applies to the lower bound with a negative exponent. Additivity property of KL divergence can directly be translated as proof of sequential composition property when the distance is used to define DP constraint as in Definition 4.

Corollary 1. *A set of conditionally independent query outputs $\{Y_j\}$ for $j = 1, \cdots, m$ given the dataset with each mechanism $P_{Y_j|X^n}$ satisfying ε_j-KL-DP also collectively satisfies KL-DP with $P_{Y_m|X^n}$ and privacy budget ε-KL-DP where $\sum_j^m \varepsilon_j = \varepsilon$.*

Proof. The collection of m conditionally independent mechanisms that satisfy ε_j-KL-DP is given by

$$D(P_{Y^m|X^n} || P_{Y^m|\tilde{X}^n}) = \sum_{j=1}^{m} D(P_{Y_j|X^n} || P_{Y_j|\tilde{X}^n}) \tag{18}$$

$$\leq \sum_{j}^{m} \varepsilon_j \tag{19}$$

which is equal to ε. (18) follows from the fact that the set of statistically independent query outputs given the database represent the mechanism $P_{Y^m|X^n}$ which is the product of the individual mechanisms for each j. Substituting the Definition 4 into (18) yields (19). The rest is a result of the following property of the logarithmic function in the definition of Kullback-Leibler divergence; $\log(a \times b) = \log a + \log b$. □

Similarly, MI-DP of Definition 5 is proven to satisfy composition theorem in [5] via employing the well-known chain rule of mutual information function.

Originally, Chernoff information of Definition 6 is not additive due to the optimization step unlike Kullback-Leibler distance or Bhattacharya distance, which is Chernoff information where $\alpha = 0.5$. Next, we state the sequential composability of Chernoff DP.

Corollary 2. *If the randomized algorithm $P_{Y_j|X^n}$ satisfies ε_j-Chernoff DP, then the set of conditionally independent randomized algorithms, $P_{Y_j|X^n}$ for $j = 1, 2, \cdots, m$, also satisfy Chernoff-DP with a privacy budget of $\sum_{j=1}^m \varepsilon_j$.*

Proof. By definition, Chernoff-DP is expanded out as follows:

$$\max_{\alpha \in (0,1)} -\log C_\alpha(P_{Y|X^n}, P_{Y|\tilde{X}^n}) \leq \varepsilon_k \tag{20}$$

The maximization on the l.h.s. of (20) can be removed without having any effect on the upper bound since it upper bounds the function for any value of α in its range.

$$-\log C_\alpha(P_{Y|X^n}, P_{Y|\tilde{X}^n}) \leq \varepsilon_k \tag{21}$$

Substituting conditionally independent randomized mechanisms $P_{Y^m|X^n}$ and $P_{Y^m|\tilde{X}^n}$ after removal of the maximization function, we have for the l.h.s.

$$-\log \int P^{\alpha}_{Y^m|X^n} P^{1-\alpha}_{Y^m|\tilde{X}^n} d\mu \tag{22}$$

$$= -\log \left(\prod_{j=1}^{m} \int P^{\alpha}_{Y_j|X^n} P^{1-\alpha}_{Y_j|\tilde{X}^n} d\mu \right) \tag{23}$$

$$\leq \sum_{j=1}^{m} \varepsilon_j \tag{24}$$

It is worth noting that, unlike ε-DP that is confined in the interval $[e^{-\varepsilon}, e^{\varepsilon}]$, Chernoff information, hence Chernoff-DP is lower bounded by 0. Additionally contrary to Chernoff-DP, Chernoff information defined by (11) is not additive. □

IV. Main Results-Comparison and Ordering of DP definitions

In this part, we re-derive Chernoff-DP via ε-DP in the form of $(\varepsilon, 0)$-closeness. It was shown in [5] and reminded by (9) that the KL-DP of Definition 4 is sandwiched between ε-DP and (ε, δ)-DP through redefining KL-DP as a function of Radon-Nikodym derivative $\frac{dP}{dQ}$. Accordingly, we have

$$D(P||Q) = \int dP(a) \log \frac{dP}{dQ}(a) \tag{25}$$

$$= \int \left[\frac{dP}{dQ}(a) \log \frac{dP}{dQ}(a) dQ(a) \right] \tag{26}$$

$$= \mathbb{E}\left[Z \log Z \right] \tag{27}$$

where (26) is obtained by setting $X \sim Q$ and in (27), the Radon-Nikodym derivative is set equal to some random variable $Z = \frac{dP}{dQ}(X)$ confined in the interval $Z \in [e^{-\varepsilon}, e^{\varepsilon}]$ following Definition 1. Namely, $(\varepsilon, \delta)-$DP ensures that the Radon-Nikodym derivative is defined over $[e^{-\varepsilon}, e^{\varepsilon}]$ when $\delta = 0$, Z is finally defined to be scattered around the endpoints $e^{-\varepsilon}$ and e^{ε} weighed by corresponding complementary probabilities to confer $\mathbb{E}[Z] = 1$. As a result, Z takes on the value e^{ε} with probability $p = \frac{1-e^{-\varepsilon}}{e^{\varepsilon}-e^{-\varepsilon}}$ and takes on $e^{-\varepsilon}$ with the complement of p when substituted into Kullback-Leibler divergence to redefine it as a function of the Radon-Nikodym derivative. KL-DP is translated into $\mathbb{E}\left[Z \log Z \right]$ and obtained as follows for P and Q satisfying ε-DP through deriving with this choice of probability distribution.

$$D(P||Q) \leq \varepsilon \frac{(e^{\varepsilon}-1)(1-e^{-\varepsilon})}{e^{\varepsilon}-e^{-\varepsilon}} \tag{28}$$

An interesting result of this relation appears in the form of symmetry between $D(P||Q)$ and $D(Q||P)$ when represented as a privacy metric.

Following similar steps, next, we plug the Radon-Nikodym derivative in Chernoff information in order to represent it in terms of $\varepsilon-$DP.

A. Chernoff DP re-written via ε-DP

α-skewed Bhattacharya affinity coefficient represented by the integral $C_\alpha(P,Q) = \int p^\alpha q^{1-\alpha} d\mu$ can be re-written w.r.t. P via a simple change of variables of integration. For $P << Q$, $C_\alpha(P,Q)$ becomes

$$\int p^\alpha q^{1-\alpha} d\mu = \int (p/q)^\alpha \, dQ \tag{29}$$

It is also possible to represent $C_\alpha(P,Q)$ based on Q to obtain the following form of α-skewed Bhattacharya affinity coefficient which yields the Chernoff information equivalent to the one given by (11).

$$\int p^\alpha q^{1-\alpha} d\mu = \int (q/p)^{1-\alpha} \, dP \tag{30}$$

It is straightforward to represent (29) as follows

$$C_\alpha(P,Q) = \int \left(\frac{dP}{dQ}(a) \right)^\alpha dQ(a) \tag{31}$$

$$= \mathbb{E}\left[Z^\alpha \right] \tag{32}$$

where in (32), we set $Z = \frac{dP}{dQ}(X)$ for $X \sim Q$. Similarly, plugging Radon-Nikodym derivative Z in $C_\alpha(P,Q)$ when the integral is based on P rather than Q, we obtain $\mathbb{E}_P\left[Z^{\alpha-1} \right]$. Using this form of $\alpha-$skewed Bhattacharya coefficient $C_\alpha(P,Q)$, finally we have

$$C_\alpha(P,Q) = \int \left(\frac{dQ}{dP}(a) \right)^{1-\alpha} dP(a) \tag{33}$$

$$= \mathbb{E}\left[Z^{\alpha-1} \right] \tag{34}$$

where $X \sim P$ in $Z = \frac{dP}{dQ}(X)$. Note that, (31) and (33) are equivalent. Next theorem defines the relation between Chernoff-DP and ε-DP via their representation as a function of Radon-Nikodym derivative dP/dQ.

Theorem 2 (Chernoff-DP and $\varepsilon-$DP relation)**.** *The following relation holds*

$$\varepsilon - \text{DP} \succeq \textit{Chernoff}\text{-DP} \tag{35}$$

where the optimal prior for $-\log C_\alpha(P,Q)$ *based on* $\mathbb{E}_Q[Z^\alpha]$ *is* $\alpha^* = \frac{1}{2\varepsilon} \log \frac{1+\varepsilon}{1-\varepsilon} - 1$. *By analogy, Chernoff-DP is maximized for* $-\log \mathbb{E}_P[Z^{\alpha-1}]$ *where the prior is* $\frac{1}{2\varepsilon} \log \frac{1+\varepsilon}{1-\varepsilon}$.

Note that, either expansion of $C_\alpha(.,.)$ leads ultimately to the same Chernoff information due to its symmetry property. Here the effect of different expansions are emphasized on the optimal value of parameter α. Theorem 2 can be interpreted as necessary condition for $\varepsilon-$DP to imply Chernoff-DP which is dependent on the relation between the prior probability α and the privacy parameter ε to hold.

Proof. To obtain Chernoff DP via ε-DP of (1), we need to re-write Chernoff DP as a function of Radon-Nikodym derivative Z. With the expectation in (32) over its range $[e^{-\varepsilon}, e^{\varepsilon}]$, $C_\alpha(P,Q)$ becomes

$$C_\alpha(P,Q) = \frac{1}{\alpha+1} \left(e^{\varepsilon(\alpha+1)} - e^{-\varepsilon(\alpha+1)} \right) \tag{36}$$

Substituting (36) into Chernoff DP, for the first expansion as a function of (32), we get

$$C(P,Q) = \max_\alpha \left\{ \log(\alpha+1) - \log\left(e^{\varepsilon(\alpha+1)} - e^{-\varepsilon(\alpha+1)} \right) \right\}, \tag{37}$$

or equivalently,

$$C(P,Q) = \max_\alpha \left\{ \log(\alpha+1) + \varepsilon(\alpha+1) - \log\left(e^{2\varepsilon(\alpha+1)} - 1 \right) \right\}. \tag{38}$$

Unfortunately, the optimization in (38) does not yield a closed-form solution for α as a function of the privacy parameter. To be able to characterize the optimal prior, we upper bound the Chernoff-DP in step (38) since the logarithmic function is monotonically increasing and the following inequality holds

$\log(x+1) \leq x$ for any positive x. Hence, the first term of (38) is upper bounded as

$$\log(\alpha+1) \leq \alpha. \tag{39}$$

Thus, we have for $C(P,Q)$, the following upper bound

$$C_{ub}(P,Q) = \max_{\alpha}\{\alpha+\varepsilon(\alpha+1)-\log\left(e^{2\varepsilon(\alpha+1)}-1\right)\} \tag{40}$$

where $C(P,Q) \leq C_{ub}(P,Q)$. In order to find the optimum value α, we take the derivative of $C_{ub}(P,Q)$ and seek for the value of α as a function of the privacy parameter ε that equals the derivative to zero as follows.

$$\frac{dC_{ub}(P,Q)}{d\alpha} = 1+\varepsilon-\frac{2\varepsilon e^{2\varepsilon(\alpha+1)}}{e^{2\varepsilon(\alpha+1)}-1} \tag{41}$$

We obtain the value of prior probability that maximizes the upper bound on Chernoff-DP as a function of the privacy parameter as $\alpha^*_{ub,I} = \frac{1}{2\varepsilon}\log\frac{1+\varepsilon}{1-\varepsilon} - 1$. Given that, α is the prior probability assigned to null hypothesis, $\alpha^*_{ub,I}$ should be confined in $(0,1]$. On the l.h.s. of $0 < \alpha^*_{ub,I} \leq 1$, we obtain $\varepsilon > 0$ whereas on the r.h.s., we obtain

$$\frac{1}{2\varepsilon}\log\left(\frac{1+\varepsilon}{1-\varepsilon}\right) \leq 2 \tag{42}$$

$$1-\frac{1-\varepsilon}{1+\varepsilon} \leq 4\varepsilon \tag{43}$$

where in (43), we lower bounded the l.h.s. of (42) by $\log x \geq 1-\frac{1}{x}$ where $x \in \mathbb{R}_{>0}$. This results in $\varepsilon \geq -.5$ which obeys the range we obtain from $\alpha^*_{ub,I} > 0$ and compatible with Definition 7.

As for the second expansion (33), where the expectation of the Radon-Nikodym derivative is based on P, we get

$$C_\alpha(P,Q) = \frac{1}{\alpha}\left(e^{\varepsilon\alpha}-e^{-\varepsilon\alpha}\right) \tag{44}$$

Plugging (44) into $C(P,Q) = \max_{\alpha}\{-\log C_\alpha(P,Q)\}$, we obtain the following expression, as in the first expansion with no explicit solution of maximization based on α.

$$C(P,Q) = \max_{\alpha}\left\{\log(\alpha)-\log\left(e^{\varepsilon\alpha}-e^{-\varepsilon\alpha}\right)\right\} \tag{45}$$

$$= \max_{\alpha}\left\{\log(\alpha)+\varepsilon\alpha-\log\left(e^{2\varepsilon\alpha}-1\right)\right\} \tag{46}$$

By upper bounding the first natural logarithm for any positive real number via $\log x \leq x-1$, we obtain $C_\alpha(P,Q) \leq \alpha - 1 + \varepsilon\alpha - \log\left(e^{2\varepsilon\alpha}-1\right)$. Plugging this expression into (46) and optimizing based on α, we have

$$\alpha^*_{ub,II} = \frac{1}{2\varepsilon}\log\frac{1+\varepsilon}{1-\varepsilon} \tag{47}$$

For the prior probability based on ε in (47) to be valid, $\alpha^*_{ub,II}$ must be confined in $(0,1]$. Accordingly, $\alpha^*_{ub,II} > 0$ is guaranteed by a positive privacy parameter, which is compatible with Definition 7. As for the upper bound on (47),

$$\frac{1}{2\varepsilon}\log\frac{1+\varepsilon}{1-\varepsilon} \leq 1 \tag{48}$$

$$1-\left(\frac{1+\varepsilon}{1-\varepsilon}\right) \leq 2\varepsilon \tag{49}$$

which again yields $\varepsilon > 0$. In (49), the l.h.s. is lower bounded via the following property of natural logarithm function $\log x \geq 1-\frac{1}{x}$ where $x \in \mathbb{R}_{>0}$. The symmetry in Chernoff information of Definition 11 is preserved in Chernoff-DP hence, substitutions of both priors into the upper bounds on two different expressions (38) and (46) result in

$$C(P,Q) \leq C^*_{ub}(P,Q) = \left(\frac{1}{2\varepsilon}+\frac{1}{2}\right)\log\frac{1+\varepsilon}{1-\varepsilon} - 1 + \log\left(\frac{2\varepsilon}{1-\varepsilon}\right) \tag{50}$$

where we denote the optimal upper bound on Chernoff-DP by $C^*_{ub}(P,Q)$. Note that, the symmetry property of Chernoff information is carried on the optimal upper bound on the DP metric due to the identical bounding step applied on both expansions. □

B. Numerical Comparison of DP Metrics

Through Definition 1 of $(\varepsilon,\delta)-$DP for $\delta = 0$, we have derived Chernoff information based privacy metric and proved its valid dependent on a relationship between optimal α and the privacy parameter ε. For the sake of demonstrating a numerical comparison between different upper bounds on $C(P,Q)$ through $\varepsilon-$DP in (38), the logarithmic term $\log\left(e^{2\varepsilon(\alpha+1)}-1\right)$ can be bounded by using $\log x \geq 1-1/x$ for $x = \log\left(e^{2\varepsilon(\alpha+1)}-1\right)$. This bound when plugged into $C(P,Q)$ upper bounds the optimal prior probability in terms of the privacy parameter by

$$\alpha^*_{alt} \leq \frac{1}{2\varepsilon}\log\left(\frac{4\varepsilon+2}{\varepsilon+1}\right) - 1 \tag{51}$$

where we denote the alternative bounds maximum point by α^*_{alt}. By analogy, for $C(P,Q)$ of (46), we get the optimal at $\alpha^*_{alt}+1$. In Figure 2, we present numerical evaluation results of two alternative upper bounds proposed on Chernoff-DP re-derived via $\varepsilon-$DP. Once again, the symmetry in the original definition is maintained in the upper bounds on the DP metric, so that the two different expansions of $C_\alpha(P,Q)$ yield upper bounds that ultimately coincide.

Figure 2 depicts the numerical comparison of various divergence based DP metrics. Accordingly, Chernoff-Cuff refers to (38) whereas UB I and UB II represent the two alternative upper bounds on $C(P,Q)$ obtained via α^*_{alt} and $\alpha^*_{ub,I}$, respectively. Additionally, $D(P||Q)$ of (28) is plotted via its product with $(1-\alpha)$ and appears as KL-Cuff in the legend of Figure 2. $D(P||Q)$ and $-D(Q||P)$ correspond to derivatives of $C(P,Q)$ for $\alpha = 0$ and $\alpha = 1$, respectively. The corresponding curves are plotted via the products $\alpha\cdot D(Q||P)$ and $(1-\alpha)\cdot D(P||Q)$. Lastly, the curve represented by $\alpha = 0.5$ in the legend is the Bhattacharya divergence which is $C(P,Q)$ for $\alpha = 0.5$.

Besides preservation of the symmetry property, Figure 2 also confirms that representation of Chernoff in terms of $\varepsilon-$DP, mirror the behavior of the curve as a function of the privacy parameter. Doubtlessly, α^*_{alt} leads to a significantly worse protection as opposed to α^*_{ub} which can be observed by comparing UB I and UB II.

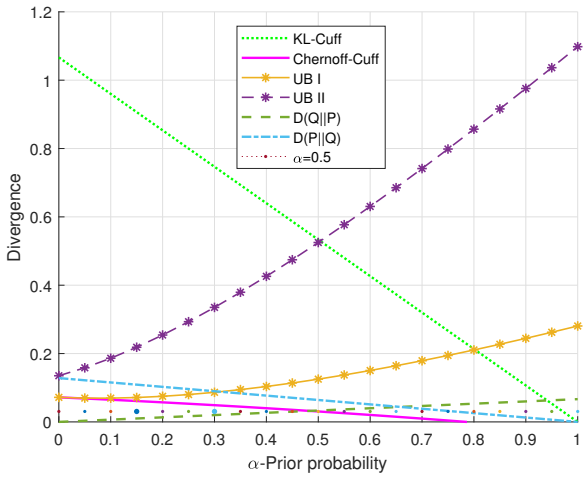

Fig. 2. Numerical comparison of different upper bounds on $C(P,Q)$ derived through $\varepsilon-$DP, KL-DP and Bhattcharya-DP which is $C(P,Q)$ with $\alpha = .5$.

V. Adversarial Classification with DP

As the discussion initiated in Section II-C, the motivation behind the interest in privacy metrics based on Chernoff and Kullback-Leibler divergences lies in best error exponents for binary classification. Imagine the following scenario, where a powerful adversary is able to benefit from privacy protection mechanism employed by the defender to avoid being correctly detected. Accordingly, the adversary is aware of the privacy protection and the main goal is to adjust the impact of its attack as a function of the privacy budget of the targeted mechanism so that the defender fails to correctly detect the attack [6]. Here the attack refers to the modification applied on the dataset by the adversary. In this part, we present a numerical comparison between KL-DP and Chernoff DP for the problem of adversarial classification in Laplace mechanisms. Laplace mechanism is reminded the reader by the following definition.

Definition 8. *Laplace mechanism [3] is defined for a function* $f : D \to \mathbb{R}^k$ *and i.i.d. Laplace random variables* $N_i \sim Lap(b = s/\varepsilon)$, $i = 1, \cdots, k$ *as follows*

$$\mathcal{Y}(x, f(.), \varepsilon) = f(x) + (N_1, \cdots, N_k) \tag{52}$$

where $\mathcal{Y}$ *represents the randomization mechanism and* s *is the global sensitivity of the function* f *that is* $\|f(x) - f(\tilde{x})\|_1 \leq s$.

An input dataset denoted by $X^n = \{X_1, \cdots, X_n\}$ is perturbed by Laplace noise $N \sim Lap(0, b)$ with the corresponding output $\mathcal{Y}(x, f(.), \varepsilon) = Y = f(X^n) + N$ where $f(.)$ denotes the query function. An adversary modifies this information by inserting or deleting one record where the modified data is denoted by X_a reflecting on the output that becomes $Y_a = f(X^n + X_a) + N$. Representing this problem with two hypotheses on the distribution of the Laplacian perturbation allows the defender to determine the threshold of correctly detecting the attacker as it also allows the adversary to fool the classifier and avoid being detected [27]. Accordingly, the hypotheses are set to decide whether the defender fails to detect the attack or correctly detects it.

In case of a linear $f(.)$, even if the adversary only has access to the published output, it is possible to determine the threshold of detection by using the likelihood ratio function. The performance criterion of such a test are different error probabilities, each of which is related either to Kullback-Leibler divergence or to Chernoff information. The main distinction between the two is the use and accessibility of prior probabilities for the opposing hypotheses, namely α and $1 - \alpha$. But ultimately, the effect of prior probabilities washes out and converges to zero [26] as a function of the size of the sequence of i.i.d. observations M.

For the problem described above, one can consider the notion of neighborhood as the datasets before and after the alteration applied by the adversary. In this case, the corresponding probability distributions to each hypothesis is the distribution of DP noise with and without the inserted record X_a considering a linear query function. Ultimately, in the classical approach, the probability distribution $N \sim Lap(\mu_0, b = s/\varepsilon)$ is tested against $Lap(\mu_1, \theta b)$ for $\theta > 1$ where in the Bayesian setting the null and alternative hypothesis are weighed by the corresponding prior probabilities α and $1-\alpha$, respectively. The difference in the mean denoted $\Delta\mu = \mu_1 - \mu_0$ is a result of the addition or deletion of X_a and $\mu_0 = 0$. Kullback-Leibler divergence between two Laplace distributions is derived in detail in [6] as

$$D(P||Q) = \log\theta - 1 + \frac{|\Delta\mu|}{\theta b} + \frac{1}{\theta} e^{-|\Delta\mu|/b} \tag{53}$$

As for the Chernoff information between two Laplace distributions, we have from [20], the following expression adapted to our problem

$$C(P,Q) = \frac{|\Delta\mu|}{\theta b} - \log\left(1 + \frac{|\Delta\mu|}{\theta b}\right). \tag{54}$$

In Figure 3, we present the numerical comparisons of (53) and (54), as a function of the privacy budget, the parameter θ that shows the change in the variance after attack and $\Delta\mu$, the shift in the mean due to the addition or (removal if negative) of X_a as a multiplier of sensitivity s. Accordingly, increasing $\Delta\mu$ with respect to the sensitivity, firstly closes the gap with the upper bound and for the extreme case of $\Delta\mu = 3 * s$ and $\theta > 1$, both Kullback-Leibler and Chernoff divergences surpass the upper bound ε. For all three scenarios, Chernoff is much tighter than Kullback-Leibler divergence as a function of privacy budget.

VI. Conclusions and Future Work

We studied the best average error exponent for adversarial classification, which is the well-known Chernoff information, in terms of the Radon-Nikodym derivative in relation with ε-DP. We showed that ε-DP implies Chernoff DP depending on the relation between the privacy parameter ε and the prior probability that conforms the classes associated with each hypothesis, which in our scenario, corresponds to whether or not the defender correctly detects the attack. The constraint

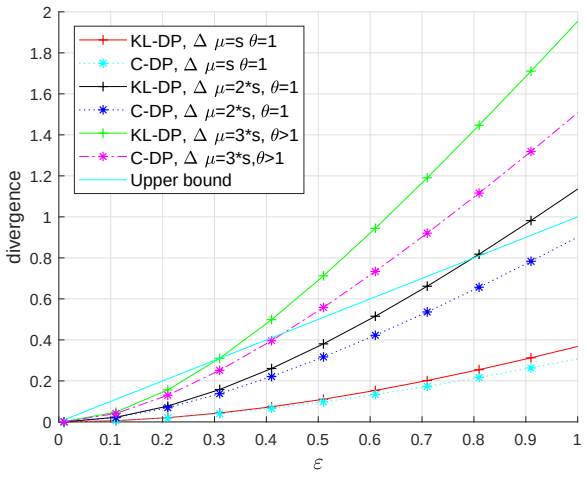

Fig. 3. Numerical comparison of KL-DP and Chernoff DP

on the privacy budget is introduced due to the prior probabilities. The optimal value to determine the best average error exponent for the prior probability and the corresponding ε could not be derived in a closed from, instead Radon-Nikodym derivative based Chernoff information is upper bounded prior to optimization. Future work will involve tighter bounding techniques.

Subsequently, we have demonstrated numerical comparison results for the well-known Kullback-Leibler divergence and Chernoff information for classification in Laplace mechanisms as a function of the privacy budget, global sensitivity and the absolute value of the modification applied on the data by the adversary. Accordingly, even when the absolute value of the impact of the attack is tripled in terms of the query's sensitivity, in the high privacy regime, i.e. for small values of ε, Chernoff information remains to obey the upper bound, unlike the Kullback Leibler divergence.

VII. Acknowledgement

The research leading to these results was funded by the 3IA Côte d'Azur Interdisciplinary Institute for Artificial Intelligence ANR-19-P3IA-0002.

References

[1] C. Dwork, "Differential privacy," in *Automata, Languages and Programming*. Berlin, Heidelberg: Springer, 2006, pp. 1–12.

[2] A. Joseph, B. Nelson, B. Rubinstein, and J. Tygar, *Adversarial Machine Learning*. Cambridge: Cambridge University Press, 2018.

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

[4] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Annals of Mathematical Statistics*, vol. 23, pp. 493–507, 1952.

[5] P. Cuff and L. Yu, "Differential Privacy as a Mutual Information Constraint," in *CCS 2016, Vienna, Austria*. New York, NY, United States: Association for Computing Machinery, Oct. 2016, pp. 43–54.

[6] A. Ünsal and M. Önen, "A statistical threshold for adversarial classification in laplace mechanisms," in *2021 IEEE Information Theory Workshop (ITW)*, 2021, pp. 1–6.

[7] I. Mironov, "Renyi differential privacy," 02 2017.

[8] W. Wang, L. Ying, and J. Zhang, "On the relation between identifiability, differential privacy and mutual information privacy," *IEEE Transactions on Information Theory*, vol. 62, pp. 5018–5029, Sep. 2016.

[9] G. Barthe and B. Köpf, "Information-theoretic bounds for differentially private mechanisms," in *Computer Security Foundations Symposium*. New York, NY, USA: IEEE, 2011, pp. 191–204.

[10] D. Mir, "Information theoretic foundations of differential privacy," in *International Symposium of Foundations on Practice of Security*. Berlin, Heidelberg: Springer, Oct. 2012, pp. 374–381.

[11] M. S. Alvim, M. E. Andrés, K. Chatzikokolakis, P. Degano, and C. Palamidessi, "Differential privacy: On the trade-off between utility and information leakage," in *Formal Aspects of Security and Trust*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 39–54.

[12] 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.

[13] D. Johnson, C. Gruner, K. Baggerly, and C. Seshagiri, "Information-theoretic analysis of neural coding," *Journal of Computational Neuroscience*, no. 10, pp. 47–69, 2001.

[14] N. Novello and A. M. Tonello, "f-divergence based classification: Beyond the use of cross-entropy," 2024.

[15] J. C. Duchi, K. Khosravi, and F. Ruan, "Information measures, experiments, multi-category hypothesis tests, and surrogate losses," *ArXiv*, vol. abs/1603.00126, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:13582051

[16] P. J. Moreno, P. P. Ho, and N. Vasconcelos, "A kullback-leibler divergence based kernel for svm classification in multimedia applications," in *Proceedings of the 16th International Conference on Neural Information Processing Systems*, ser. NIPS'03. Cambridge, MA, USA: MIT Press, 2003, p. 1385–1392.

[17] A. O. Hero, B. Ma, O. J. J. Michel, and J. D. Gorman, "Alpha-divergence for classification, indexing and retrieval (revised 2)," 2002. [Online]. Available: https://api.semanticscholar.org/CorpusID:12727488

[18] F. Nielsen, "An information-geometric characterization of chernoff information," *IEEE Signal Processing Letters*, vol. 20, Mar. 2013.

[19] F. Nielsen, "Revisiting chernoff information with likelihood ratio exponential families," *Entropy*, vol. 24, no. 10, 2022. [Online]. Available: https://www.mdpi.com/1099-4300/24/10/1400

[20] D. Johnson and S. Sinanovic, "Symmetrizing the kullback-leibler distance," 02 2003.

[21] P. Billingsley, *Probability and Measure*. New York: Wiley, 1995.

[22] O. Nikodym, "Sur une généralisation des intégrales de m. j. radon," *Fundamenta Mathematicae*, vol. 15, no. 1, pp. 131–179, 1930. [Online]. Available: http://eudml.org/doc/212339

[23] 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.

[24] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.

[25] S. Kullback and R. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, 1951.

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

[27] A. Ünsal and M. Önen, "Calibrating the attack to sensitivity in differentially private mechanisms," *Journal of Cybersecurity and Privacy*, vol. 2, no. 4, October 2022.

[28] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Advances in Cryptology - EUROCRYPT 2006*, S. Vaudenay, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 486–503.

[29] P. Kairouz, S. Oh, and P. Viswanath, "The composition theorem for differential privacy," in *32nd International Conference on Machine Learning*. JMLR, Inc. and Microtome Publishing (United States), 2015, pp. 4037–4049.