A Statistical Threshold for Adversarial Classification in Laplace Mechanisms

Ayşe ÜNSAL, Melek ÖNEN

Digital Security Dept., EURECOM

November 22, 2021
3IA Workshop, Toulouse

EURECOM
Sophia Antipolis

UNIVERSITÉ CÔTE D'AZUR

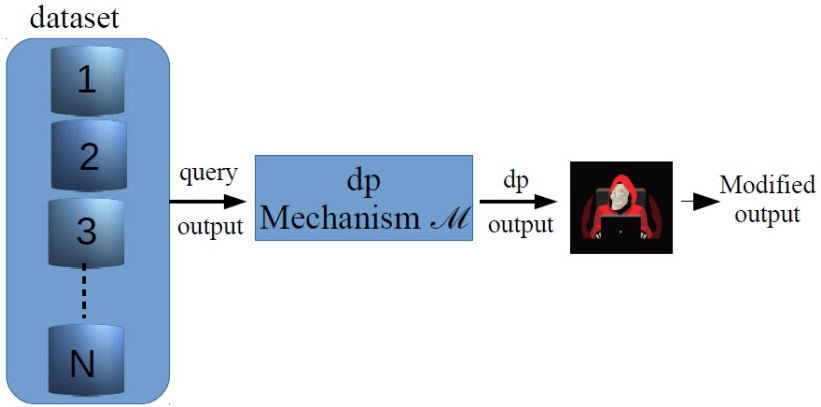
3iA Côte d'Azur
Interdisciplinary Institute for Artificial Intelligence

Outline

- Motivation
- Preliminaries
- Problem definition and performance criteria
- Threshold(s) to avoid being detected
- ROC curves
- Kullback-Leibler differential privacy
- References

Motivation

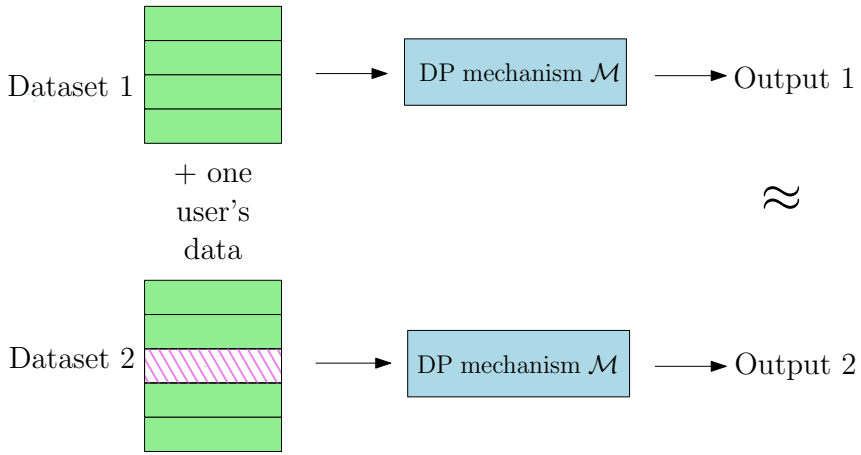
- Adversarial classification with differential privacy (DP)



- Adversary's conflicting goals:
  - s/he gives false data by modifying the released information, with the biggest possible difference from the real data
  - avoid being detected $\rightarrow$ adversary knows about the DP.
- On the defender's end, the mechanism wants to preserve DP and to detect adversarial examples

Preliminaries

- Differential privacy: the absence or presence of a single database item does not affect the outcome of the analysis



- $(\epsilon, \delta)-$ DP:

Definition (Dwork & Roth 2014)

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.

$$\Pr\left[\mathcal{Y}(x) \in S\right] \leq \Pr\left[\mathcal{Y}(\tilde{x}) \in S\right] \exp\{\epsilon\} + \delta$$

Preliminaries

- Laplace mechanism $\rightarrow (\epsilon, 0)-$ DP:

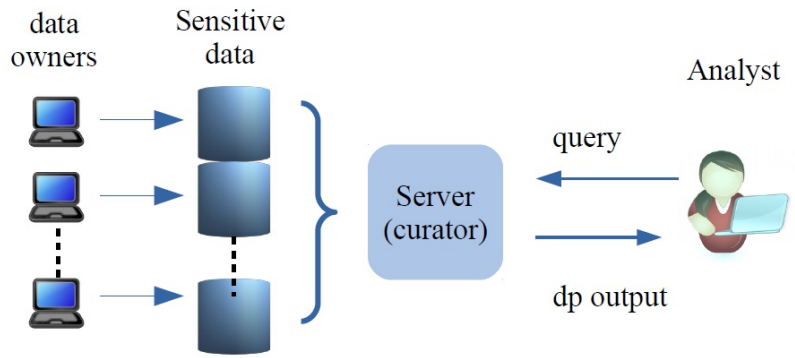
Definition (Dwork et al. 2006)

Laplace mechanism is defined for a function $f : D \rightarrow \mathbb{R}^k$ as follows

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

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

- DP- the global setting: the noise is added by a *trusted* central server who has access to raw data



Problem definition

- Query output is $f(x) = \sum_{i=1}^{n} X_i$ where the dataset is $\mathbf{X} = \{X_1, \cdots, X_n\}$
- The noisy output is defined by $Y_0 = \sum_{i=1}^{n} X_i + Z$ where $Z \sim Lap(s/\epsilon)$
- An adversary adds the record X_a to the dataset

- What is the statistical threshold for detecting the adversary's attack?

$$H_0 : \text{defender does not detect } X_a$$
$$H_1 : \text{defender detects } X_a$$

Performance criteria

- The probability of false-alarm (type I error)

$$P_{FA} = \alpha = \Pr\left[H_0 \text{ reject} | H_0 \text{ is true}\right]$$

- The power of the test (correct detection)

$$P_D = \bar{\beta} = \Pr\left[H_0 \text{ reject} | H_1 \text{is true}\right]$$

- The corresponding likelihood ratio for this problem yields

$$\Lambda = \frac{\mathcal{L}(\text{Lap}(\mu_1, b_1); z)}{\mathcal{L}(\text{Lap}(\mu_0, b_0); z)} \underset{H_1}{\overset{H_0}{\lessgtr}} \kappa$$

where κ is some positive number to be determined.

Threshold to avoid being detected-One sided

Theorem

The threshold of the best critical region of size α for deciding between H_0 and H_1 for a Laplace mechanism with the largest possible power $\bar{\beta}$ is given as a function of the probability of false alarm, privacy parameter ϵ and global sensitivity s as follows

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

Then, the adversary's hypothesis testing problem for $\mu_1 - \mu_0 > 0$ is $Y_0 \underset{H_1}{\overset{H_0}{\lessgtr}} k + f(x)$ *where $f(.)$ denotes the query function.*

By analogy for negative bias, we have $Y_0 \underset{H_0}{\overset{H_1}{\gtrless}} k + f(x)$.

Threshold(s) to avoid being detected-Two-sided

- Two sided test

$$
\begin{aligned}
&H_0 : \mu = \mu_0, b = s/\epsilon \\
&H_1 : \text{at least one of the equalities does not hold} \\
&\quad (\mu = \mu_1, b = (\theta s)/\epsilon)
\end{aligned} \tag{1}
$$

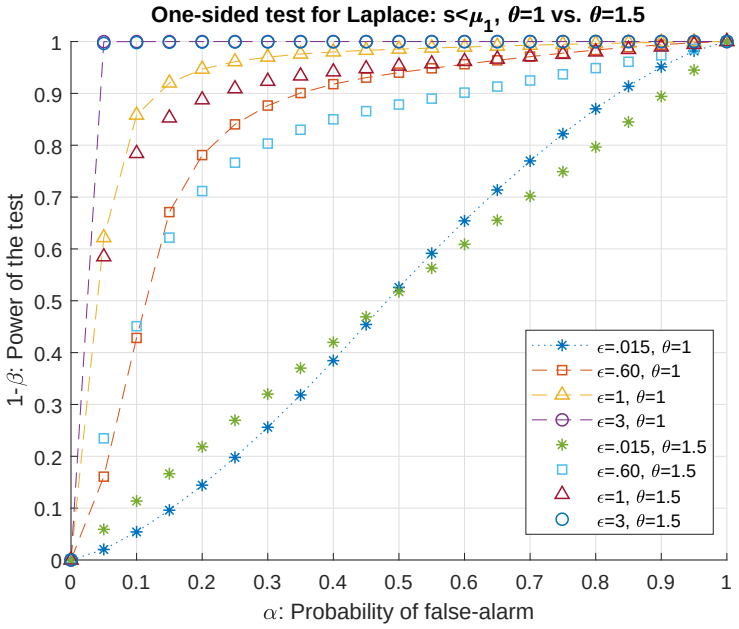
Theorem

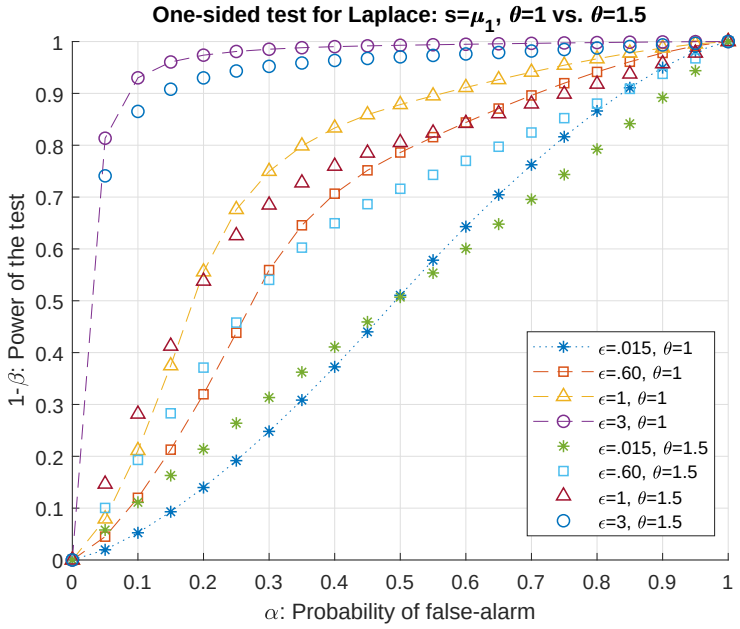
The threshold of the best critical region of size α for choosing between H_0 and H_1 of the two-sided hypothesis testing problem with the largest power $\bar{\beta}$ is

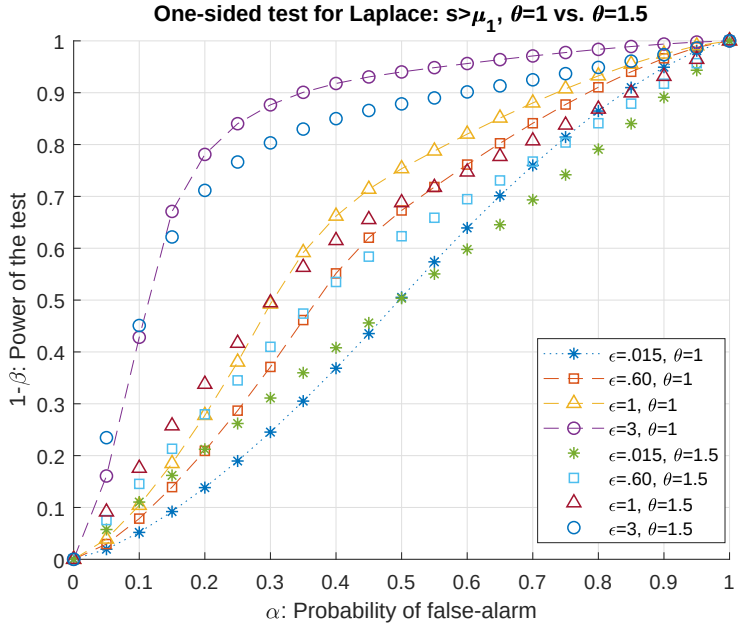
$$
\begin{aligned}
k_1 &= \mu_0 - (s/\epsilon)\log\alpha \\
k_2 &= \mu_0 + (s/\epsilon)\log\alpha
\end{aligned}
$$

Then according to the adversary's hypothesis testing problem, the defender fails to detect the attack when Y_0 is confined in $(f(x)+k_2, f(x)+k_1)$.

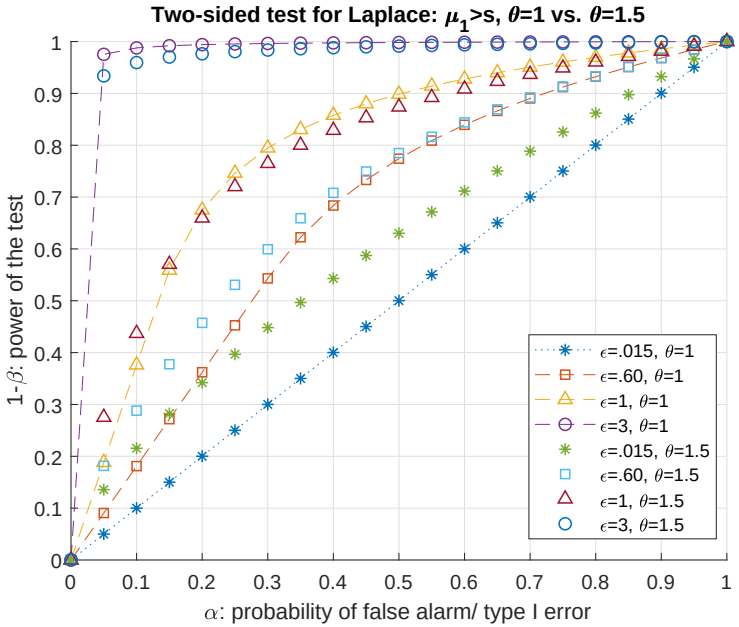
ROC Curves-One sided Test

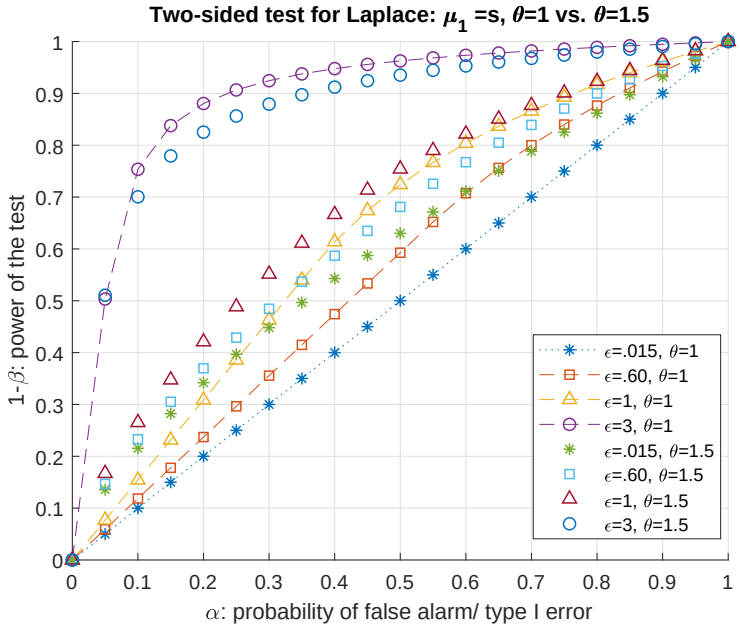


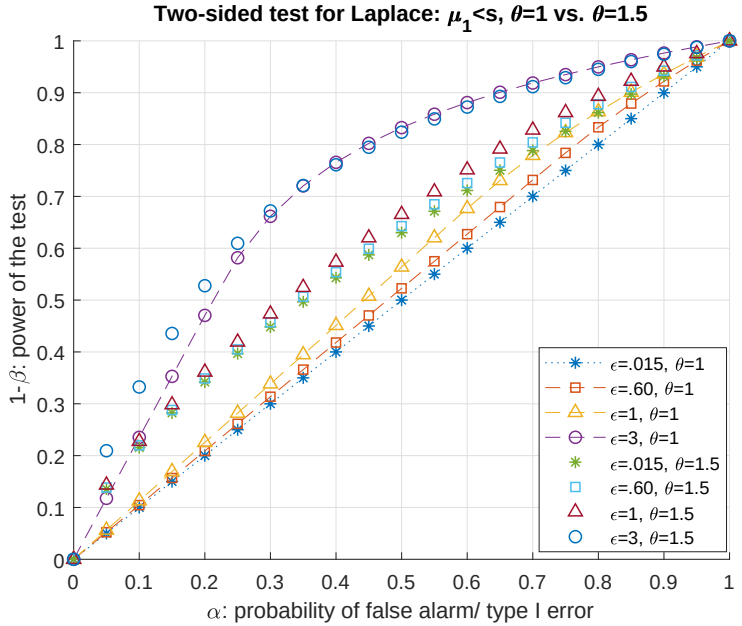




ROC Curves-Two sided Test





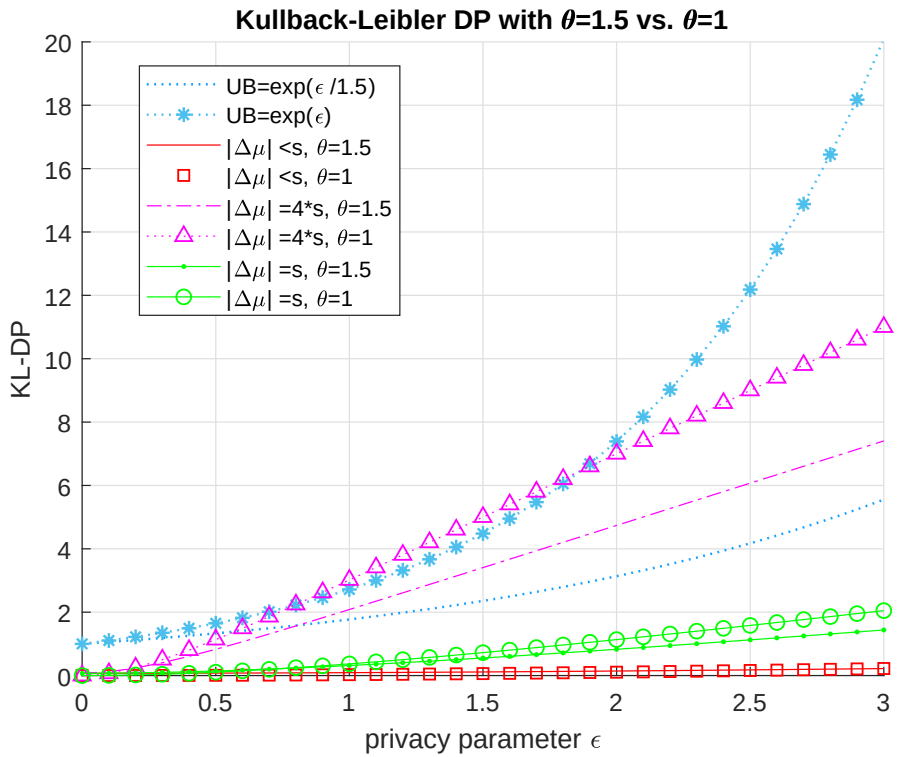


Kullback-Leibler DP for Adversarial Classification

Definition (Cuff & Yu 2016)

For a randomized mechanism $P_{Y|X}$ that guarantees $\epsilon-$ KL-DP, if the following inequality holds for all its neighboring datasets x and $\tilde{x}$.

$$D(P_{Y|X=x}||P_{Y|X=\tilde{x}}) \leq \exp\{\epsilon\}$$



References

- A. Ünsal and M. Önen, ”A Statistical Threshold for Adversarial Classification in Laplace Mechanisms”, IEEE ITW 2021
- C. Liu, X. He, T. Chanyaswad, S. Wang and P. Mittal, ”Investigating Statistical Privacy Frameworks from the Perspective of Hypothesis Testing”, Privacy Enhancing Technologies, pgs. 233-254, 2019
- J. Giraldo, A.A. Cardenas, M. Kantarcioglu and J. Katz, ”Adversarial Classification under Differential Privacy”, Network and Distributed Systems Security Symposium 2020, Feb. 2020, San Diego, CA, USA
- P. Cuff and L. Yu, ”Differential Privacy as a Mutual Information Constraint”, ACM SIGSAC Conference on Computer and Communications, Oct 24-28 2016, Vienna, Austria
- C. Dwork, Roth, A. ”The Algorithmic Foundations of Differential Privacy”, Foundations and Trends in Theoretical Computer Science, pgs. 211-407, 2014
- C. Dwork, F. McSherry, K. Nissim, and A. Smith, ”Calibrating Noise to Sensitivity in Private Data Analysis”, Theory of Cryptography Conference, 2006, pp. 265–284.