Information Bounds and Convergence Rates for Side-Channel Security Evaluators

Loïc Masure, Gaëtan Cassiers, Julien Hendrickx, François-Xavier Standaert

IACR Transactions on Cryptographic Hardware and Embedded Systems
ISSN 2569-2925, Vol. 2023, No. 3, pp. 522–569. DOI:10.46586/tches.v2023.i3.522-569

Information Bounds and Convergence Rates for Side-Channel Security Evaluators

Loïc Masure[1], Gaëtan Cassiers[2]*,
Julien Hendrickx[1], François-Xavier Standaert[1]

[1] UCLouvain, ICTEAM, Crypto Group, Louvain-la-Neuve, Belgium
firstname.lastname@uclouvain.be
[2] TU Graz, Graz, Austria, firstname.lastname@iaik.tugraz.at

Abstract. Current side-channel evaluation methodologies exhibit a gap between inefficient tools offering strong theoretical guarantees and efficient tools only offering heuristic (sometimes case-specific) guarantees. Profiled attacks based on the empirical leakage distribution correspond to the first category. Bronchain *et al.* showed at Crypto 2019 that they allow bounding the worst-case security level of an implementation, but the bounds become loose as the leakage dimensionality increases. Template attacks and machine learning models are examples of the second category. In view of the increasing popularity of such parametric tools in the literature, a natural question is whether the information they can extract can be bounded.
In this paper, we first show that a metric conjectured to be useful for this purpose, the hypothetical information, does not offer such a general bound. It only does when the assumptions exploited by a parametric model match the true leakage distribution. We therefore introduce a new metric, the training information, that provides the guarantees that were conjectured for the hypothetical information for practically-relevant models. We next initiate a study of the convergence rates of profiled side-channel distinguishers which clarifies, to the best of our knowledge for the first time, the parameters that influence the complexity of a profiling. On the one hand, the latter has practical consequences for evaluators as it can guide them in choosing the appropriate modeling tool depending on the implementation (*e.g.*, protected or not) and contexts (*e.g.*, granting them access to the countermeasures' randomness or not). It also allows anticipating the amount of measurements needed to guarantee a sufficient model quality. On the other hand, our results connect and exhibit differences between side-channel analysis and statistical learning theory.

Keywords: Profiled Attacks · Perceived Information · Training Information

1 Introduction

Evaluating the security of a cryptographic implementation against side-channel attacks is a complex problem. Since their introduction by Kocher *et al.* in the late nineties [KJJ99], a broad literature has focused on analyzing physical leakage in order to perform concrete attacks efficiently and to assess physical security on theoretically sound bases.

A first step towards such sound bases is the separation between non-profiled and profiled attacks. While Kocher's seminal work and early variants like Brier *et al.*'s Correlation Power Analysis (CPA) exploit an *a-priori* leakage model [BCO04], it has been shown that profiling the target device (*i.e.*, leveraging an open sample to estimate a leakage model)

[*] The work was done in parts while being a Research Fellow of the Belgian Fund for Scientific Research (FNRS-F.R.S.) with UCLouvain and in parts while being with Lamarr Security Research.

Received: 2023-01-15 Accepted: 2023-03-15 Published: 2023-06-09

can significantly improve the attacks' efficiency. Chari *et al.* introduced profiled attacks, and stated that such attacks are "the strongest form of side-channel attack possible in an information theoretic sense" [CRR02]. This statement seeded a line of works on worst-case side-channel security, *i.e.*, the security level reached when universally quantifying over the adversary. Standaert *et al.* observed that profiled attacks are critical to estimate the worst-case security of an implementation [SMY09]. Whitnall *et al.* extended this observation and proved that profiling is in general necessary for this purpose (*i.e.*, there is no generic attack strategy enabling us to recover secret information from a physically observable device's leakage without any a priori knowledge about the device's leakage distribution) [WOS14]. Heuser *et al.* finally proved that a generalized version of Chari et al.'s strategy, namely distinguishing thanks to the probability distribution of the leakage conditioned on the targeted secret, is indeed optimal in an information theoretic sense [HRG14].

A second step towards sound side-channel security evaluations is the acknowledgment that even in the profiled evaluation setting, performing an optimal attack in the sense of Heuser *et al.* is a highly non-trivial task. The main reason is that the true leakage distribution of a device is in general unknown and can be quite complex to estimate, especially in the presence of countermeasures like masking [CJRR99].

As a result, one can summarize the evaluation problem in two questions:

1. What is the data complexity of the attack using an optimal profiled model?
2. What is the profiling data complexity to estimate this optimal model?

Here, both data complexities are defined in terms of number of measured traces.

The first question is standard in the cryptographic setting. It aims at determining the level of security that can be guaranteed against an informed adversary. Since running an attack to evaluate its complexity for highly secure cryptographic implementations can be prohibitively expensive, an increasingly standard evaluation approach consists in using information theoretic metrics for this purpose. In particular, the Mutual Information (MI) can be used to bound the data complexity of worst-case attacks [DFS15, dCGRP19, MRS22, BCG$^+$23]. The difficulty of estimating the MI [Pan03], which we elaborate later in this paper, has led Renauld *et al.* to identify the Perceived Information (PI) as a metric capturing the amount of information that can be extracted from physical leakage thanks to the adversary/evaluator's (parametric) model, possibly biased by estimation or assumption errors [RSV$^+$11]. Durvaux *et al.* therefore formalized leakage certification as the problem of assessing the distance between the PI and the MI [DSV14].

Bronchain *et al.* showed that the PI is in general (*i.e.*, for any leakage distribution, including for masked implementations) a lower bound for the MI and that an upper bound is obtained by estimating the empirical Hypothetical Information (eHI), which is the amount of information that would be extractable from a device if the true distribution was identical to the one of a measured evaluation dataset [BHM$^+$19]. They additionally showed that, when increasing the dataset size, the expected value of the eHI asymptotically converges towards the MI. Unfortunately, the practical impact of these results is limited since the required dataset size grows with the number of points in the leakage traces, becoming very quickly impractical. The informal workaround proposed by Bronchain *et al.* is to use the HI estimated with a parametric model in such cases. Informally, and while the non-empirical HI loosens the formal link with the MI, the goal is to use the parametric HI as an upper bound for the complexity of the evaluator's best attack. They conjectured that this HI is an upper bound of the PI estimated with the same model.

The second question is less standard in the cryptographic setting. It rather aims at determining whether a worst-case attack is somewhat "practical". In other words, despite the profiling of a leakage model is a one-time effort, could it be so complex that estimating an accurate model becomes unrealistic. To the best of our knowledge, investigations in this direction have been less formal so far. Numerous profiling techniques have been introduced and evaluated based on specific case studies. These include extensions of Chari *et al.*.'s Template Attacks (TA) [CRR02, SLP05, GLP06, APSQ06, SA08, SKS09, CK13, CK14] and a steadily increasing (and not exhaustive) list of works leveraging machine (and deep) learning [HGM$^+$11, HZ12, LMBM13, LBM14, LPB$^+$15, MPP16, CDP17, CCC$^+$19, ZBHV20, WAGP20, ZBD$^+$21]. Recently, Masure *et al.* showed that these profiling strategies are not disconnected: by optimizing the appropriate loss function, evaluation approaches based on machine learning and deep learning actually target the same goal as TA, namely maximizing the PI [MDP20]. However, a systematic characterization of the parameters that influence the profiling phase of a side-channel attack, which would answer the practicality question, is still missing. For example, how does the convergence of a machine learning model depend on the physical leakage characteristics (noise level, number of dimensions, security order), number of classes and number of profiling traces? And are some statistical tools better suited depending on the contexts?

Our contributions regarding these two main questions are twofold:

Regarding the first question, we falsify and fix the conjecture of Bronchain *et al.* Precisely, we show that the parametric HI is not always an upper bound of the parametric PI. Since our counterexample corresponds to realistic leakage distributions (namely, mixture distributions that happen with masked implementations), we then propose a new metric, the Training Information (TI_N), that eliminates this limitation. While the HI can be viewed as a measure of a parametric model tested against itself, the TI_N is a measure of a parametric model tested against (the empirical distribution of) its training samples. We show that for parametric leakage models that optimize the appropriate loss function, the TI_N upper bounds the "learnable information" (LI) defined as the supremum of the PI over a parametric class of models, and that for $N \to \infty$, the PI and TI_N converge towards the LI. Like the HI, the TI_N does not offer guarantees against assumption errors when it is computed for parametric models: the LI may be smaller than the MI. But it offers an easy way to bound estimation errors (i.e., LI − PI) for practically relevant classes of distinguishers. Besides, it can be used for both generative and discriminative models (while the HI was limited to the first ones). This allows evaluators to gauge how much their attacks can be improved by collecting more profiling traces, and to stop their measurement campaigns when the gain becomes small. In other words, this new metric answers the question: how much information can be learned with my leakage model?

Regarding the second question, we initiate a study of the convergence rate of the TI_N and PI metrics for practically-relevant profiling techniques. Namely, we consider simple representatives of two widely-used profiled attack families. For the Gaussian templates, we consider the original attack of Chari *et al.* [CRR02], denoted in this paper as gTA, and its variant with *pooled* covariance matrix estimation [CK13], denoted as p-gTA. For the deep learning attacks, we analyze a Multi-Layer Perceptron (MLP) with L layers and W weights to fit, trained with a negative log-likelihood loss function. Although less common in side-channel attacks, we also consider the k^{th}-order logistic regression, denoted as LR_k, which is interesting since this model is similar to Gaussian templates but its training process is closer to the one of the MLP. Our results are synthesized in Table 1.

On the one hand, this table positively answers our question regarding the practicality of the profiling phase in a security evaluation. It shows that there are profiling tools for which the estimation error is inversely proportional to $\sqrt{N}$ (N being the number of profiling traces) for any (even protected) implementation (e.g., MLP and LR_k). It also shows that the convergence rate of the models depends on their hyperparameters but not on the physical leakage characteristics (i.e., the true leakage distribution), and consolidates the general intuition that side-channel security evaluations are a trade-off between the genericity and the efficiency of the profiling. On the other hand, the table shows that there are statistical tools that are better suited depending on the evaluation contexts. For

Table 1: Convergence of the PI of different profiling tools (the $\widetilde{\mathcal{O}}(\cdot)$ notation ignores log terms). The "Fast regime" column assumes that, for some ideally chosen values of the parameters, the model can perfectly match the true leakage distribution.

)

Model	Fast Regime	General Bound
MLP	$\widetilde{\mathcal{O}}\left(\frac{QWL}{N}\right)$	$\widetilde{\mathcal{O}}\left(\sqrt{\frac{QWL}{N}}\right)$
k^{th}-order logistic regression (LR_k)	$\widetilde{\mathcal{O}}\left(\frac{QD^k}{N}\right)$	$\tilde{\mathcal{O}}\left(\sqrt{\frac{Q\cdot D^k}{N}}\right)$
Gaussian templates (gTA)	$\mathcal{O}\left(\frac{QD^2}{N}\right)$	
Pooled Gaussian templates ($\mathsf{p\text{-}gTA}$)	$\mathcal{O}\left(\frac{QD}{N}\right)$ for $Q = 2$	

Q denotes the number of profiled classes, D the dimensionality of the traces, and N the number of traces acquired for profiling, *i.e.*, quantify the *sample complexity* of profiling.

example, the convergence rate of LR_k for a security order k leads the modeling error to scale in $\mathcal{O}(D^k)$. By contrast, for a circuit of complexity k (*e.g.*, the masking of a sensitive variable that would leak $D = k$ samples corresponding to the shares), it is always possible to build an MLP whose complexity $W \cdot L$ scales as $\mathsf{poly}(D = k)$ [SB14, Thm. 20.3]. So if an evaluator has to profile higher-order leakages, leveraging MLPs leads to a more efficient profiling than trying to profile moments of the leakage distribution with LR_k.

As discussed in Section 7, we hope these theoretical results can help evaluators operating within a limited time frame towards finding the best trade-off in their model selection, by anticipating and optimizing the models' profiling complexity.

1.1 Related Works

The use of information theoretic metrics to guide/compare profiled attacks dates back to [SKS09]. In a work from Cosade 2021 [PBP21], Picek *et al.* show that this intuition does not only hold for the number of profiling traces but also for the number of epochs used in the training phase of a machine learning model. Ito *et al.* show that the direct optimization of security metrics such as the Success Rate (SR) or Guessing Entropy (GE) [SMY09] can slightly improve an optimization guided by information theoretic metrics in some contexts, at the cost of some computational overheads [IUH22]. It follows previous observations that security metrics and information theoretic metrics can sometimes lead to comparatively different outcomes (*e.g.*, for low noise levels or small number of attack traces) [SPAQ06, PHJ+19]. Yet, since information theoretic metrics are inversely proportional to the asymptotic complexity of a side-channel attack phase, the concrete impact of such an observation is also limited. For example, the experiments performed in [IUH22] show some gains for attacks that succeed in 400 traces, but these gains already vanish for attacks succeeding in more than 1,000 traces. So while such results are interesting to push the optimization of concrete attacks in specific contexts, they do not contradict the general relevance of information theoretic metrics for side-channel security evaluations. Finally, Cristiani *et al.* investigate the so-called *Neural-based* MI *estimation* (MINE) [CLM20]. They leverage the variational formulation of the MI allowing to train an MLP to maximize a lower bound of the MI, similarly to the PI [CT12, Eq. (8.93)]. This research follows the observation of Mather *et al.* [MOBW13] that an evaluator may estimate the complexity of her best attack without having to mount it. Analyzing whether this complementary approach could be used to upper bound the information leakage like the TI_N and assessing its convergence rate are interesting scopes for further investigation.

2 Background

Notations. In the following, we denote random variables (resp., random vectors) by upper-case (resp., bold upper-case) letters X (resp., $\boldsymbol{X}$). We denote by the same calligraphic letter $\mathcal{X}$ the observation domain of the corresponding random variable (resp., random vector). We denote observations of a random variable (resp., random vector) by the corresponding lower-case roman letter x (resp., $\boldsymbol{x}$). If a random variable X is discrete, we denote by $\Pr(X = x)$ its probability mass function (pmf), for which we will use the shortcut notation $\mathsf{p}(x)$. We note $\mathcal{P}(\mathcal{V})$ the set of probability distributions over a random variable of domain $\mathcal{V}$. If p and m denote two distributions over the same support, the Kullback - Leibler (KL) divergence is denoted by $\mathsf{D}_{\mathsf{KL}}(\mathsf{p} \parallel \mathsf{m}) = \underset{X \sim \mathsf{p}}{\mathbb{E}}\left[\frac{\mathsf{p}(X)}{\mathsf{m}(X)}\right]$. We use the notation $\mathcal{O}(f(n))$ to hide constant factors in n, and the notation $\widetilde{\mathcal{O}}(f(n))$ to additionally hide log factors in n. For a square matrix A, we denote by $\|A\|_*$ its spectral norm (*i.e.*, the greatest of its eigenvalues in absolute value) and by $\|A\|_F$ its Frobenius norm.

2.1 Information Theoretic Metrics

Let Y be a discrete uniform random variable over a domain $\mathcal{Y}$, denoting the sensitive intermediate computation targeted by the attacker/evaluator, and $\boldsymbol{L}$ be a discrete random vector over a domain $\mathcal{L}$, denoting the corresponding physical measurement of the leakage of Y. During its attack, the adversary/evaluator, who knows the distribution of Y, acquires a *profiling* set $\mathcal{S}_N$ made of N observations $(y, \boldsymbol{l})$ of the joint probability distribution of $(Y, \boldsymbol{L})$. We consider the problem of estimating a *discriminative* model $\mathsf{m}(y \mid \boldsymbol{l})$ for the true conditional Probability Mass Function (PMF) $\Pr(Y = y \mid \boldsymbol{L} = \boldsymbol{l})$, for which we will use the shortcut notation $\mathsf{p}(y \mid \boldsymbol{l})$. In some cases, we also care about a *generative* model $\mathsf{m}(\boldsymbol{l} \mid y)$ for the true PMF $\Pr(\boldsymbol{L} = \boldsymbol{l} \mid Y = y)$, denoted for short as $\mathsf{p}(\boldsymbol{l} \mid y)$. We note that, since the distribution of Y is known, a generative model naturally induces a discriminative model (using Bayes' rule). We further define a distance metric Δ between a generative model m and a discriminative model m' (a probability distribution p may also be used in place of one (or two) of the models):

$$\Delta_{\mathsf{m}}^{\mathsf{m}'} = \mathsf{H}(Y) + \sum_{y \in \mathcal{Y}, \boldsymbol{l} \in \mathcal{L}} \mathsf{m}(y, \boldsymbol{l}) \cdot \log_2\left(\mathsf{m}'(y \mid \boldsymbol{l})\right), \tag{1}$$

where $\mathsf{H}(Y)$ is the entropy of Y. Thanks to this notation, we can express the Mutual Information (MI) between the random variables Y and $\boldsymbol{L}$ as

$$\mathsf{MI}(Y; \boldsymbol{L}) = \Delta_{\mathsf{p}}^{\mathsf{p}}.$$

The MI is a relevant evaluation metric for side-channel attacks since the (measurement) complexity of a worst-case side-channel attack targeting a secret key, *e.g.*, $y = \mathsf{S}(x \oplus k)$ where x denotes a plain text, k denotes a secret key chunk, and S denotes an S-box, is inversely proportional to $\mathsf{MI}(Y; \boldsymbol{L})$ [DFS19, dCGRP19]. However, this metric cannot be computed directly since the true leakage distribution (*i.e.*, $\mathsf{p}(\boldsymbol{l} \mid y)$) is in general unknown. One solution is to estimate it, which is known to be a difficult problem [Pan03]. Alternatively, the amount of information that can be extracted from the leakages thanks to a model can be quantified by the *Perceived Information* (PI) given by

$$\mathsf{PI}(Y; \boldsymbol{L}; \mathsf{m}) = \Delta_{\mathsf{p}}^{\mathsf{m}}.$$

The authors in [BHM+19] additionally considered the Hypothetical Information (HI):

$$\mathsf{HI}(Y; \boldsymbol{L}; \mathsf{m}) = \Delta_{\mathsf{m}}^{\mathsf{m}},$$

and the empirical Hypothetical Information (eHI) defined as

$$\mathsf{eHI}_N(Y;\boldsymbol{L}) = \Delta_{\tilde{\mathsf{e}}_{\mathcal{S}_N}}^{\tilde{\mathsf{e}}_{\mathcal{S}_N}} \ ,$$

where $\tilde{\mathsf{e}}$ denotes the operator that maps a profiling set $\mathcal{S}_N$ to the corresponding *empirical distribution*, *i.e.*, $\tilde{\mathsf{e}}_{\mathcal{S}_N}(y,\boldsymbol{l}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{(y,\boldsymbol{l})=(y_i,\boldsymbol{l}_i)}$. Whenever there is no ambiguity, we will replace the notation $\tilde{\mathsf{e}}_{\mathcal{S}_N}$ by $\tilde{\mathsf{e}}_N$. Based on these quantities, their main result is twofold. First, the PI is always upper bounded by the MI regardless of the tested model m, with equality if and only if m coincides with the true leakage distribution p. Second, the eHI may be used to bound the MI as follows:

$$\underset{\tilde{\mathsf{e}}_{N-1}}{\mathbb{E}}\left[\mathsf{eHI}_{N-1}(Y;\boldsymbol{L})\right] \geq \underset{\tilde{\mathsf{e}}_{N}}{\mathbb{E}}\left[\mathsf{eHI}_{N}(Y;\boldsymbol{L})\right] \geq \mathsf{MI}(Y;\boldsymbol{L}) \ . \tag{2}$$

Note that the bound is for the expectation of the HI over the model estimations. It only holds for the empirical distribution $\tilde{\mathsf{e}}_N$ and the authors also show that

$$\underset{\tilde{\mathsf{e}}_{N}}{\mathbb{E}}\left[\mathsf{eHI}_{N}(Y;\boldsymbol{L})\right] \underset{N\to\infty}{\longrightarrow} \mathsf{MI}(Y;\boldsymbol{L}) \ . \tag{3}$$

By contrast, the PI bound is true for any model.

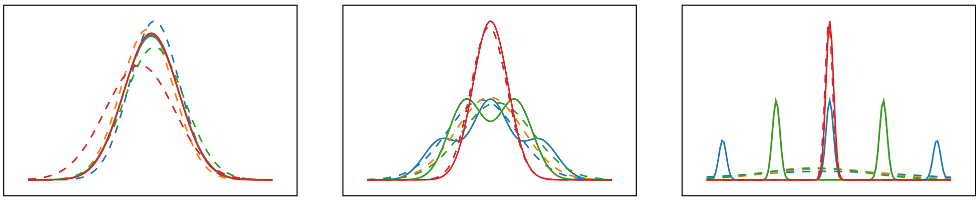
3 Limitations of the HI

One important question left open by Bronchain *et al.* is whether the properties of the HI generalize to parametric leakage models. This question is important since, as experimentally observed in [BHM+19], assessing the security of an implementation with an empirical model (and the corresponding bounds) rapidly becomes too expensive. In this section, we consolidate this HI proposal in two directions. First, we give a counter-example contradicting that the HI is in general (*i.e.*, for any model) an upper bound for the PI. In our example, it appears that this conjecture only holds when the parametric model used in the bound corresponds to the true leakage function to a sufficient extent. This will lead us to introduce a new metric to fix this issue in Section 4. Second, we formalize the observation that empirical models converge too slowly for being a practical alternative in (multivariate) side-channel security evaluations. For this purpose, we reconsider the convergence of the eHI towards the MI. Bronchain *et al.* proved a monotone convergence of the expectation. However, in practice the profiling dataset acquisition is usually performed a single time by the evaluators. Accordingly, stronger notions of convergence (*e.g.*, in probability) are better suited to argue about the profiling phase of a side-channel attack. We give such a stronger result in Section 3.2, while also showing that an evaluation based on the eHI suffers from very slow convergence rates. In particular, it suffers from a bias that grows exponentially with the trace dimensionality.

3.1 Inconsistency with Non-Empirical Models

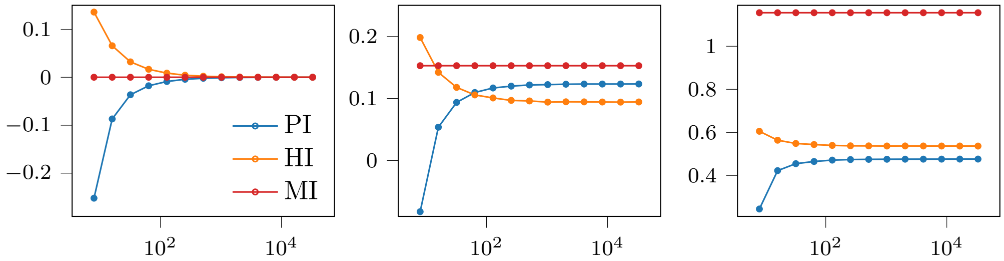
In [BHM+19], the authors proposed the gHI (*i.e.* the HI computed for a Gaussian model) as a surrogate of the eHI enabling a faster convergence. We next show empirically that we can actually observe all three possible cases for the convergence of the PI and HI in a quite realistic context: either they both converge to the same asymptotic value, or the HI converges strictly above the PI, or the HI converges strictly below the PI.

We illustrate the three cases by measuring the gHI against true distributions that are not Gaussian. In particular, we use discretized univariate Gaussian mixture models which are relevant in the context of masked implementations. Concretely, the leakage is the sum of a Gaussian noise and the Hamming weight of the sharing $(x \oplus r, r)$ for the n-bit word



(a) SNR $= 0.02$ **(b)** SNR $= 2$ **(c)** SNR $= 200$

Figure 1: True distributions (continuous lines) and models (dashed lines) trained with 20 samples for each of the 4 classes (*i.e.* $n = 2$ bits). The X axis is the value of the leakage and the Y axis axis is its probability density.



(a) SNR $= 0.02$ **(b)** SNR $= 2$ **(c)** SNR $= 200$

Figure 2: gPI, gHI and MI (Y axis, in bits) for 2-bit masked variable as a function of the number of traces used to train the Gaussian model (X axis).

x, masked with a uniformly random n-bit word r. The model, for each leakage class (*i.e.* $x = 0$ and $x = 1$) is a Gaussian fitted using maximum likelihood estimators. In Figure 1, we show the leakage (continuous lines) and the models (dashed lines) for two distinct values of the SNR, computed as the ratio between the variance of the Hamming weight of an n-bit uniformly random variable, and the variance of the Gaussian noise [Man04].

In Figure 2, we show the corresponding gPI, gHI and MI. In addition to the observation of the aforementioned three cases, we can look at the relationship between the gPI/gHI and the MI. When the true distribution is close to Gaussian (Figure 1a), both gPI and gHI converge to the MI, as conjectured. However, in the other cases, the gPI and gHI are below the MI. This is explained by the inability of the Gaussian model to accurately represent the distinctive features of the classes, and thus to exhibit good class discrimination. Visually, the more dissimilarity between the true leakage and the model (*i.e.*, from left to right in Figure 1), the wider the gap between HI and MI (from left to right in Figure 2).

3.2 Slow Convergence of the Empirical Model

We now formalize the observation that empirical models converge too slowly for being a practical alternative in side-channel security evaluations.

3.2.1 Convergence of the Expectation.

We first state that the bias of eHI scales exponentially in the dimensionality of the traces D and linearly in $\frac{Q}{N}$, with Q the number of classes and N the number of profiling traces.

Theorem 1. *Consider an evaluator sampling N traces from a D-dimensional leakage with an ω-bit resolution, related to a sensitive intermediate computation over Q classes,*

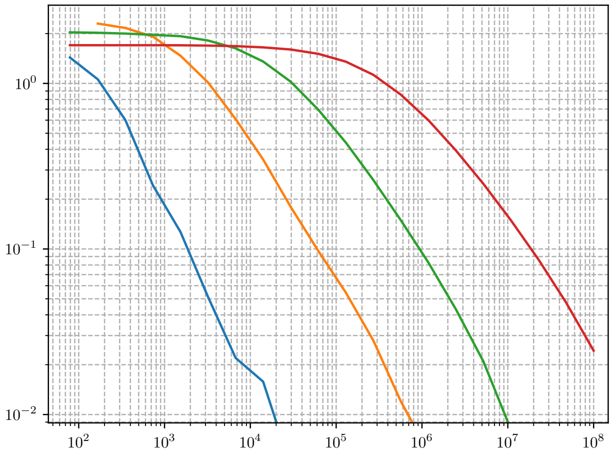


Figure 3: eHI − MI (y-axis) with respect to the number of profiling traces N (x-axis) for $D = 1$ (blue), 2 (orange), 3 (green), and 4 (red). Here, $\omega = 4$ and $Q = 16$.

assumed to be uniformly distributed. Then, the eHI *satisfies the following inequalities:*

$$\mathsf{MI}(Y;\boldsymbol{L}) \leq \mathbb{E}\left[\mathsf{eHI}_N\right] \leq \mathsf{MI}(Y;\boldsymbol{L}) + \frac{BQ}{N} \quad , \tag{4}$$

where B denotes the number of bins in the empirical distribution. In particular, here $B = 2^{\omega D}$. Moreover,

$$\Big(\mathbb{E}\left[\mathsf{eHI}_N\right] - \mathsf{MI}(Y;\boldsymbol{L})\Big) \cdot \frac{N}{BQ} \xrightarrow[N\to\infty]{} 1/2 \quad . \tag{5}$$

The proof of this statement is directly inspired from Paninski's work [Pan03], and is detailed in Appendix A. Note that as a consequence of Equation 5, the upper bound of Equation 4 is asymptotically tight, thereby meaning that the lower bound is asymptotically loose. Since there is no unbiased estimator of the MI [Pan03, Prop. 8], this is unavoidable (otherwise removing the right term of Equation 4 would have given an unbiased estimator of the MI). We illustrate this result with the auxiliary source code released by Bronchain *et al.* with the paper [BHM+19].[1] Figure 3 depicts the absolute difference between eHI_N and MI with respect to the number N of profiling traces, simulated according to a "Hamming weight + Gaussian noise" leakage model, with a trace dimensionality ranging from 1 to 4.We can see that every curve has the same slope of roughly -1 with a constant offset between each other, which confirms the theoretical expectations of Theorem 1.

3.2.2 Convergence in Probability.

So far we provided a speed of convergence of the expectation of the eHI towards the MI. As already mentioned, such a result is not directly representative of an evaluation context where the profiling phase is (ideally) performed once. For example, the results shown in Figure 3 depict the convergence of eHI for *one* simulation, whereas Theorem 1 only ensures that the shape of the curves observed in Figure 3 are the ones that are expected *on average*, *i.e.* over several simulations. It might however be possible that by (lack of) chance, one could observe different results for one particular eHI computation. We next eliminate this

[1] https://github.com/obronchain/Leakage_Certification_Revisited

limitation by discussing/proving a stronger notion of convergence, namely the convergence in probability. Incidentally, Bronchain *et al.* already proved the convergence in probability, in the proof of [BHM+19, Lemma 2, p. 10], although not claimed as a theoretical result in their paper. In this section, we additionally provide upper bounds on the rate of convergence in probability. We state hereafter that the deviation between the eHI and its expected value converges towards 0 at a speed $\mathcal{O}\left(\frac{\log(N)}{\sqrt{N}}\right)$.

Theorem 2. *For all $\delta > 0$, the inequality*

$$\left|\mathsf{eHI}_N - \mathbb{E}\left[\mathsf{eHI}_N\right]\right| \leq \log_2(N)\sqrt{\frac{8\log(4/\delta)}{N}} \tag{6}$$

holds with probability at least $1-\delta$, and furthermore

$$\mathbb{E}\left[\left|\mathsf{eHI}_N - \mathbb{E}\left[\mathsf{eHI}_N\right]\right|\right] \in \Theta\left(\frac{1}{\sqrt{N}}\right) \quad .$$

The proof of Theorem 2 is provided in Appendix A and is also directly inspired by Paninski's work [Pan03]. Interestingly, the convergence rate of Equation 6 does not depend on D, while the bias increases exponentially with D. When the number of dimensions is large, the bias will therefore dominate for practical N, despite the faster convergence rate of the bias with respect to N. In that case, the eHI is thus an upper-bound of the MI with high probability, although so loose that it is of little interest. Overall we conclude that the eHI converges too slowly for many practical use-cases, which calls for a better solution (which is not provided by the non-empirical HI, as discussed in Section 3.1).

4 Introducing the Training Information

The previous section showed the HI metric limitations both in terms of its ability to bound the information that can be extracted with parametric models and in terms of the convergence rate that its instantiation with the empirical function leads to. In this section, we introduce a new metric to circumvent these limitations, which we call the Training Information (TI_N). Like the eHI, it upper-bounds the PI while also having much better quantitative convergence properties. To explain the intuition behind the TI_N, we recall that the eHI is the quantity $\Delta_{\tilde{\mathsf{e}}_N}^{\tilde{\mathsf{e}}_N}$, where Δ is the operator defined in Equation 1, whereas the HI, in its general form (*i.e.*, defined for an arbitrary model m), is given by $\Delta_{\mathsf{m}}^{\mathsf{m}}$, and the PI is given by $\Delta_{\mathsf{p}}^{\mathsf{m}}$, where p denotes the true (unknown) leakage distribution. The main goal of the TI_N is to base the metric on a parametric model (enabling faster convergence), while keeping an upper bound for the PI. For this purpose, the eHI upper-bounds the MI by *overfitting*: it builds an ideal discriminative model $\tilde{\mathsf{e}}_N$ (in the superscript) based on some samples, then *evaluates it* on the same samples (in the subscript). We define the TI_N as $\Delta_{\tilde{\mathsf{e}}_N}^{\mathsf{m}}$, where m is trained on the same sample set as the one used to compute $\tilde{\mathsf{e}}_N$. Since the TI_N is based on a model instead of the empirical distribution, it carries the possible biases induced by the choice of possible models (*e.g.*, Gaussian distributions). Hence it cannot upper-bound the MI in general (*e.g.*, if the true distribution is not Gaussian). However, we can still relate the TI_N and the PI to a meaningful quantity that we name the Learnable Information (LI for short). The LI is the maximum amount of information that can be extracted from a given leakage distribution using a family of models, and the gap between the LI and the MI corresponds to the "assumption error" of the evaluator/attacker's model [DSV14]. Informally, we have the following inequalities: $\mathsf{PI} \leq \mathsf{LI} \leq \mathsf{TI}$. We next formalize the concepts of LI and TI_N in Section 4.1, then prove the above inequalities and prove that the expectation of the TI_N converges in Equation 4.2.

4.1 Definition and Rationale

We first formalize the notion of "family of models" as follows.

Definition 1 (Hypothesis class)**.** A *hypothesis class* $\mathcal{H}$ is a – possibly infinite – collection of discriminative models $\mathsf{m} : \mathcal{L} \to \mathcal{P}(\mathcal{Y})$, where $\mathcal{L}$ denotes the input space of the random vector $\boldsymbol{L}$ of the side-channel trace, and $\mathcal{Y}$ denotes the finite set of all hypothetical values of the target discrete random variable Y.

The output of m can be seen as a possible discrete probability distribution of the target random variable Y, while an hypothesis class can be understood as "a model where the parameters are not yet fixed" (*e.g.* the set of MLPs with a given structure is an hypothesis class). Using this notion of hypothesis class, we next define the LI.

Definition 2 (Learnable Information)**.** Let $\mathcal{H}$ be a hypothesis class. The *learnable information* on Y from leakage $\boldsymbol{L}$ using a model from $\mathcal{H}$ is defined as the quantity:

$$\mathsf{LI}(Y;\boldsymbol{L};\mathcal{H}) = \sup_{\mathsf{m}\in\mathcal{H}} \mathsf{PI}(Y;\boldsymbol{L};\mathsf{m}) \quad . \tag{7}$$

In order to introduce the training information, we need two more definitions.

Definition 3 (Learning Algorithm)**.** A *learning algorithm* $\mathcal{A}$ for a hypothesis class $\mathcal{H}$ is a function

$$\mathcal{A} : \bigcup_{N=1}^{\infty} (\mathcal{Y} \times \mathcal{L})^N \to \mathcal{H}, \tag{8}$$

taking as an input a set $\mathcal{S}_N$ of N acquisitions drawn from the (unknown) joint probability distribution of $(Y, \boldsymbol{L})$ and returning a model $\mathsf{m} = \mathcal{A}(\mathcal{S}_N)$ from the hypothesis class $\mathcal{H}$.

It is worth noticing that in a profiled attack scenario, the adversary can be defined by its underlying learning algorithm. Hence, in this paper, we denote interchangeably by $\mathcal{A}$ either an adversary, or its corresponding learning algorithm. The following definition states how we compare different learning attackers, *i.e.*, learning algorithms.

Definition 4 (Regret)**.** Let $\mathcal{A}$ be an attacker, *i.e.*, a learning algorithm. The *regret* of $\mathcal{A}$ is the following quantity:

$$\mathsf{R}\left(\mathcal{A}\right) = \mathsf{MI}(Y;\boldsymbol{L}) - \mathsf{PI}(Y;\boldsymbol{L};\mathcal{A}(\mathcal{S}_N)) \quad . \tag{9}$$

By definition, the regret is always non-negative, and equals 0 if and only if the learning algorithm outputs the exact leakage model, *i.e.* $\mathcal{A}(\mathcal{S}_N) = \mathsf{p}$. We can now give the formal definition of TI_N, based on the Δ operator.

Definition 5 (Training Information)**.** Let $\mathcal{S}_N$ be a set of N samples drawn from a distribution over $(Y, \boldsymbol{L})$. The *training information* by $\mathcal{A}$ with N traces is defined as the following quantity:

$$\mathsf{TI}_N(Y;\boldsymbol{L};\mathcal{A}) = \Delta_{\tilde{\mathsf{e}}_{\mathcal{S}_N}}^{\mathcal{A}(\mathcal{S}_N)} \quad . \tag{10}$$

Since TI_N is defined for any learning algorithm, regardless of their performances, there is no prior reason why TI_N could be an upper bound of MI nor PI. Nevertheless, this is possible by adding a few more assumptions, in particular assuming that the learning algorithm is a TI_N maximizer as we next formalize.

Definition 6 (TI_N maximizer)**.** Let $\mathcal{H}$ a hypothesis class and let $\mathcal{S}_N$ be the dataset of N traces. The TI_N *maximizer for the hypothesis class* $\mathcal{H}$ is the learning algorithm $\mathcal{A}_{\mathcal{H}}$ such that $\mathcal{A}_{\mathcal{H}}(\mathcal{S}_N) = \widehat{\mathsf{m}}_N$, where $\widehat{\mathsf{m}}_N$ is defined as

$$\widehat{\mathsf{m}_{\mathcal{S}_N}} = \operatorname*{argmax}_{\mathsf{m}\in\mathcal{H}} \Delta_{\tilde{\mathsf{e}}_{\mathcal{S}_N}}^{\mathsf{m}} \quad . \tag{11}$$

For conciseness, we will replace the notation $\widehat{\mathsf{m}_{\mathcal{S}_N}}$ by $\widehat{\mathsf{m}}_N$ in the remaining of this paper.

4.2 Bound and Convergence of the TI_N

Provided with the TI_N maximizer of a hypothesis class, it is possible to derive properties similar to the ones conjectured for the gHI by Bronchain *et al.* [BHM+19]. The first one that we give hereafter tells that the maximum TI_N over a hypothesis class is an upper bound in expectation of the LI for the same hypothesis class. The second one tells that, for a TI_N maximizer, the expectation of the TI_N is monotonically decreasing. These two results imply that the expectation of the TI_N converges to an upper bound of the LI.

Proposition 1. *Let $\mathcal{H}$ be a hypothesis class, and N be a positive integer. Then*

$$\mathsf{LI}(Y;\boldsymbol{L};\mathcal{H}) \leq \mathbb{E}\left[\mathsf{TI}_N(Y;\boldsymbol{L};\mathcal{A}_{\mathcal{H}})\right] \ , \tag{12}$$

where the expectation is taken over the profiling set $\mathcal{S}_N$ of size N.

Proof. According to Definition 5 and Definition 6, for any model $\mathsf{m} \in \mathcal{H}$, if $\widehat{\mathsf{m}}_N$ denotes the maximum likelihood for $\mathcal{H}$, it holds that

$$\Delta_{\tilde{\mathsf{e}}_N}^{\widehat{\mathsf{m}}_N} \geq \Delta_{\tilde{\mathsf{e}}_N}^{\mathsf{m}} \ . \tag{13}$$

Since the expectation is monotone, non-decreasing, it follows that

$$\mathbb{E}\left[\mathsf{TI}_N(Y;\boldsymbol{L};\widehat{\mathsf{m}}_N)\right] = \mathbb{E}\left[\Delta_{\tilde{\mathsf{e}}_N}^{\widehat{\mathsf{m}}_N}\right] \geq \mathbb{E}\left[\Delta_{\tilde{\mathsf{e}}_N}^{\mathsf{m}}\right] \tag{14}$$

Since the $\Delta_{\mathsf{a}}^{\mathsf{b}}$ operator is linear with respect to a, it follows that

$$\mathbb{E}\left[\Delta_{\tilde{\mathsf{e}}_N}^{\mathsf{m}}\right] = \Delta_{\mathsf{p}}^{\mathsf{m}} = \mathsf{PI}(Y;\boldsymbol{L};\mathsf{m}) \ . \tag{15}$$

Since the latter holds regardless the choice for m we may arbitrarily take the model that maximizes the PI, which gives Equation 12. □

Proposition 2. *Let $\mathcal{H}$ be a hypothesis class, and N be a positive integer. Then*

$$\mathbb{E}\left[\mathrm{TI}_{N-1}(Y;\boldsymbol{L};\mathcal{A}_{\mathcal{H}})\right] \geq \mathbb{E}\left[\mathsf{TI}_N(Y;\boldsymbol{L};\mathcal{A}_{\mathcal{H}})\right] \ , \tag{16}$$

where the expectation is taken over the profiling set $\mathcal{S}_N$ of size N.

Proof. We first remark that we can extend the definition of the TI_N-maximizer to learn from an empirical distribution: let $\mathsf{e} \in \mathcal{P}(\mathcal{Y},\mathcal{L})$, we define

$$\widehat{\mathsf{m}}_{\mathsf{e}} = \underset{\mathsf{m}\in\mathcal{H}}{\operatorname{argmax}}\, \Delta_{\mathsf{e}}^{\mathsf{m}} \ .$$

We shall show that the function $\gamma : \tilde{\mathsf{e}}_N \mapsto \Delta_{\tilde{\mathsf{e}}_N}^{\widehat{\mathsf{m}}_{\tilde{\mathsf{e}}_N}}$ is convex. The theorem then follows from Lemma 2 of Bronchain *et al.* [BHM+19]. For any $\mathsf{e}, \mathsf{e}' \in \mathcal{P}(\mathcal{Y},\mathcal{L})$, $\alpha \in [0,1]$, let $\mathsf{e}'' = \alpha\mathsf{e} + (1-\alpha)\mathsf{e}'$. We show that $\gamma(\mathsf{e}'') \leq \alpha\gamma(\mathsf{e}) + (1-\alpha)\gamma(\mathsf{e}')$. First, using the linearity of $\Delta_{\mathsf{e}}^{\mathsf{m}}$ with respect to e, we have

$$\gamma(\mathsf{e}'') = \Delta_{\mathsf{e}''}^{\widehat{\mathsf{m}}_{\mathsf{e}''}} = \alpha\, \Delta_{\mathsf{e}}^{\widehat{\mathsf{m}}_{\mathsf{e}''}} + (1-\alpha)\, \Delta_{\mathsf{e}'}^{\widehat{\mathsf{m}}_{\mathsf{e}''}} \ .$$

Since $\widehat{\mathsf{m}}_{\mathsf{e}}$ and $\widehat{\mathsf{m}}_{\mathsf{e}'}$ are TI_N-maximizers, $\Delta_{\mathsf{e}}^{\widehat{\mathsf{m}}_{\mathsf{e}''}} \leq \Delta_{\mathsf{e}}^{\widehat{\mathsf{m}}_{\mathsf{e}}}$ and $\Delta_{\mathsf{e}'}^{\widehat{\mathsf{m}}_{\mathsf{e}''}} \leq \Delta_{\mathsf{e}'}^{\widehat{\mathsf{m}}_{\mathsf{e}'}}$, which gives

$$\gamma(\mathsf{e}'') \leq \alpha\, \Delta_{\mathsf{e}}^{\widehat{\mathsf{m}}_{\mathsf{e}}} + (1-\alpha)\, \Delta_{\mathsf{e}'}^{\widehat{\mathsf{m}}_{\mathsf{e}'}} = \alpha\gamma(\mathsf{e}) + (1-\alpha)\gamma(\mathsf{e}') \ .$$

□

Proposition 1 and Proposition 2 together show that the TI_N satisfies the same monotone convergence of its expectation as the one satisfied by the eHI, previously shown by Bronchain *et al.* [BHM+19]. Moreover, Proposition 1 tells us that the asymptotic TI_N is an upper bound of LI. It is therefore interesting to discuss whether, like in Bronchain *et al.*'s works, it is possible to get stronger notions of convergence, with the hope to get faster convergence rates than the one satisfied by eHI. Section 5 will be devoted to this question.

5 Convergence Rate of TI-Maximizing Distinguishers

So far, the metrics for a TI_N-maximizer operating on a hypothesis class $\mathcal{H}$ follow

$$\mathsf{PI}(Y;\boldsymbol{L};\mathcal{A}_{\mathcal{H}}) \leq \mathsf{LI}(Y;\boldsymbol{L};\mathcal{H}) \underset{\mathbb{E}}{\leq} \mathsf{TI}_N(Y;\boldsymbol{L};\mathcal{A}_{\mathcal{H}}) \underset{\mathbb{E}}{\leq} \mathrm{TI}_{N-1}(Y;\boldsymbol{L};\mathcal{A}_{\mathcal{H}}) \enspace ,$$

where the first inequality is unconditionally true [BHM+19], whereas the last two inequalities hold in expectation only (see Equations (12), (16)). In this section, we are interested in whether both the TI_N and the PI converge towards the quantity of interest, namely the LI. And if so, what convergence rate could we expect for the gaps between those metrics? At a very high level, the answer to both questions depends on the combination of three factors: the *richness* of the hypothesis class $\mathcal{H}$, how it is likely to depict well the true leakage model, and how *smooth* the metric we aim to optimize (*i.e.* the TI_N here) is. Depending on those factors, we may observe a *fast* convergence (*i.e.*, at a rate $\widetilde{\mathcal{O}}(1/N)$), a *slow* rate (*i.e.*, at a rate $\widetilde{\mathcal{O}}\left(1/\sqrt{N}\right)$), or no convergence at all. Which case fits to our problem? This section aims at addressing this question. To this end, we need first to formally introduce in Section 5.1 the hypothesis classes that we will consider in this paper. Then, we will have the necessary material to state in Section 5.2 the convergence rates.

5.1 Definition of our Problem

For the remaining of Section 5, we consider a hypothesis class $\mathcal{H}$ that is the family of concatenations of real-valued functions belonging to a given set $\mathcal{F}$ (that we will describe thereafter), composed with a *softmax* function

$$\sigma(\boldsymbol{x}) = \frac{1}{\sum_{i=1}^{Q} e^{\boldsymbol{x}_i}} \begin{pmatrix} e^{\boldsymbol{x}_1} \\ \vdots \\ e^{\boldsymbol{x}_Q} \end{pmatrix}, \boldsymbol{x} \in \mathbb{R}^Q \enspace . \tag{17}$$

We assume that each real-valued function $f \in \mathcal{F}$ can be fully described by a parameter vector $\boldsymbol{\theta}$. In other words, each function $\mathsf{m} \in \mathcal{H}$ can be written as

$$\mathsf{m}_{\boldsymbol{\Theta}}(\boldsymbol{l}) = \sigma \begin{pmatrix} f(\boldsymbol{l};\boldsymbol{\theta}_1) \\ \vdots \\ f(\boldsymbol{l};\boldsymbol{\theta}_Q) \end{pmatrix} \enspace , \tag{18}$$

where $\boldsymbol{\Theta}$ is the concatenation of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_Q$. We denote by $\mathcal{H}^{\intercal}$ the space $\boldsymbol{\Theta}$ belongs to.

Remark 1. The softmax function σ remains invariant by applying the same shift to all its entries. It follows that if the elementary class $\mathcal{F}$ is a group, one may fix one of the $f(\boldsymbol{l};\boldsymbol{\theta}_i)$ to the constant function 1, without changing the resulting hypothesis class $\mathcal{H}$.

This definition covers a broad family of models, such as Logistic Regression models with polynomial basis of degree k (LR_k for short) and deep neural networks, among which we particularly focus on MLP s (without loss of generality).

In the case of an LR_k-attacker, the elementary class $\mathcal{F}$ is the set of all polynomial transformations of degree at most k over the leakage space $\mathcal{L} \subset \mathbb{R}^D$. As an example, in the case of LR_1, the mapping

$$\boldsymbol{l}, \boldsymbol{\theta}_i \mapsto f(\boldsymbol{l};\boldsymbol{\theta}_i) = B_i^{\intercal}\boldsymbol{l}' \tag{19}$$

is an affine form, where $B_i \in \mathbb{R}^{D+1}$ and $\boldsymbol{l}' = (\boldsymbol{l}, 1)$. Here, $\boldsymbol{\theta}_i$ corresponds to B_i. In the case of LR_2, the mapping

$$\boldsymbol{l}, \boldsymbol{\theta}_i \mapsto f(\boldsymbol{l};\boldsymbol{\theta}_i) = \boldsymbol{l}'^{\intercal} A_i \boldsymbol{l}', \tag{20}$$

where $A_i \in \mathbb{R}^{(D+1)^2}$ is a quadratic form. Here, $\boldsymbol{\theta}_i = A_i$.
Finally, in the case of MLP s, the mapping

$$\boldsymbol{l}, \boldsymbol{\theta}_i \mapsto f(\boldsymbol{l}; \boldsymbol{\theta}_i) = \phi_L\left(\cdot; \boldsymbol{\Theta}_i^{(L)}\right) \circ \ldots \circ \phi_1\left(\cdot; \boldsymbol{\Theta}_i^{(1)}\right)(\boldsymbol{l}) \tag{21}$$

is a composition of L *layers* ϕ_i, each being the composition of a linear mapping, defined by the weight matrix $\boldsymbol{\Theta}_i^{(j)}$, with an element-wise non-linear function (a.k.a. *activation*) – except the L-th layer which is not composed with any activation function, since this role will be played by the whole softmax function. Here, $\boldsymbol{\theta}_i = (\boldsymbol{\Theta}_i^{(1)}, \ldots, \boldsymbol{\Theta}_i^{(L)})$. In the rest of the paper, we assume that the total number of entries in the weight matrices equals W.

Whereas MLPs are now widely used for profiled side-channel analysis, LR models have not been considered so far in the literature to the best of our knowledge.[2] However, LR models may be of great interest thanks to their connection to Gaussian templates. Indeed, we claim that the hypothesis class of Gaussian templates (resp., pooled Gaussian templates [CK13]) is included in LR_2 (resp., LR_1). This will be shown in Section 6. A similar correspondence could be investigated for the inclusion of so-called side-channel attacks of order k [SM16, MS16] in LR_k. We discuss in Section 6 the main difference between LR and Gaussian templates approaches, which is the nature of the underlying learning algorithm $\mathcal{A}$ used to find the right model from $\mathcal{H} = \mathsf{LR}_k$ (for $k = 1, 2$).

5.2 Convergence Rates for TI_N-Maximizers

As briefly stated in introduction of Section 5, the convergence rate of the TI_N and the PI towards the LI depends on three factors, namely the richness of $\mathcal{H}$, how it depicts well the true leakage distribution, and the smoothness of the metrics to optimize. When considering only the first and the last criteria, it is possible to prove the convergence in probability of the PI and the TI_N to the LI, with rate $\widetilde{\mathcal{O}}\left(\sqrt{\frac{P}{N}}\right)$, where P is a constant depicting the richness of $\mathcal{H}$. However, formalizing the concept of richness in this case requires some involved discussion, that the interested reader may find in Appendix B.

Instead, we propose to introduce some assumption about the second criterion, as it will allow us to derive much more intuitive, and much more efficient results. Indeed, some recent advances in statistical learning theory have seen the emergence of proofs of convergence under the so-called *central condition* [vEGM$^+$15], a rather general requirement that allows us to derive fast convergence rates. Here as well, we will not elaborate much about the exact meaning of this assumption. Instead, and for readability purpose, we provide hereafter a stronger assumption which is significantly easier to grasp.

Lemma 1 ([vEGM$^+$15, Example 2.2])**.** *Let $\mathcal{H}$ be a hypothesis class and let p be the true leakage model to be estimated. If $\mathsf{p} \in \mathcal{H}$, then the central condition holds.*

Van Erven *et al.* argue that even if $\mathsf{p} \notin \mathcal{H}$, this condition is often verified [vEGM$^+$15, Example 2.2], up to some (possibly high [MG22]) constant factors in the bounds. That is why we will assume in this section that the hypothesis of Lemma 1 holds true.

5.2.1 Fast convergence of PI towards LI

We now state the fast convergence rates for the different hypothesis classes that we consider in this section. The following corollaries 1 and 2, are proven in Appendix C.

Corollary 1. *Let LR_k for $k = 1, 2$ be a TI_N-maximizer attack using logistic regression for profiling. Suppose that*

[2] Logistic Regression models without polynomial transformation can actually be seen as the simplest MLP model, *i.e.*, without any hidden layer, nor activation layer, excepted the output softmax.

- *For all* $\boldsymbol{l} \in \mathcal{L} \subset \mathbb{R}^D$*,* $\|\boldsymbol{l}\|_2 \leq R$*, for some* $R \in \mathbb{R}$*.*
- *For all* $1 \leq i \leq Q$*,* $\|\boldsymbol{\theta}_i\|_2 \leq S$*, for some* $S \in \mathbb{R}$*.*

If LR_k *verifies the assumption of Lemma 1, and* $N \geq 5$*, the gap* $\mathsf{LI} - \mathsf{PI}$ *is bounded by*

$$\frac{8}{N}\left(2(R^2+1)^{k/2}S + \log(Q)\right)\left((D+1)^k Qh + \log\left(\frac{1}{\delta}\right)\right) + \frac{1}{N} \quad . \tag{22}$$

with $h = \log\left(32QSN(R^2+1)^{k/2}\right)$*.*

If $\mathsf{p} \in \mathsf{LR}_k$ (for $k = 1$ or $k = 2$), then $\mathsf{LI}(Y;\boldsymbol{L};\mathsf{LR}_k) = \mathsf{MI}(Y;\boldsymbol{L})$. In other words, the regret of an LR_k attacker is bounded by $\widetilde{\mathcal{O}}\left(\frac{D^kQ}{N}\right)$ if we assume that every real parameter and every leakage value is bounded by a constant.

Corollary 2. *Let* $\mathcal{A}$ *be a* TI_N*-maximizer attacker using* MLP *as defined in Equation 21 with ReLU activation function for profiling. Suppose that*

- *For all* $\boldsymbol{l} \in \mathcal{L} \subset \mathbb{R}^D$*,* $\|\boldsymbol{l}\|_2 \leq R$*, for some* $R \in \mathbb{R}$*.*
- *For all* $1 \leq i \leq L$ *and for all* $1 \leq j \leq Q$*,* $\left\|\boldsymbol{\Theta}_i^{(j)}\right\|_F \leq S$*, for some* $S \in \mathbb{R}_{\geq 1}$*.*

If MLP *verifies the assumption of Lemma 1, and* $N \geq 5$*,*

$$\mathsf{LI}(Y;\boldsymbol{L};\mathsf{MLP}) - \mathsf{PI}(Y;\boldsymbol{L};\mathcal{A}_{\mathsf{MLP}}) \leq \frac{8B}{N}\left(WQ\log(16BN) + \log\left(\frac{1}{\delta}\right)\right) + \frac{1}{N} \quad , \tag{23}$$

where $B = 2Q^{3/2}RLS^{L+1}$*.*

If $\mathsf{p} \in \mathsf{MLP}$, then $\mathsf{LI}(Y;\boldsymbol{L};\mathsf{MLP}) = \mathsf{MI}(Y;\boldsymbol{L})$. In other words, the regret of an MLP attacker is bounded by $\widetilde{\mathcal{O}}\left(\frac{LW^{2L+3}DQ^{5/2}}{N}\right)$ if we assume that every real parameter and every leakage value is bounded by a constant.

5.2.2 Fast convergence of TI_N towards LI

So far we have shown that under the central condition (Lemma 1) — in other words under the assumption that $\mathsf{LI} = \mathsf{MI}$ — the regret of a TI_N-maximizer, *i.e.* the gap between the MI and the PI enjoys a fast convergence rate with high probability towards 0. Since we have shown in Section 4 that for this learning algorithm, the TI_N is monotonically decreasing and converges to the LI, we may wonder what is its convergence rate. We show in Appendix B that the TI_N converges in probability towards the LI at a rate $\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{N}}\right)$, and a faster convergence rate cannot hold in general. To see why, let us take a counter-example in which the hypothesis class $\mathcal{H}$ contains only the true leakage model p, so we trivially have the equality $\mathsf{PI} = \mathsf{LI} = \mathsf{MI}$. Yet, since $\mathcal{H}$ is a singleton, the TI_N-maximizer is constant, so the TI_N can be expressed as an empirical mean. According to the well-known central limit theorem, the rate of convergence in probability cannot be faster than $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$.

Nevertheless, the latter theoretical counter-example does not reflect what an evaluator can observe in practice. Indeed, the slow convergence rate comes from the variance in the TI_N: its deviation converges slowly (as a consequence of the central limit theorem), regardless of whether the TI_N-maximizer is good or not. On the other hand, similarly to the conclusion of Section 3, the gap between the TI_N and the LI is dominated by its statistical bias, which converges towards 0 at a *fast* rate. More precisely, Proposition 3 (Appendix C) analyzes the training gap

$$\mathsf{TG}_N(Y;\boldsymbol{L};\mathcal{A}_{\mathcal{H}}) = \mathsf{TI}_N(Y;\boldsymbol{L};\mathcal{A}_{\mathcal{H}}) - \mathsf{PI}(Y;\boldsymbol{L};\mathcal{A}_{\mathcal{H}})$$

and shows that

$$\mathop{\mathbb{E}}_{\mathcal{S}_N}\left[\mathsf{TG}_N(Y;\boldsymbol{L};\mathcal{A}_{\mathcal{H}})\right] \in \mathcal{O}\left(\frac{h}{N}\right)$$

where h depends on the richness of the hypothesis class $\mathcal{H}$. Proposition 3 also bounds the deviation of the training gap:

$$\mathop{\mathbb{E}}_{\mathcal{S}_N}\left[\left|\mathsf{TG}_N(Y;\boldsymbol{L};\mathcal{A}_{\mathcal{H}}) - \mathop{\mathbb{E}}_{\mathcal{S}_N}\left[\mathsf{TG}_N(Y;\boldsymbol{L};\mathcal{A}_{\mathcal{H}})\right]\right|\right] \in \mathcal{O}\left(\frac{h}{N} + \frac{1}{\sqrt{N}}\right) .$$

In most practical cases, similarly to Section 3, we observe that $h \gg N$, hence the dominant term in the deviation is proportional to the bias.

The overall picture. To summarize, combining the results of Section 4.2, Equation 4.2, and this section, we come to the following picture for the TI_N-maximizer regarding the convergence w.r.t N:

$$\begin{aligned} \mathsf{LI}(Y;\boldsymbol{L};\mathcal{H}) - \widetilde{\mathcal{O}}\left(\frac{1}{N}\right) \underset{\text{h.p.}}{\leq} \mathsf{PI}(Y;\boldsymbol{L};\mathcal{A}_{\mathcal{H}}) \ &\leq \mathsf{LI}(Y;\boldsymbol{L};\mathcal{H}) \\ \mathsf{LI}(Y;\boldsymbol{L};\mathcal{H}) \underset{\mathbb{E}}{\leq} \mathsf{TI}_N(Y;\boldsymbol{L};\mathcal{A}_{\mathcal{H}}) &\underset{\mathbb{E}}{\leq} \mathsf{LI}(Y;\boldsymbol{L};\mathcal{H}) + \widetilde{\mathcal{O}}\left(\frac{1}{N}\right) \ , \end{aligned}$$

where $\underset{\text{h.p.}}{\leq}$ denotes an inequality that holds with high probability, and $\underset{\mathbb{E}}{\leq}$ denotes an inequality verified by the expectations of both hand-sides.

6 Gaussian Templates

The assumption $\mathsf{p} \in \mathcal{H}$, which is key to obtain the fast convergence rate of the previous section, is actually a fairly common assumption made in side-channel security evaluations. One of the most popular models is the Gaussian template where $\mathcal{H}$ is the set of multivariate Gaussian distributions.[3] The Gaussian template attack (gTA for short), however, is not a TI_N maximizer, since the parameters (mean and covariance) of the templates are chosen as the empirical average and covariance, raising the question whether we can still derive similar bounds to what has been done in Section 5? In this section, we compute the convergence rates of gTA, first for the original and most generic template attack [CRR02], then in the particular case where the covariance matrix is known to be diagonal — a.k.a. the so-called *naive Bayes* classifier [PHG17, PSK$^+$18] — and finally for the pooled gTA (*i.e.* the covariance is the same for all values of y) [CK13]. Formally, we assume that the leakage distribution $\mathsf{f}_y(\cdot)$ for each of the Q different classes y has a Gaussian distribution of mean μ_y and covariance Σ_y. For each class y, the adversary estimates a D-dimensional Gaussian generative model $\widehat{\mathsf{f}}_y(\cdot)$ (the template) according to the empirical mean vector $\widehat{\mu}_y$ and the empirical covariance matrix $\widehat{\Sigma}_y$. Without loss of generality, we assume that for each class, the adversary has acquired N/Q traces during the profiling phase in order to build each template $\widehat{\mathsf{f}}_y(\cdot)$. The discriminative model derived from this Gaussian model — computed thanks to the Bayes rule — is used to mount a key recovery attack.

One may then remark that LR_2 covers the set of discriminative models derived from gTA. To see this, define each elementary function $f(\boldsymbol{l};\boldsymbol{\theta}_i) = -\frac{1}{2}(\boldsymbol{l}-\mu_i)^\intercal\Sigma_i^{-1}(\boldsymbol{l}-\mu_i) = \boldsymbol{l}'^\intercal A_i \boldsymbol{l}'$ for some $A_i \in \mathbb{R}^{(D+1)^2}$. Thus, the corresponding LR_2 model $\mathsf{m}_{\boldsymbol{\Theta}}$ coincides with the Gaussian template. Likewise, if we further assume that the covariance matrix is the same for all

[3] Other popular generative models used in the side-channel literature are restricted classes of Gaussian templates (*e.g.*, Schindler's stochastic model [SLP05]), Gaussian templates after pre-processing (*e.g.*, Linear discriminant analysis [APSQ06]) or generalizations (*e.g.*, Gaussian mixtures [LP07]).

classes, the quadratic term $-\frac{1}{2}\boldsymbol{l}^{\intercal}\Sigma_i^{-1}\boldsymbol{l}$ is common to all functions $f(\boldsymbol{l};\boldsymbol{\theta}_i)$ and can be subtracted without change to the model $\mathsf{m}_{\boldsymbol{\Theta}}$. We deduce that the set of *pooled* Gaussian templates is equal to the hypothesis class of LR_1.[4] In other words, despite a gTA (resp., p-gTA) adversary differs from an LR_2 (resp., LR_1) adversary, since they do not use the same learning algorithm, the hypothesis class of the former one lies in the hypothesis class of the latter one. It is therefore interesting to compare their convergence rates, *e.g.* by comparing their respective regrets (*i.e.*, the gap between the LI and the PI since it follows from the Gaussian assumption that $\mathsf{LI} = \mathsf{MI}$). This is the aim of this section.

Remark 2. The Gaussian TA (resp., pooled TA) is identical to the quadratic (resp., linear) discriminant analysis (QDA/LDA), which are well-known machine learning models. However, most of the literature focuses on the success rate metric (*e.g.* [Efr75, HTF09]), and is not directly adaptable to information theoretic metrics. To the best of our knowledge, there is no existing bound on the convergence of the LDA/QDA that applie to the PI.

6.1 gTA convergence

Let us start with a convergence bound for the gTA, which is the most general Gaussian templates model. The proof of the following corollary is given in Section D.1.

Corollary 3. *For any $\delta > 0$, the regret $\mathsf{R}\left(\mathit{gTA}\right)$ of an attacker instantiating a Gaussian template attack is upper-bounded by $\mathcal{O}\left(\frac{QD^2}{N}\log\left(\frac{1}{\delta}\right)\right)$ with probability at least $1-\delta$.*

In other words, to be able to control the estimation error of the MI when profiling with a gTA, the attacker/evaluator must ensure that the number of profiling traces scales with the squared dimensionality of the traces times the number of classes.

6.2 On the tightness of the bound

So far, we have emphasized an upper bound of the regret of a gTA attacker. It is then interesting to assess whether this upper bound is tight or not. Namely, can we derive tighter bounds of our regret, for any actual multivariate Gaussian leakage? We argue that without further assumption regarding the knowledge of the attacker, we cannot get better bounds. The convergence rate emphasized in Corollary 3 essentially comes from the error terms due to the estimation of the empirical covariance matrix, namely $\log\left(\det\left(\widehat{\Sigma}\right)\right)$ and $\mathrm{Tr}\left(\widehat{\Sigma}^{-1}\right) - D$. However, the sum of both error terms scale with $\Theta\left(\frac{QD^2}{N}\right)$ in expectation (the proof is given in Section D.1.1). Despite this negative argument, it is still possible to obtain faster convergence, provided that the attacker has more prior knowledge concerning the leakage, and more particularly concerning the shape of the covariance matrix. We next emphasize two particular cases that are often considered in side-channel analysis.

6.2.1 The Covariance Matrix is Diagonal: Naive Bayes

The Naive Bayes model has sometimes been used in SCA [PHG17, PSK+18]. It assumes a Gaussian multivariate distribution with diagonal covariance matrix for the leakage function. This reduces the covariance estimation to the estimation of the variance in each dimension, leading to a faster convergence, as stated by the next corollary, proven in Section D.2.

Corollary 4. *The regret of an attacker instantiating a Gaussian template attack knowing that the covariance matrices are all diagonal is upper-bounded by $\mathcal{O}\left(\frac{QD}{N}\log\left(\frac{1}{\delta}\right)\right)$.*

[4] Even though the hypothesis classes of p-gTA and LR_1 are the same, the LR_1 model is more general (due to its different training). Indeed, Efron argues that the model LR_1 could coincide with the template attacks with exponential family distribution sets, with common nuisance parameter [Efr75].

6.2.2 Choudary and Kuhn's Pooled Template Attacks.

For gTA-based side-channel attacks, the bottleneck task is the estimation of the covariance matrices. Choudary and Kuhn considered this problem at Cardis'13 and emphasized that if $N/Q \leq D$, the empirical covariance matrices admit some zero singular values, so they are not invertible [CK13]. To circumvent this numerical issue, they proposed to pool all the covariance matrices into one common matrix for all the classes, leading to the pooled Gaussian templates attack (p-gTA). This assumption is also known under the name of *homoscedasticity* and it leads to mounting a *Linear Discriminant Analysis* (LDA) classification under the statistical learning terminology. Despite its popular success in SCA [SA08, LPB+15, CDP15, CDP16, BS20], less has been done regarding the analysis of this approach since Choudary and Kuhn's paper. Yet, using a p-gTA addresses the necessary condition emphasized by Choudary and Kuhn so that the attack works, but does not ensure any sufficient condition. Can we find another explanation to the success of p-gTA? At first glance, using Q times more traces to estimate the pooled covariance matrix would induce a $\mathcal{O}(D^2/N)$ convergence for the estimation of the covariance, while keeping $\mathcal{O}(QD/N)$ convergence for the means estimation. This would result in a $\mathcal{O}(\max\{D^2/N, QD/N\})$ bound in Corollary 3 for the ultimate regret of pooled template attacks. However, we conjecture that the latter upper bound can even be tightened to $\mathcal{O}(QD/N)$, becoming fully linear in the trace dimensionality, despite the D^2 matrix coefficients to estimate. Our conjecture is grounded on the similarity with the LR_1 model and on a proof in the particular case where $Q = 2$, stated next and proven in Section D.3.

Corollary 5. *The regret of an attacker instantiating p-gTA for $Q = 2$, is upper bounded by $\mathcal{O}\left(\left(\Delta^2 + 1\right) \frac{D+1}{N}\right)$ where $\Delta^2 = \left(\mu_1 - \mu_0\right)^\intercal \Sigma^{-1} \left(\mu_1 - \mu_0\right)$ denotes the Mahalanobis distance between the two centroids.*

7 Case Study and Practical Use

So far, we have studied the PI and TI_N for different classes of models. We finally discuss the impact of these results for the SCA practitioner. First, we briefly explain in Section 7.1 how the theoretical bounds could be used by an evaluator. Then, we illustrate in Section 7.2 our bounds and their use on simulated and experimental data.

7.1 Discussion on the practical use

Let us illustrate the properties of the TI_N and discuss its practical usage in a side-channel evaluation context. Suppose that an evaluator has a target security level claim to verify, *e.g.*, expressed in bits leaked per trace.[5] If an evaluator wants to verify this claim, she can run a profiling with a TI maximizer as a learning algorithm. Figure 4 sketches the different situations that an evaluator may face after acquiring a profiling dataset (with a given amount of traces) and a validation dataset, then running the attack.

In the first case (left of the figure), the PI is higher than the target security level. Therefore, the evaluator can conclude that the device under evaluation does not satisfy the security requirement. Furthermore, the gap between the PI and the TI captures the potential improvement of the attack that beats the target security level.

In the third case (right of the figure), the opposite situation holds. The TI is below the target and measures the guaranteed security level. Furthermore, the gap between the PI and the TI captures the potential improvement of the guaranteed security level. It is remarkable that this conclusion holds even if the PI of the model trained by the evaluator is negative, that is, independently of whether this model is useful to mount an attack.

[5] Which can be converted into a success rate using bounds like [DFS15, dCGRP19].

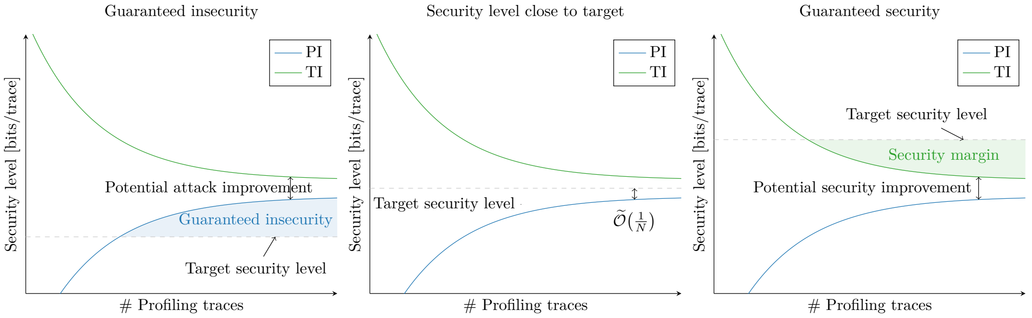


Figure 4: Illustration of security evaluations results.

In the remaining case (middle plot), the target security level lies between the PI and the TI, for the given amount of profiling traces. While it is in general less conclusive, our tools also allow interesting statements in this case. Indeed, we know that the actual security level is also between the PI and the TI. Let us denote the target security level by T and let $\varepsilon = \mathsf{TI}_N - \mathsf{PI}$. We then know that the actual security level belongs to the interval $[\mathsf{PI}, \mathsf{TI}_N] \subset [T - \varepsilon, T + \varepsilon]$. Let us moreover assume that $\varepsilon \leq \alpha T$ for some α chosen by the evaluator. We can then claim that the security level of the implementation belongs to $[(1-\alpha)T, (1+\alpha)T]$, i.e., T with an error margin of $(\alpha/100)\%$. This brings us to the relevance of knowing the convergence rate of the PI and the TI_N. Indeed, this approach is practical only if the evaluator can easily make ε small. Thanks to the bounds given in Section 5 and Section 6, this requirement is satisfied: ε converges at a fast $\mathcal{O}\left(\frac{1}{N}\right)$ rate, where N is the number of profiling traces. Moreover, our quantitative bounds in these sections (see, e.g., Corollary 2) show that the constants behind the $\mathcal{O}(\cdot)$ notation are reasonably small. Therefore, a practical use for the convergence rates is to extrapolate the guarantees that can be obtained with a number of profiling traces: from a given target security level T and an uncertainty α, the evaluator can have a bound on the number of profiling traces she will need to conclude her experiments with confidence.

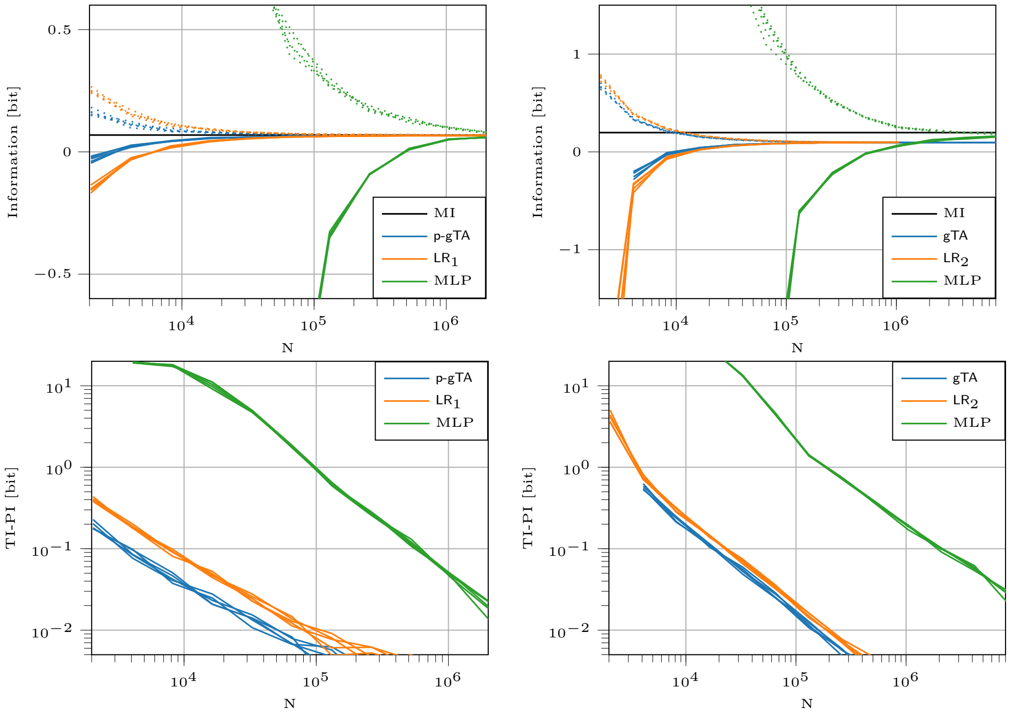
7.2 Illustration on simulated & experimental data

It now remains to illustrate our bounds with concrete data. For this purpose, we consider both simulated leakages and a public dataset of real measurements.

7.2.1 Setup & Models

Simulation setup. For our simulated experiments, we consider the Hamming weight leakage of an 8-bit secret in two settings. The first one (denoted as "hardware") corresponds to a typical hardware implementation: no masking and low SNR. The second one (denoted as "software") corresponds to a protected software implementation: 2-shares Boolean masking and high SNR (each share independently leaking its Hamming weight). These simulations have 1 and 2 points in the leakage traces and the noise is Gaussian.

Public dataset. For the experimental validation, we take Bhasin *et al.*'s AES-HD dataset in its extended version [BJP20], which is an unprotected AES implemented on FPGA. The dataset is made of 500,000 traces of 1250 time samples, of which 450,000 traces are used for the training, *i.e.*, maximizing the TI, whereas the remaining is used for validation, *i.e.*, estimating the PI. The target intermediate value is the first byte of the AES state before the AddRoundkey operation of the last round, for which the full dataset exhibits an SNR peak up to 0.016 [ZBHV20, Fig. 18]. Since the last AES round is clearly identifiable on



(a) Hardware: not masked, SNR=0.1

(b) Software: masked, SNR=10.

Figure 5: Convergence of information metrics. In the upper part of the figure, the dotted lines represent the TI while the solid lines represent the PI.

the raw traces, we assume the evaluator/adversary to be able to restrict its target window over 100 Points of Interest (PoIs) around the SNR peak.

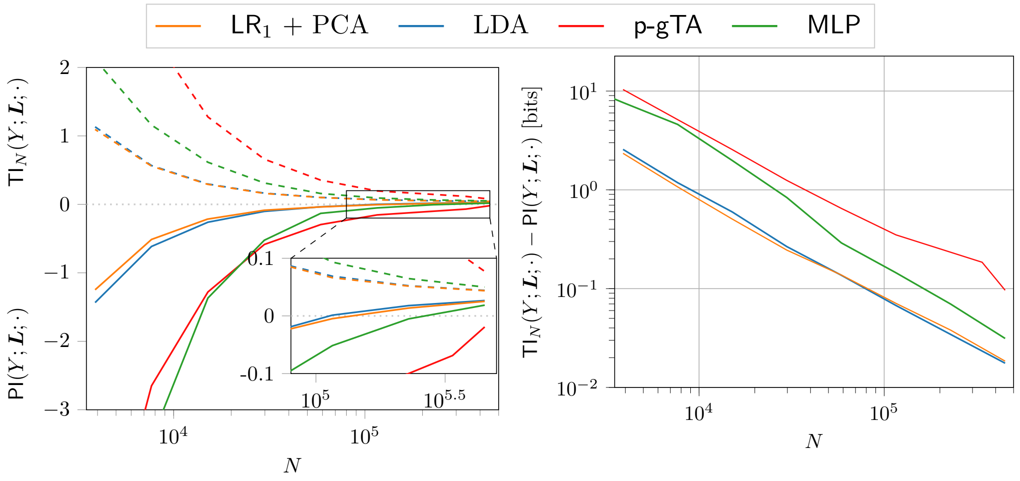
Models. In the "hardware" setting, we evaluate the linear models: LR_1 and p-gTA, as well as an MLP (single hidden layer with 100 neurons in the simulations, and 10 neurons in the experiments). The p-gTA is done using the LDA from SCALib[6], and we also consider (for the experimental dataset) a variant of the p-gTA with reduction to a 10-dimensional linear subspace (also known as LDA [SA08]). The logistic regression is done with the implementation in scikit-learn[7], and for the experimental dataset, we apply a Principal Component Analysis (PCA) to reduce it to 20 dimensions, which simplifies the optimization [APSQ06]. The TI maximization of the MLP is done thanks to the Adam optimizer [KB15] implemented on the Pytorch framework [PGM+19] with a 10^{-4} learning rate, without weight decay and a full batch, for $10,000$ epochs (*i.e.*, a high number, in order to best maximize the TI). In the "software" setting, the leakage function is non-linear, we evaluate the LR_2, gTA and MLP models (with the same hyper-parameters).

7.2.2 Results

The TI_N and PI of these models for varying number of training traces are shown in Figure 5 for the simulations (the training is repeated for 5 different training sets) and on Figure 6 for the experiments on AES-HD. Additionally for the simulations, since the true distribution is known, the MI is also shown. These figures lead to the following observations.

[6] scalib.readthedocs.io/.

[7] https://scikit-learn.org/.



(a) Learning curves. **(b)** Gap trend vs. profiling complexity.

Figure 6: Experiments on AES-HD.

In the upper part of Figure 5, we see that the variance of the TI_N is quite small compared to its bias (w.r.t. the LI). This is a consequence of the $\log(\frac{1}{\delta})$ terms in Corollaries 1, 2 and 3.[8] Next, considering the lower part of Figure 5 which depicts the gap between the TI and the PI, we see that the slope in the logarithmic plots is close to -1, which means that the gap is inversely proportional to N, as proven Section 5 and Section 6.[9] Interestingly, this holds true even when the PI and the TI are not yet close to their limit, and over a wide range of training set sizes (more than two orders of magnitudes), which confirms the practical interest of the extrapolation proposed in Section 7.1.

The same observation can be made on Figure 6b, depicting the gap between the TI and the PI on the AES-HD dataset. So concretely, an evaluator could estimate how many traces are needed for her profiling from the beginning of a learning curve (*i.e.*., when reaching the linear regime), which we illustrate with a concrete example. If the evaluator (who does not know the MI) wants to assess whether the target leaks less than 0.1 bit/trace when profiled with a linear model such as the p-gTA or the LR$_1$, Figure 6a tells us that the she can stop the acquisition campaign and conclude after 100,000 traces. Furthermore, she can estimate this number with a much smaller dataset of $\approx 10,000$ traces, by extrapolating the gap $\varepsilon = \mathsf{TI}_N - \mathsf{PI}$, knowing that it is inversely proportional to N.

We finally remark that for the software simulation, the MLP model has a higher LI than the LR$_2$ and gTA models, meaning that it better models the true distribution. This increased versatility comes at a cost: training it requires at least two orders of magnitude more traces than the simpler models (roughly matching the bounds given in Table 1).

8 Concluding Remarks

This paper provides new information theoretic metrics and bounds together with a study of the convergence rates for practically-relevant profiled attacks. Besides their interest for helping side-channel security evaluators in selecting the profiling tools that best match their target device and time constraints, our results also show connections and differences between statistical learning theory and side-channel analysis. For example, in order to

[8] The hypothesis $\mathsf{p} \in \mathcal{H}$ is not satisfied for the gTA applied to a Gaussian mixture. Hence Corollary 3 does not apply, but its conclusion seems to hold here. The p-gTA results are in line with our conjecture.

[9] For the gTA and LR$_2$, $\mathsf{p} \in \mathcal{H}$ does not hold, but convergence is still in $1/N$, as discussed in Section 5.

obtain convergence rates, we observed that the evaluator's goal, namely maximizing the PI to estimate the highest lower bound on MI, could be rephrased as a machine learning problem, using information theoretic metrics as loss functions. Accordingly, the TI_N metric is nothing but the *empirical risk* studied in learning theory, and the TI_N-maximizer in the profiling SCA view coincides with the *Empirical Risk Minimizer* (ERM), one of the most studied algorithms in machine learning. Yet, and somewhat surprisingly, the IT metrics that are most relevant for side-channel security evaluations are less investigated optimization goals than security metrics (like the accuracy) in the machine learning literature. So our study puts forward both the interest of leveraging the broad scope of theoretical results established in statistical learning theory over the past few years, and the need to adapt them to needs that are somewhat specific to security evaluations. Eventually, an interesting meta-conclusion of our results is that the profiling data complexity to estimate a model does not fundamentally differ from the attack data complexity using this model, since the profiling error we need to reach is proportional to the security level. This motivates shortcut approaches to profiling as proposed in [ABB$^+$20], and suggests that making security claims based on the profiling complexity of an implementation (*i.e.*, contradicting the relevance of such shortcuts) could only be sound if showing that the model estimation problem is computationally hard, which is an interesting open problem.

Acknowledgments

Gaëtan Cassiers and François-Xavier Standaert are respectively Research Fellow and Senior Associate Researcher of the Belgian Fund for Scientific Research (FNRS-F.R.S.). This work and its presentation were funded in parts by the ERC consolidator grant 724725 (SWORD) and the ERC Advanced Grant 101096871 (BRIDGE).

A Proofs of Section 3

Proof of Theorem 1. It is worth reminding that the left inequality of Equation 4 has already been shown by Bronchain *et al.* [BHM$^+$19, Thm. 5]. Nevertheless, we provide here a simpler alternative proof, by taking inspiration from the work of Paninski [Pan03, Prop. 1] with slight modifications adapted to our context, thereby showing the right inequality. First, we note that the eHI can be restated as follows:

$$\begin{aligned}
\mathsf{eHI}_N(Y;\boldsymbol{L}) \quad &= \quad \mathsf{MI}(Y;\boldsymbol{L}) && (24)\\
&+ \quad \sum_{y,\boldsymbol{l}} (\tilde{\mathsf{e}}_N(y,\boldsymbol{l}) - \mathsf{p}(y,\boldsymbol{l}))\log_2(\mathsf{p}(y \mid \boldsymbol{l})) && (25)\\
&+ \quad \sum_{\boldsymbol{l}} \tilde{\mathsf{e}}_N(\boldsymbol{l})\mathsf{D}_{\mathsf{KL}}(\tilde{\mathsf{e}}_N(\cdot|\boldsymbol{l}) \parallel \mathsf{p}(\cdot \mid \boldsymbol{l})) \ , && (26)
\end{aligned}$$

where $\mathsf{D}_{\mathsf{KL}}(\cdot \parallel \cdot)$ denotes the KL divergence. This re-statement is of great interest, since the first sum is unbiased – since $\tilde{\mathsf{e}}_N(y,\boldsymbol{l})$ admits $\mathsf{p}(y,\boldsymbol{l})$ as expected value – whereas the second sum is positively biased – because each of its term are positive thanks to the KL divergence. Hence the first inequality of Equation 4.

It now remains to upper bound the second sum in expectation in order to get the upper bound on the bias of eHI. To this end, as suggested by Paninski [Pan03, Proposition 1], we use the fact that

$$0 \le \mathop{\mathbb{E}}_{\tilde{\mathsf{e}}_N}\left[\mathsf{D}_{\mathsf{KL}}(\tilde{\mathsf{e}}_N(\cdot|\boldsymbol{l}) \parallel \mathsf{p}(\cdot \mid \boldsymbol{l}))\right] \le \log\left(1 + \frac{Q-1}{N}\right) \ . \tag{27}$$

Finally, we have

$$\begin{aligned}\mathbb{E}\left[\mathsf{eHI}_N - \mathsf{MI}\right] &= \sum_{\boldsymbol{l}} \mathop{\mathbb{E}}_{\tilde{\mathsf{e}}_N}\left[\tilde{\mathsf{e}}_N(\boldsymbol{l}) \cdot \mathsf{D}_{\mathsf{KL}}(\tilde{\mathsf{e}}_N(\cdot|\boldsymbol{l}) \parallel \mathsf{p}(\cdot \mid \boldsymbol{l}))\right] \\ &\leq \sum_{\boldsymbol{l}} \mathop{\mathbb{E}}_{\tilde{\mathsf{e}}_N}\left[\mathsf{D}_{\mathsf{KL}}(\tilde{\mathsf{e}}_N(\cdot|\boldsymbol{l}) \parallel \mathsf{p}(\cdot \mid \boldsymbol{l}))\right] \\ &\leq |\mathcal{L}| \log\left(1 + \frac{Q-1}{N}\right) \\ &\leq |\mathcal{L}| \frac{Q-1}{N} \quad .\end{aligned}$$

We conclude the proof by observing that $|\mathcal{L}|$ is the number of bins. In addition, Equation 5 is a direct consequence of [Pan03, Thm. 5]. □

Proof of Theorem 2. Notice that

$$\mathsf{eHI}_N = \mathsf{H}(Y) + \widehat{\mathsf{H}(\boldsymbol{L})} - \widehat{\mathsf{H}(Y, \boldsymbol{L})} \quad , \tag{28}$$

where $\widehat{\mathsf{H}(\boldsymbol{L})} = -\sum_{\boldsymbol{l} \in \mathcal{L}} \tilde{\mathsf{e}}_N(\boldsymbol{l}) \log(\tilde{\mathsf{e}}_N(\boldsymbol{l}))$, and likewise for $\widehat{\mathsf{H}(Y, \boldsymbol{L})}$. Subtracting the expected value of the eHI, we get

$$\left|\mathsf{eHI}_N - \mathbb{E}\left[\mathsf{eHI}_N\right]\right| \leq \left|\widehat{\mathsf{H}(\boldsymbol{L})} - \mathbb{E}\left[\widehat{\mathsf{H}(\boldsymbol{L})}\right]\right| + \left|\widehat{\mathsf{H}(Y, \boldsymbol{L})} - \mathbb{E}\left[\widehat{\mathsf{H}(Y, \boldsymbol{L})}\right]\right| \quad . \tag{29}$$

Now, using McDiarmid's inequality [AK01, Thm. 1], we have that for all $\epsilon > 0$

$$\Pr\left(\left|\widehat{\mathsf{H}(\boldsymbol{L})} - \mathbb{E}\left[\widehat{\mathsf{H}(\boldsymbol{L})}\right]\right| > \frac{\epsilon}{2}\right) \leq 2 \exp\left(-\frac{\epsilon^2 N}{8 \log_2(N)^2}\right) \quad . \tag{30}$$

Likewise, the very same inequality holds to upper bound $\left|\widehat{\mathsf{H}(Y, \boldsymbol{L})} - \mathbb{E}\left[\widehat{\mathsf{H}(Y, \boldsymbol{L})}\right]\right|$. Hence, for all $\epsilon > 0$

$$\Pr\left(\left|\mathsf{eHI}_N - \mathbb{E}\left[\mathsf{eHI}_N\right]\right| > \epsilon\right) \leq 4 \exp\left(-\frac{\epsilon^2 N}{8 \log_2(N)^2}\right) \quad . \tag{31}$$

Denoting by δ the right hand-side of Equation 31, we get the main result.

Finally, the property

$$\left|\mathsf{eHI}_N - \mathbb{E}\left[\mathsf{eHI}_N\right]\right| \in \Theta\left(\frac{1}{\sqrt{N}}\right)$$

is proven in [AK01] (Section 4.1). □

A.0.1 On the Effect of Discretization.

It is worth emphasizing that the latter analysis has been done assuming discrete probability distributions for the leakage. Thereby, one may wonder whether those results extend to the case where the leakage is modeled by continuous probability distributions. At first sight, the latter result would become useless, as it would imply the oscilloscope resolution ω to tend towards infinity. Unfortunately, it is hardly likely to obtain tight convergence bounds in this case, because of the so-called *curse of dimensionality*, which – informally – states that the convergence rate of non-parametric density estimation methods would slow down at least exponentially with D [Sto82, Sto83]. Moreover, with nonparametric density estimation methods, there is a risk that, depending on the choice of the kernel, the HI no longer upper-bound the MI.

B Proofs of Section B.2

B.1 Characterizing the Complexity of $\mathcal{H}$: the Pseudo-Dimension

In the next section, we will present several upper bounds on the TI_N towards the LI. It is expected that those bounds will depend on the *complexity* – or the *richness* – of the underlying hypothesis class $\mathcal{H}$. Intuitively, the more parameters in $\boldsymbol{\Theta}$ to fit, the slower the convergence. It turns out that it is possible to characterize this complexity. This characterization, named *Pseudo-Dimension*, is defined in this section, and we provide some examples of pseudo-dimensions for several classes of interest for this study. We will therefore be able to provide some convergence rates in the next sections that depend on the pseudo-dimension.

We first need an intermediate definition of a *pseudo-shattering*.

Definition 7 (Pseudo-shattering [AB02, Def. 11.1])**.** Let $\mathcal{F}$ be a set of functions mapping from a domain $\mathcal{L}$ to $\mathbb{R}$ and suppose that $\mathcal{S}_N = \{\boldsymbol{l}_1, \dots, \boldsymbol{l}_N\} \subset \mathcal{L}$ for some positive integer N. Then, $\mathcal{S}_N$ is *pseudo-shattered* by $\mathcal{F}$ if there are real numbers $r_1, \dots, r_N$ such that for all $\boldsymbol{b} \in \{0,1\}^N$ there is a function $f_{\boldsymbol{b}} \in \mathcal{F}$ such that for all $1 \leq i \leq N$,

$$f_{\boldsymbol{b}}(\boldsymbol{l}_i) \begin{cases} \leq r_i \text{ if } \boldsymbol{b}_i = 0 \\ > r_i \text{ if } \boldsymbol{b}_i = 1 \end{cases} . \tag{32}$$

We say that $r = (r_1, \dots, r_N)$ *witnesses* the shattering.

An example of pseudo-shattering is depicted in Figure 7. We consider $\mathcal{F}$ as the set of affine functions in $\mathbb{R}$. When $\mathcal{S}_N = \{\boldsymbol{l}_1, \boldsymbol{l}_2\}$, we can exhibit a function from $\mathcal{F}$ satisfying Equation 32 for any 2-bit vector $\boldsymbol{b} \in \{0,1\}^2$. However, we can notice that when adding $\boldsymbol{l}_3$ to $\mathcal{S}_N$, the new profiling set cannot be shattered anymore, since the binary vector $\boldsymbol{b} = (0,0,1)$ provides a counter-example where Equation 32 is not satisfied. It can be verified that no matter the choice of r_3, one will always find such a binary vector $\boldsymbol{b}$ breaking the condition of Equation 32. Intuitively, this states that $\mathcal{F}$ is not *rich* enough to shatter any set of 3 leakages or more. Hence the choice of quantifying the richness of $\mathcal{F}$ by the maximum amount of leakages that can be shattered by $\mathcal{F}$, as formalized hereafter.

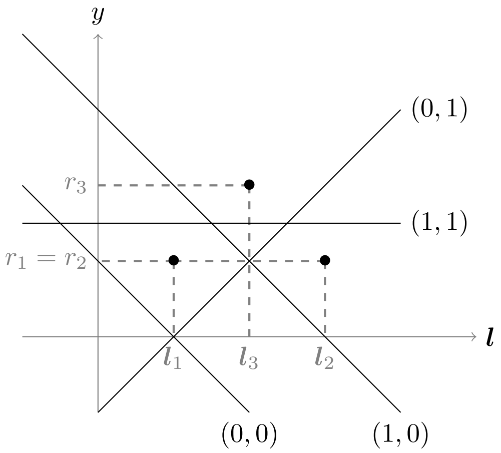


Figure 7: Illustration of the pseudo-shattering by the set $\mathcal{F}$ of affine functions of $\mathcal{L} = \mathbb{R}$. The tuples denote the different values of $\boldsymbol{b}$. $\{\boldsymbol{l}_1, \boldsymbol{l}_2\}$ is pseudo-shattered by $\mathcal{F}$, while $\{\boldsymbol{l}_1, \boldsymbol{l}_2, \boldsymbol{l}_3\}$ is not.

Definition 8 (Pseudo-dimension [AB02, Def. 11.2])**.** Suppose that $\mathcal{F}$ is a set of functions from a domain $\mathcal{L}$ to $\mathbb{R}$. Then, $\mathcal{F}$ has *pseudo-dimension* N if N is the largest integer such that any subset $\mathcal{S}_N$ of $\mathcal{L}$ of cardinality N is pseudo-shattered by $\mathcal{F}$. If no such maximum exists, we say that $\mathcal{F}$ has infinite pseudo-dimension. The pseudo-dimension of $\mathcal{F}$ is denoted $\mathsf{P}_{\mathsf{dim}}(\mathcal{F})$.

As an example, it is known that if $\mathcal{F}$ is a finite dimensionality vector space of functions from an input space $\mathcal{L}$ onto $\mathbb{R}$, then $\mathsf{P}_{\mathsf{dim}}(\mathcal{F})$ is the dimensionality of $\mathcal{F}$ [AB02, Thm. 11.4]. We give hereafter the pseudo-dimension of the two classes considered in this work, namely the Logistic regression and the MLP.

Theorem 3 (Pseudo-dimension of LR_k [AB02, Thm. 11.8])**.** *Let $\mathcal{F}$ be the class of all polynomial transformations on $\mathbb{R}^D$ of degree at most k. Then*

$$\mathsf{P}_{\mathsf{dim}}(\mathcal{F}) = \binom{D+k}{k} \quad . \tag{33}$$

Theorem 4 (Pseudo-dimension of MLP [BHLM19])**.** *Let $\mathcal{F}$ be the class of* MLP *with real-valued output with piece-wise linear activation function, W parameters and L layers. Then, there exists two constants $c > 0, C > 0$ such that*

$$cWL\log(W/L) \leq \mathsf{P}_{\mathsf{dim}}(\mathcal{F}) \leq CWL\log(W) \quad . \tag{34}$$

Put in another way, this means that the pseudo-dimension of parametric models is roughly proportional to the number of real-valued parameters to fit.[10]

B.2 Convergence Rate for TI Maximizers

We are now ready to present our main result for TI_N maximizers.

Theorem 5. *Let $\mathcal{H}$ be a hypothesis class to model the leakage of an intermediate computation of Q hypothetical values, such that the corresponding elementary class $\mathcal{F}$ of functions $\mathcal{L} \to [-V, V]$ (with $V \geq \frac{1}{2}$) has pseudo-dimension $\mathsf{P}_{\mathsf{dim}}$. Define the following quantities:*

$$h = \log\Big(e\,(2V + \log(Q))\,Q^{3/2}\Big) + \frac{\log(e\,\mathsf{P}_{\mathsf{dim}}+1)}{\mathsf{P}_{\mathsf{dim}}} + \frac{\log(2)}{\mathsf{P}_{\mathsf{dim}}\,Q}$$
$$\eta = \log\left(\frac{64\,(2V + \log(Q))^2}{N}\right) + \log\left(\mathsf{P}_{\mathsf{dim}}\,Qh + \log\left(\frac{1}{\delta}\right)\right)$$

where N denotes the number of profiling traces. Define also the following quantity

$$\epsilon_{\mathsf{P}_{\mathsf{dim}},Q,V,N,\delta} = 8\,(2V + \log(Q))\sqrt{\frac{\log\left(\frac{1}{\delta}\right) + \mathsf{P}_{\mathsf{dim}}\,Q\left(h + \frac{\eta}{2}\right)}{N}} \quad .$$

Then, for all $0 < \delta \leq 1$, the inequality

$$\sup_{\mathsf{m}\in\mathcal{H}} \left|\Delta^{\mathsf{m}}_{\tilde{\mathsf{e}}_N} - \mathsf{PI}(Y; \boldsymbol{L}; \mathsf{m})\right| \leq \epsilon_{\mathsf{P}_{\mathsf{dim}},Q,V,N,\delta} \tag{35}$$

holds with probability at least $1 - \delta$.

We prove Theorem 5 in Appendix B. Corollary 6 follows from this result.

[10] This rule of thumb is not always true for other classes of models beyond the scope of this study. The interested reader may find counter-examples in [Vap98, pp. 159-160].

Corollary 6. *Let $\mathcal{A}_{\mathcal{H}}$ be a TI_N-maximizer adversary that profiles with N traces and considers a hypothesis class $\mathcal{H}$ such that the corresponding elementary class $\mathcal{F}$ has pseudo-dimension $\mathsf{P}_{\mathsf{dim}}$. The following inequalities*

$$\begin{aligned} 0 \leq \ & \mathsf{LI}(Y;\boldsymbol{L};\mathcal{H}) - \mathsf{PI}(Y;\boldsymbol{L};\widehat{\mathsf{m}}_N) \leq 2\epsilon_{\mathsf{P}_{\mathsf{dim}},Q,V,N,\delta} \\ -3\epsilon_{\mathsf{P}_{\mathsf{dim}},Q,V,N,\delta} \leq & \mathsf{TI}_N(Y;\boldsymbol{L};\mathcal{A}_{\mathcal{H}}) - \mathsf{LI}(Y;\boldsymbol{L};\mathcal{H}) \leq \epsilon_{\mathsf{P}_{\mathsf{dim}},Q,V,N,\delta} \end{aligned}$$

hold with probability $1-\delta$ (except the first one that always holds), and the slack $\epsilon_{\mathsf{P}_{\mathsf{dim}},Q,V,N,\delta}$ belongs to $\widetilde{\mathcal{O}}\left(V\sqrt{\frac{\mathsf{P}_{\mathsf{dim}}\,Q}{N}}\right)$.

Proof. The first inequality is a direct consequence of the definition of the LI. The second one is a direct consequence of Theorem 5 and Theorem 6 (proven in Appendix), while the two last ones follow from Corollary 7. □

Putting the pseudo-dimensions of our models of interest in this Corollary gives our generic convergence results ($\forall\,\mathsf{p}$ in Table 1).

B.3 Proof of Theorem 5

In this section, we prove Theorem 5. The proof is done in several steps that we briefly describe hereafter before diving into the details.

1. We bound the gap between $\mathsf{TI}_N(Y;\boldsymbol{L};\widehat{\mathsf{m}}_N)$ and $\mathsf{PI}(Y;\boldsymbol{L};\widehat{\mathsf{m}}_N)$ with a *uniform* bound, *i.e.*, not specific to any $\mathsf{m} \in \mathcal{H}$. We are now reduced to show that the gap uniformly converges towards 0.
2. We invoke a theorem stating that the uniform convergence rate is upper bounded by a quantity depending on the so-called *covering numbers* that we will define.
3. We will then introduce some properties of covering numbers in order to reduce the problem to bounding the covering number of the different $\mathcal{F}_i$.
4. The covering numbers can actually be bounded by the pseudo-dimension introduced in Section B.1.
5. We now have all the ingredients to state the theorem and its corollary.

B.4 Uniform Convergence

Definition 9 (Uniform Convergence)**.** Let $\mathcal{H}$ be a hypothesis class. We say that $\mathcal{H}$ has the *uniform convergence* property if for any probability distribution over $(Y,\boldsymbol{L})$, and for any $\epsilon,\delta > 0$, the following inequality is satisfied:

$$\Pr\left(\sup_{\mathsf{m}\in\mathcal{H}} \left|\Delta^{\mathsf{m}}_{\tilde{\mathsf{e}}_N} - \mathsf{PI}(Y;\boldsymbol{L};\mathsf{m})\right| \geq \epsilon\right) \leq \delta \ . \tag{36}$$

Theorem 6 (Uniform Convergence implies Learnability)**.** *With the same notations as in Definition 9, the inequality*

$$\mathsf{LI}(Y;\boldsymbol{L};\mathcal{H}) - \mathsf{PI}(Y;\boldsymbol{L};\widehat{\mathsf{m}}_N) \leq 2\sup_{\mathsf{m}\in\mathcal{H}} \left|\mathsf{PI}(Y;\boldsymbol{L};\mathsf{m}) - \Delta^{\mathsf{m}}_{\tilde{\mathsf{e}}_N}\right| \tag{37}$$

is satisfied.

Proof. Let $\mathsf{m} \in \mathcal{H}$ be fixed, and let us denote $\widehat{\mathsf{m}}_N = \mathcal{A}_{\mathcal{H}}(\tilde{\mathsf{e}}_N)$. By Definition 5, we have $\mathsf{TI}_N(Y; \boldsymbol{L}; \mathcal{A}_{\mathcal{H}}) = \Delta_{\tilde{\mathsf{e}}_N}^{\widehat{\mathsf{m}}_N} \geq \Delta_{\tilde{\mathsf{e}}_N}^{\mathsf{m}}$, therefore

$$\begin{aligned}
\Delta_{\mathsf{p}}^{\mathsf{m}} - \Delta_{\mathsf{p}}^{\widehat{\mathsf{m}}_N} &= \left(\Delta_{\mathsf{p}}^{\mathsf{m}} - \Delta_{\tilde{\mathsf{e}}_N}^{\widehat{\mathsf{m}}_N}\right) + \left(\Delta_{\tilde{\mathsf{e}}_N}^{\widehat{\mathsf{m}}_N} - \Delta_{\mathsf{p}}^{\widehat{\mathsf{m}}_N}\right) \\
&\leq \left(\Delta_{\mathsf{p}}^{\mathsf{m}} - \Delta_{\tilde{\mathsf{e}}_N}^{\mathsf{m}}\right) + \left(\Delta_{\tilde{\mathsf{e}}_N}^{\widehat{\mathsf{m}}_N} - \Delta_{\mathsf{p}}^{\widehat{\mathsf{m}}_N}\right) \\
&\leq \left|\Delta_{\mathsf{p}}^{\mathsf{m}} - \Delta_{\tilde{\mathsf{e}}_N}^{\mathsf{m}}\right| + \left|\Delta_{\mathsf{p}}^{\widehat{\mathsf{m}}_N} - \Delta_{\tilde{\mathsf{e}}_N}^{\widehat{\mathsf{m}}_N}\right| \\
&\leq 2 \sup_{\mathsf{m}' \in \mathcal{H}} \left|\Delta_{\mathsf{p}}^{\mathsf{m}'} - \Delta_{\tilde{\mathsf{e}}_N}^{\mathsf{m}'}\right| \quad .
\end{aligned}$$

Since the right hand-side does not depend on the fixed m, taking the supremum of the left hand side with respect to m, concludes the proof. □

In other words, it suffices to prove the uniform convergence for our hypothesis class $\mathcal{H}$ to show that the PI converges towards its supremum. Interestingly, the uniform convergence of $\mathcal{H}$ is also a necessary condition [ABCH97, Thm. 4.2].[11]

Corollary 7. *Let* $\epsilon = \sup_{\mathsf{m} \in \mathcal{H}} \left|\mathsf{PI}(Y; \boldsymbol{L}; \mathsf{m}) - \Delta_{\tilde{\mathsf{e}}_N}^{\mathsf{m}}\right|$, *the following inequalities hold*

$$-3\epsilon \leq \mathsf{TI}_N(Y; \boldsymbol{L}; \mathcal{A}_{\mathcal{H}}) - \mathsf{LI}(Y; \boldsymbol{L}; \mathcal{H}) \leq \epsilon \tag{38}$$

Proof. We first prove the first inequality:

$$\begin{aligned}
\mathsf{LI}(Y; \boldsymbol{L}; \mathcal{H}) - \mathsf{TI}_N(Y; \boldsymbol{L}; \mathcal{A}_{\mathcal{H}}) &= \mathsf{LI}(Y; \boldsymbol{L}; \mathcal{H}) - \mathsf{PI}(Y; \boldsymbol{L}; \widehat{\mathsf{m}}_N) \\
&\quad + \mathsf{PI}(Y; \boldsymbol{L}; \widehat{\mathsf{m}}_N) - \mathsf{TI}_N(Y; \boldsymbol{L}; \mathcal{A}_{\mathcal{H}}) \\
&\leq 2 \sup_{\mathsf{m} \in \mathcal{H}} \left|\mathsf{PI}(Y; \boldsymbol{L}; \mathsf{m}) - \Delta_{\tilde{\mathsf{e}}_N}^{\mathsf{m}}\right| \\
&\quad + \sup_{\mathsf{m} \in \mathcal{H}} \left|\mathsf{PI}(Y; \boldsymbol{L}; \mathsf{m}) - \Delta_{\tilde{\mathsf{e}}_N}^{\mathsf{m}}\right|
\end{aligned}$$

where the bound on the first term comes from Theorem 6 and the bound on the second term follows from the definition of $\mathsf{TI}_N(Y; \boldsymbol{L}; \mathcal{A}_{\mathcal{H}})$.

Next, we prove the second inequality

$$\begin{aligned}
\mathsf{TI}_N(Y; \boldsymbol{L}; \mathcal{A}_{\mathcal{H}}) - \mathsf{LI}(Y; \boldsymbol{L}; \mathcal{H}) &= \left(\mathsf{TI}_N(Y; \boldsymbol{L}; \mathcal{A}_{\mathcal{H}}) - \mathsf{PI}(Y; \boldsymbol{L}; \widehat{\mathsf{m}}_N)\right) \\
&\quad - \left(\mathsf{LI}(Y; \boldsymbol{L}; \mathcal{H}) - \mathsf{PI}(Y; \boldsymbol{L}; \widehat{\mathsf{m}}_N)\right) \\
&\leq \sup_{\mathsf{m} \in \mathcal{H}} \left|\mathsf{PI}(Y; \boldsymbol{L}; \mathsf{m}) - \Delta_{\tilde{\mathsf{e}}_N}^{\mathsf{m}}\right| - 0
\end{aligned}$$

where the bound on the second term follows from the definition of the LI. □

B.5 Bounding Uniform Convergence with Covering Numbers

We now turn to emphasize uniform bounds, which, thanks to Corollary 7, will enable us to draw bounds on the gap between TI_N and LI. The main idea of the results that we will present in this section is to reduce the uniform convergence for infinite hypothesis classes to the uniform convergence for finite hypothesis classes, provided further assumptions. To this end, we need to introduce the concept of *covering numbers*.

Definition 10 (Covering of a set [SB14, Def. 27.1])**.** Let $\mathcal{A}$ be a normed vector space with respect to the $\|\cdot\|_1$ norm, and $\epsilon > 0$. We say that $\mathcal{A}$ is ϵ*-covered by a set* $\mathcal{A}'$*, with respect to the* $\|\cdot\|_1$ *norm*, if for all $\boldsymbol{a} \in \mathcal{A}$, there exists a vector $\boldsymbol{a}' \in \mathcal{A}'$ such that $\|\boldsymbol{a} - \boldsymbol{a}'\|_1 \leq \epsilon$. We define by $N_1(\epsilon, \mathcal{A})$ the cardinality of the smallest $\mathcal{A}'$ that ϵ-covers $\mathcal{A}$.

[11] For more general learning problem, the uniform convergence may be not necessary (see counter-example in [Vap98, Sec. 3.12]). Nevertheless, a relaxed form of uniform convergence, called *one-sided* convergence, becomes the necessary and sufficient condition for a learning algorithm to be consistent [Vap98, Thm. 3.2].

In a nutshell, an ϵ-covering of a set $\mathcal{A}$ can be seen as a *representative* finite sample of $\mathcal{A}$, in the sense that any point from $\mathcal{A}$ is ϵ-close from at least one element from the covering. Therefore, any analysis that is done over the covering is likely to still hold (up to an error margin depending on at most ϵ) over the whole set $\mathcal{A}$.

Beyond metric spaces, covering numbers can also be defined for functional spaces, such as the ones we consider here. The following definition formally states this idea.

Definition 11 (Covering number of a hypothesis class [AB02, Sec. 10.4])**.** Let $\mathcal{H}$ be a set of functions from an input space $\mathcal{L}$ to a subset of $\mathbb{R}^Q$. Given a sequence $\mathcal{S}_N = (\boldsymbol{l}_1, \dots, \boldsymbol{l}_N) \in \mathcal{L}^N$ of input data, we let $\mathcal{H}_{\mathcal{S}_N}$ be the following set:

$$\mathcal{H}_{\mathcal{S}_N} = \left\{(f(\boldsymbol{l}_1), \dots, f(\boldsymbol{l}_N)) \in \mathbb{R}^{N \times Q} : f \in \mathcal{H}\right\}$$

For a positive number ϵ, we define the *covering number of $\mathcal{H}$ for accuracy ϵ and number of data N* as the quantity

$$\mathcal{N}_1(\epsilon, \mathcal{H}, N) = \max_{\mathcal{S}_N \in \mathcal{L}^N} N_1(\epsilon, \mathcal{H}_{\mathcal{S}_N}) \quad . \tag{39}$$

Covering numbers are crucial in statistical learning theory. This is formally stated by Theorem 7 hereafter.

Theorem 7 ([Hau92, Thm. 3])**.** *Let $\mathcal{H}$ be a permissible*[12] *hypothesis class of functions from $\mathcal{L}$ to $\mathcal{P}(\mathcal{Y})$, such that for all $\mathsf{m} \in \mathcal{H}$, and $y, \boldsymbol{l} \in \mathcal{Y} \times \mathcal{L}$, $0 \leq -\log(\mathsf{m}[y]) \leq B$. Assume $N \geq 1$. Suppose that $\mathcal{S}_N$ is generated by N independent random draws according to any joint probability distribution on $\mathcal{Y} \times \mathcal{L}$. Then*

$$\Pr\left(\sup_{\mathsf{m} \in \mathcal{H}} \left|\mathsf{PI}(Y; \boldsymbol{L}; \mathsf{m}) - \Delta^{\mathsf{m}}_{\tilde{\mathsf{e}}_N}\right| > \epsilon\right) \leq 2\,\mathcal{N}_1(\epsilon, \log \circ \mathcal{H}, 2N)\, e^{-\frac{\epsilon^2 N}{64 B^2}} \quad , \tag{40}$$

where $\log \circ \mathcal{H}$ *denotes the set of functions* $\{y, \boldsymbol{l} \mapsto -\log(\mathsf{m}[y]) : \mathsf{m} \in \mathcal{H}\}$.

It now remains to see when Theorem 7 provides non-trivial bounds. Indeed, assuming that $(\log \circ \mathcal{H})_{\mathcal{S}_N}$ is a subset of $[0, B]^N$, for some $B > 0$, then the covering number $\mathcal{N}_1(\epsilon, \log \circ \mathcal{H}, N)$ can itself be trivially bounded by $\left(\frac{BN}{\epsilon}\right)^N$. Unfortunately, in that case, the right hand-side of Equation 40 tends to infinity with $N \to \infty$, if ϵ is small enough. In other words, without further assumption, Theorem 7 is a rather tautological result, and further conditions on $\mathcal{H}$ must be set for sound bounds.

Hopefully, we will see in Section B.7 that for some classes of functions, we can get tighter bounds for covering numbers, yielding non-trivial worst-case of uniform convergence rates. Before going further through our reasoning, we need a few technical lemmas concerning covering numbers. Those technical results will be helpful to derive the aimed bounds.

B.6 A Few Properties about Covering Numbers

In this section, we introduce some technical lemmas that will be helpful for bounding the covering numbers. We start with the *contraction* lemma that leverages the Lipschitz property of a function.

Lemma 2 (Contraction)**.** *Let $\mathcal{A}, \mathcal{B}$ be two sets, and $\phi : \mathcal{A} \to \mathcal{B}$ be a ρ-Lipschitz function for a given norm $\|\cdot\|$ respectively induced on $\mathcal{A}, \mathcal{B}$. That is, for $\boldsymbol{a}, \boldsymbol{b} \in \mathcal{A}$, the following inequality holds:*

$$\|\phi(\boldsymbol{a}) - \phi(\boldsymbol{b})\|_{\mathcal{B}} \leq \rho \|\boldsymbol{a} - \boldsymbol{b}\|_{\mathcal{A}} \quad . \tag{41}$$

Then, if N denotes the covering number with respect to the considered norm, the inequality

$$N_1(\rho\epsilon, \phi \circ \mathcal{A}) \leq N_1(\epsilon, \mathcal{A}) \tag{42}$$

is valid.

[12] A very loose condition, see [Hau92, Footnote 11].

Lemma 2 is inspired by the proof given by Shalev-Shwartz and Ben-David [SB14, Lemma 27.2] who showed the result for the $\|\cdot\|_2$ norm. We observe however that the result can be generalized to any norm.

Proof. By definition, there exists a minimal ϵ-covering of $\mathcal{A}$ of size $N_1(\epsilon, \mathcal{A})$. Then, for any $\boldsymbol{a} \in \mathcal{A}$, there exists $\boldsymbol{a'}$ from the covering $\mathcal{A}'$ such that the following inequality holds:

$$\|\boldsymbol{a} - \boldsymbol{a'}\| \leq \epsilon \ . \tag{43}$$

Define $\mathcal{B} = \phi \circ \mathcal{A}$ and $\mathcal{B}' = \phi \circ \mathcal{A}'$. It follows from the Lipschitz property of ϕ that:

$$\|\phi(\boldsymbol{a}) - \phi(\boldsymbol{a'})\| \leq \rho\|\boldsymbol{a} - \boldsymbol{a'}\| \leq \rho\epsilon \ . \tag{44}$$

Hence, $\mathcal{B}'$ is a $(\rho\epsilon)$-cover of $\mathcal{B}$. □

Corollary 8 (Contraction)**.** *Using the same notations as in Lemma 2, if ϕ is a ρ-Lipschitz function (with respect to a given norm), then for any set of functions $\mathcal{F}$, one can bound the covering numbers of $\phi \circ \mathcal{F}$ as follows:*

$$\mathcal{N}_1(\rho\epsilon, \phi \circ \mathcal{F}, N) \leq \mathcal{N}_1(\epsilon, \mathcal{F}, N) \ . \tag{45}$$

Proof. Recalling that $\mathcal{N}_1(\epsilon, \mathcal{F}, N)$ is by definition the maximum value of $N_1(\epsilon, \mathcal{A})$ over all the sets $\mathcal{A}$ of size N in the image set of $\mathcal{F}$, the result straightforwardly follows from Lemma 2. □

Informally, Corollary 8 tells us that the smoother the function ϕ – in the sense that the lower its Lipschitz constant ρ – the less are needed to get an ϵ-cover of the image set by considering the image of the ϵ-cover of the input space. Therefore, it is useful to reduce the covering numbers computation of an hypothesis class if the latter one is a set of composed smooth functions. The direct application of Corollary 8 is to bound the covering number of $\log \circ \mathcal{H}$ with the covering number of $\mathcal{F}^Q$ defined as the set $\{h : \mathcal{L} \to \mathbb{R}^Q : \sigma \circ h \in \mathcal{H}\}$, *i.e.*, such that $\sigma \circ \mathcal{F}^Q = \mathcal{H}$. Let us first observe that the Lipschitz constant of the composed function $\log \circ \sigma$ is bounded by the square root of the number of its entries, as stated by Lemma 3.

Lemma 3. *For all $1 \leq i \leq Q$, the function $\boldsymbol{x} \in \mathbb{R}^Q \mapsto \log(\sigma(\boldsymbol{x})_i)$ is $\sqrt{Q}$-Lipschitz in the $\|\cdot\|_1$ and $\|\cdot\|_2$ norms.*

Proof. Denote by ϕ the considered function. Since $\|\boldsymbol{x}\|_2 \leq \|\boldsymbol{x}\|_1$, it suffices to show that ϕ is $\sqrt{Q}$-Lipschitz in the $\|\cdot\|_2$ norm. Moreover, it is known that the Lipschitz constant in the latter norm is bounded by the supremum over the range of $\boldsymbol{x}$ of the $\|\cdot\|_2$ norm of the gradient of ϕ. For $1 \leq j \leq Q$, the partial derivative of ϕ with respect to $\boldsymbol{x}_j$ is $\delta_{i,j} - \sigma(\boldsymbol{x})_j$, where $\delta_{i,j}$ denotes the Kronecker symbol. Since both $\delta_{i,j}$ and $\sigma(\boldsymbol{x})_j$ are bounded in $[0, 1]$, it implies that the Lipschitz constant is bounded by $\sqrt{Q}$. □

Corollary 9. *For all $\epsilon > 0$, and for all $N \geq 1$, the following inequality holds:*[13]

$$\mathcal{N}_1(\epsilon, \log \circ \mathcal{H}, N) \leq \mathcal{N}_1\left(\frac{\epsilon}{\sqrt{Q}}, \mathcal{F}^Q, N\right) \ . \tag{46}$$

Thanks to Corollary 9, we are now reduced to bound the covering number of the set $\mathcal{F}^Q$, which we now address. We start by defining the set of functions $\mathcal{F}^Q$ previously introduced as a *free product* of Q elementary sets of functions.

[13] A similar result can be found in [AB02, Lemma 17.6]

Definition 12 (Free product)**.** Let $\mathcal{A} = \mathcal{A}_1 \times \ldots \times \mathcal{A}_Q$ be the Cartesian product of Q metric spaces (for the L^1 distance). Let $\mathcal{F}_i$ be a family of functions from $\mathcal{L}$ into $\mathcal{A}_i$. The *free product* of the $\mathcal{F}_i$ is the class of functions

$$\mathcal{F}^Q = \{\boldsymbol{f} = (f_1, \ldots, f_Q) : f_i \in \mathcal{F}\} \ ,$$

where $\boldsymbol{f} = (f_1, \ldots, f_Q) : \mathcal{L} \to \mathcal{A}$ is the function defined by

$$(f_1, \ldots, f_Q)(\boldsymbol{l}) = \begin{pmatrix} f_1(\boldsymbol{l}) \\ \vdots \\ f_Q(\boldsymbol{l}) \end{pmatrix} \ .$$

We may now properly bound the covering number of $\mathcal{F}^Q$ in terms of covering numbers of the $\mathcal{F}_i$, thanks to Lemma 4.

Lemma 4 ([Hau92, Lemma 7])**.** *If $\mathcal{F}_1, \ldots, \mathcal{F}_Q$ are defined as above, then*

$$\mathcal{N}_1\big(\epsilon, \mathcal{F}^Q, N\big) \leq \prod_{i=1}^{k} \mathcal{N}_1\bigg(\frac{\epsilon}{Q}, \mathcal{F}_i, N\bigg) \ . \tag{47}$$

Proof. For each $1 \leq i \leq Q$, let $\mathcal{U}_i$ be an $\frac{\epsilon}{Q}$-cover for $\mathcal{F}_i$. Let

$$\mathcal{U} = \{(f_1, \ldots, f_Q) : f_i \in \mathcal{U}_i, 1 \leq i \leq Q\} \ . \tag{48}$$

Let us show that $\mathcal{U}$ is an ϵ-cover for $\mathcal{F}$. That is, let $\boldsymbol{g} = (g_1, \ldots, g_Q) \in \mathcal{H}$, and let us show that there exists $\boldsymbol{f} \in \mathcal{U}$ such that $\|\boldsymbol{g} - \boldsymbol{f}\|_1 \leq \epsilon$. For all $1 \leq i \leq Q$, since $\mathcal{U}_i$ is an $\frac{\epsilon}{Q}$-cover of $\mathcal{F}_i$, we know that there exists $f_i \in \mathcal{U}_i$ such that $\|g_i - f_i\|_1 \leq \frac{\epsilon}{Q}$. Let us consider then $\boldsymbol{h} = (h_1, \ldots, h_k)$. Notice that

$$\|\boldsymbol{g} - \boldsymbol{f}\|_1 = \sum_{i=1}^{Q} \|g_i - f_i\|_1 \leq Q \cdot \frac{\epsilon}{Q} \leq \epsilon \ . \tag{49}$$

Hence, $\mathcal{U}$ is an ϵ-cover for $\mathcal{F}^Q$. It now remains to notice that the cardinality of $\mathcal{U}$ is the product of cardinalities for $\mathcal{U}_i, 1 \leq i \leq Q$. □

B.7 Bounding the Covering Numbers of $\mathcal{F}$ with $\mathsf{P}_{\mathsf{dim}}(\mathcal{F})$

We finally come to the link between covering numbers and pseudo-dimensions, thanks to the following results.

Theorem 8 ([Hau95, Thm. 1])**.** *Let $\mathcal{F}$ be a non-empty set of real functions mapping from a domain $\mathcal{L}$ to the real interval $[0, 1]$ and suppose that $\mathcal{F}$ has finite pseudo-dimension $\mathsf{P}_{\mathsf{dim}}(\mathcal{F})$. Then*

$$\mathcal{N}_1(\epsilon, \mathcal{F}, N) \leq e(\mathsf{P}_{\mathsf{dim}}(\mathcal{F}) + 1) \left(\frac{2e}{\epsilon}\right)^{\mathsf{P}_{\mathsf{dim}}(\mathcal{F})} \tag{50}$$

for all $\epsilon > 0$.

Corollary 10. *Let $\mathcal{F}$ be a non-empty set of real functions mapping from a domain $\mathcal{L}$ to the real interval $[0, B]$ and suppose that $\mathcal{F}$ has finite pseudo-dimension $\mathsf{P}_{\mathsf{dim}}(\mathcal{F})$. Then*

$$\mathcal{N}_1(\epsilon, \mathcal{F}, N) \leq e(\mathsf{P}_{\mathsf{dim}}(\mathcal{F}) + 1) \left(\frac{2eB}{\epsilon}\right)^{\mathsf{P}_{\mathsf{dim}}(\mathcal{F})} \tag{51}$$

for all $\epsilon > 0$.

Proof. Straightforward, by applying Corollary 8 to $\mathcal{F}$ and $\phi \circ \mathcal{F}$, where $\phi(\boldsymbol{x}) = \frac{1}{B}\boldsymbol{x}$. □

Comparing with the trivial bound $\left(\frac{BN}{\epsilon}\right)^N$ discussed before, Corollary 10 provides a much tighter bound since it no longer depends on the amount N of profiling data. This noticeable property is the cornerstone of statistical learning theory, in the sense that it makes the results from Theorem 7 much more useful now.

B.8 Putting all Together

Now we have characterized every element in the upper bound of Theorem 7 in terms of pseudo-dimension of $\mathcal{F}$, we may gather all those results to come back to a concrete bound. Let us denote $P = \Pr\left(\sup_{\mathsf{m}\in\mathcal{H}} \left|\mathsf{PI}(Y; \boldsymbol{L}; \mathsf{m}) - \Delta^{\mathsf{m}}_{\tilde{\mathsf{e}}_N}\right| > \epsilon\right)$. Applying Theorem 7, it comes that

$$\begin{aligned}
P \cdot e^{\frac{\epsilon^2 N}{64 B^2}} &\overset{(40)}{\leq} 2\mathcal{N}_1(2\epsilon, \log \circ \mathcal{H}, 2N) \\
&\overset{(46)}{\leq} 2\mathcal{N}_1\left(2\frac{\epsilon}{\sqrt{Q}}, \mathcal{F}^Q, 2N\right) \\
&\overset{(47)}{\leq} 2\mathcal{N}_1\left(2\frac{\epsilon}{Q^{3/2}}, \mathcal{F}, 2N\right)^Q \\
&\overset{(51)}{\leq} 2\left((e\,\mathsf{P}_{\mathsf{dim}}(\mathcal{F}) + 1)\left(\frac{eBQ^{3/2}}{\epsilon}\right)^{\mathsf{P}_{\mathsf{dim}}(\mathcal{F})}\right)^Q \quad .
\end{aligned}$$

Let

$$\begin{aligned}
\alpha &= \frac{N}{64B^2} \\
\beta &= \frac{1}{2}\,\mathsf{P}_{\mathsf{dim}}(\mathcal{F})\,Q \\
\gamma &= \mathsf{P}_{\mathsf{dim}}(\mathcal{F})\,Q\log\left(eBQ^{3/2}\right) + Q\log(e\,\mathsf{P}_{\mathsf{dim}}(\mathcal{F}) + 1) + \log(2) \quad ,
\end{aligned}$$

the latter inequality can be rephrased as

$$P \leq \exp\left(-\alpha\epsilon^2 - \beta\log\left(\epsilon^2\right) + \gamma\right) \quad . \tag{52}$$

Let $\delta > 0$. We would like to find a sufficient condition such that $P \leq \delta$. It suffices to find a sufficient condition such that

$$\alpha\epsilon^2 + \beta\log\left(\epsilon^2\right) \geq \gamma + \log\left(\frac{1}{\delta}\right) \quad . \tag{53}$$

Let

$$\begin{aligned}
\epsilon_0^2 &= \max\left(\frac{\gamma + \log\left(\frac{1}{\delta}\right)}{\alpha}, 0\right) \\
\epsilon^2 &= \epsilon_0^2 + \max\left(-\frac{\beta}{\alpha}\log\left(\epsilon_0^2\right), 0\right) \quad ,
\end{aligned}$$

we shall show that Equation 53 is satisfied. Using the above definitions, we have

$$\epsilon^2 \geq \epsilon_0^2 - \frac{\beta}{\alpha}\log\left(\epsilon_0^2\right) \geq \frac{\gamma + \log\left(\frac{1}{\delta}\right)}{\alpha} - \frac{\beta}{\alpha}\log\left(\epsilon_0^2\right) \quad .$$

Moreover, since $\epsilon^2 \geq \epsilon_0^2$, it holds that $\frac{\beta}{\alpha}\log\left(\epsilon^2\right) \geq \frac{\beta}{\alpha}\log\left(\epsilon_0^2\right)$. Finally, summing the two above equations gives Equation 53.

It now remains to replace the bound B of the loss function by a more practical bound on the output range of each elementary class $\mathcal{F}$. This is stated by the following lemma.

Lemma 5. *Let $\boldsymbol{x} \in \mathbb{R}^Q$ such that for all i, $|\boldsymbol{x}_i| \leq V$. Then,*

$$0 \leq -\log(\sigma(\boldsymbol{x})) \leq 2V + \log(Q) \quad . \tag{54}$$

Proof.

$$\begin{aligned} -\log(\sigma(\boldsymbol{x})) = \log\left(1 + \sum_{j \neq i} e^{\boldsymbol{x}_j - \boldsymbol{x}_i}\right) &\leq \log\big(1 + (Q-1)\, e^{2V}\big) \\ &\leq \log\big(Qe^{2V}\big) = 2V + \log(Q) \quad . \end{aligned}$$

□

This result allows us to replace B with $2V + \log(Q)$ in the definitions of α and β, which, along with the hypothesis $V \geq \frac{1}{2}$, allows us to observe that $\gamma \geq 1$, hence we can remove the max in the definition of ϵ_0: $\epsilon_0^2 = \left(\gamma + \log\left(\frac{1}{\delta}\right)\right)/\alpha$.

Finally, taking the complement probability in Equation 52, and expliciting the expression of ϵ gives Theorem 5.

C Proofs of fast rate

C.1 Convergence of the PI

Theorem 9 ([Meh17, Thm. 1], restated)**.** *Let $\mathcal{H} = \{\mathsf{m}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \mathcal{H}^{\mathsf{T}}\}$ such that $\boldsymbol{\theta} \in \mathcal{H}^{\mathsf{T}} \subset \mathbb{R}^{\mathsf{P}}$ is a convex set satisfying $\sup_{\boldsymbol{\theta}',\boldsymbol{\theta}} \|\boldsymbol{\theta}' - \boldsymbol{\theta}\|_2 \leq T$. Suppose, for all $y, \boldsymbol{l} \in \mathcal{Y} \times \mathcal{L}$, that the mapping $\boldsymbol{\theta} \mapsto \log(\mathsf{m}(y \mid \boldsymbol{l}))$ is U-Lipschitz. Suppose that the true leakage model p belongs to $\mathcal{H}$ and that for all $y \in \mathcal{Y}, \boldsymbol{l} \in \mathcal{L}, \mathsf{m} \in \mathcal{H}$ $\left|\log\left(\frac{\mathsf{m}(y|\boldsymbol{l})}{\mathsf{p}(y|\boldsymbol{l})}\right)\right| \leq B$. Then, if $N \geq 5$, with probability at least $1 - \delta$, the TI_N-maximizer returns a model $\widehat{\mathsf{m}}_N$ such that*

$$\mathsf{MI}(Y; \boldsymbol{L}) - \mathsf{PI}(Y; \boldsymbol{L}; \widehat{\mathsf{m}}_N) \leq \frac{1}{N} 8B \left(\mathsf{P}\log(16UTN) + \log\left(\frac{1}{\delta}\right)\right) + \frac{1}{N} \quad . \tag{55}$$

Remark 3. In Theorem 9, we assumed that the true leakage model belongs to the hypothesis class. Such a requirement can often be relaxed [vEGM+15, Example 2.2], up to a multiplicative constant in the convergence rates.

We introduce hereafter a few technical lemmas that will be useful to derive the proofs.

Lemma 6. *Let $\boldsymbol{l} \in \mathcal{L}$ be such that $\|\boldsymbol{l}\|_2 \leq R$. Let $\boldsymbol{\Theta}$ be a parameter vector such that $\mathsf{m}_{\boldsymbol{\Theta}} \in \mathcal{H}$, where $\mathcal{H}$ denotes the hypothesis class of an LR_2 attacker. Then, for all $y \in \mathcal{Y}$ and for all $\boldsymbol{l} \in \mathcal{L}$, the mapping $\boldsymbol{\Theta} \mapsto \log\Big(\sigma(\mathsf{m}_{\boldsymbol{\Theta}}(\boldsymbol{l}))_y\Big)$ is ρ-Lipschitz for the norm $\|\cdot\|_2$ with $\rho \leq \sqrt{Q}(R^2 + 1)$.*

Proof. Using Lemma 3, we get that for all $(y, \boldsymbol{l})$,

$$\left|\log\Big(\sigma(\mathsf{m}_{\boldsymbol{\Theta}}(\boldsymbol{l}))_y\Big) - \log\Big(\sigma(\mathsf{m}_{\boldsymbol{\Theta}'}(\boldsymbol{l}))_y\Big)\right| \leq \sqrt{Q}\sqrt{\sum_{i=1}^{Q} \left(\mathsf{m}_{\boldsymbol{\Theta}}(\boldsymbol{l})_i - \mathsf{m}_{\boldsymbol{\Theta}'}(\boldsymbol{l})_i\right)^2} \quad . \tag{56}$$

Since m is an LR_2 model, $\mathsf{m}_{\boldsymbol{\Theta}}(\boldsymbol{l})_i = \boldsymbol{l}'^{\mathsf{T}} A_i \boldsymbol{l}'$ where $\boldsymbol{l}' = (\boldsymbol{l}, 1)$. Therefore, using Cauchy-Schwartz' inequality, we get

$$\begin{aligned} |\mathsf{m}_{\boldsymbol{\Theta}}(\boldsymbol{l})_i - \mathsf{m}_{\boldsymbol{\Theta}'}(\boldsymbol{l})_i| &= |\boldsymbol{l}'^{\mathsf{T}}(A_i - A_i')\boldsymbol{l}'| \\ &\leq \|\boldsymbol{l}'\|_2^2 \, \|A_i - A_i'\|_* \\ &\leq (R^2 + 1)\, \|A_i - A_i'\|_F \\ &= (R^2 + 1)\, \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_i'\|_2 \end{aligned}$$

Injecting this bound into Equation 56 gives the desired result. □

Lemma 7. *With the same notations has before, if now we are considering an* LR_1 *attacker, then the resulting mapping becomes ρ-Lipschitz with*

$$\rho \leq \sqrt{Q(R^2+1)} \ .$$

Proof. We now have $\mathsf{m}_{\boldsymbol{\Theta}}(\boldsymbol{l})_i = B_i \boldsymbol{l}'$ (still with $\boldsymbol{l}' = (\boldsymbol{l}, 1)$), and thus

$$|\mathsf{m}_{\boldsymbol{\Theta}}(\boldsymbol{l})_i - \mathsf{m}_{\boldsymbol{\Theta}'}(\boldsymbol{l})_i| \leq \|\boldsymbol{l}'\|_2 \, \|B_i - B_i'\|_2 \leq \sqrt{R^2+1} \, \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_i'\|_2 \ .$$

Injecting this bound into Equation 56 concludes the proof. □

Restatement of Theorem 9. The original version of Mehta's theorem [Meh17, Thm. 1] required the loss function to be *exp-concave*,[14] instead of the true leakage model p belonging to $\mathcal{H}$. Nevertheless, Mehta's proof relies on another more general assumption, the so-called *η-central condition*. This central condition is implied either by assuming the loss function to be η-exp-concave, or in the particular case where the loss function is the log-loss, by assuming that the true leakage distribution p belongs to $\mathcal{H}$ [vEGM+15, Example 2.2]. In the latter case, the parameter η is set to 1. Beside, the supremum of PI can be replaced by MI, since we assume $\mathsf{p} \in \mathcal{H}$. The remaining of Mehta's proof remains unchanged. □

Proof of Corollary 1 for LR_1*.* This is a direct application of Theorem 9, by properly setting the parameters of the theorem. First, observe that $\mathcal{H}^\intercal \subset \mathbb{R}^{(D+1)\times Q}$ so $\mathsf{P} = (D+1)Q$, and taking $T = 2S\sqrt{Q}$ satisfies $\sup_{\boldsymbol{\Theta}', \boldsymbol{\Theta}} \|\boldsymbol{\Theta}' - \boldsymbol{\Theta}\|_2 \leq T$.

Next, the condition $\left|\log\left(\frac{\mathsf{m}(y|\boldsymbol{l})}{\mathsf{p}(y|\boldsymbol{l})}\right)\right| \leq B$ is satisfied if both $\log(\mathsf{m}(y \mid \boldsymbol{l})) - \log(\mathsf{p}(y \mid \boldsymbol{l})) \leq B$ and $\log(\mathsf{p}(y \mid \boldsymbol{l})) - \log(\mathsf{m}(y \mid \boldsymbol{l})) \leq B$. Since $\mathsf{p}(y \mid \boldsymbol{l}) \leq 1$ and $\mathsf{m}(y \mid \boldsymbol{l}) \leq 1$ the condition reduces to $-\log(\mathsf{p}(y \mid \boldsymbol{l})) \leq B$ and $-\log(\mathsf{m}(y \mid \boldsymbol{l})) \leq B$. Furthermore, $\mathsf{p} \in \mathcal{H}$, it only remains to find B such that $-\log(\mathsf{m}(y \mid \boldsymbol{l})) \leq B$ for all $\mathsf{m} \in \mathcal{H}$. Using Lemma 5 and the observation that $|B_i \boldsymbol{l}'| \leq \sqrt{R^2+1}S$ (where $\boldsymbol{l}' = (\boldsymbol{l}, 1)$), we get that $B = 2\sqrt{R^2+1}S + \log(Q)$ satisfies the condition.

Finally, using Lemma 7, we get that the Lipschitz constant L is upper bounded by $\sqrt{Q(R^2+1)}$. Putting all together into Equation 55 gives the desired result. □

Proof of Corollary 1 for LR_2*.* This is a direct application of Theorem 9, by properly setting the parameters of the theorem. As previously, we have $\mathsf{P} = (D+1)Q$ and $T = 2S\sqrt{Q}$. Furthermore, using the same reasoning as before, but using the bound $|B_i \boldsymbol{l}'| \leq (R^2+1)S$, we get $B = 2(R^2+1)S + \log(Q)$. Finally, using Lemma 6, we get that $L \leq \sqrt{Q}(R^2+1)$. Putting all together into Equation 55 gives the desired result. □

Proof of Corollary 2. This is a direct application of Theorem 9, by properly setting the parameters of the theorem to fit the different assumptions.

First, recall from Section 5.1 that our class of models is composed of Q MLPs, each being made of W real parameters by assumption. Hence, $\mathcal{H}^\intercal \subset \mathbb{R}^{W \times Q}$ so $\mathsf{P} = WQ$.

Second, we bound $\sup_{\boldsymbol{\theta}, \boldsymbol{\theta}'} \|\boldsymbol{\theta}' - \boldsymbol{\theta}\|$. Notice that for each MLP ϕ_y plugged to the entries of the softmax, $\|\boldsymbol{\theta}_i\| \leq LS$ (we use $l2$ norms in this proof), so using the triangle inequality, we get that for all $\boldsymbol{\theta}, \boldsymbol{\theta}'$,

$$\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| \leq T = 2SQL \ . \tag{57}$$

Third, we show the Lipschitzness of MLPs. Using Lemma 3, we get that for all $(y, \boldsymbol{l})$,

$$\left|\log\Big(\sigma(\mathsf{m}_{\boldsymbol{\theta}}(\boldsymbol{l}))_y\Big) - \log\Big(\sigma(\mathsf{m}_{\boldsymbol{\theta}'}(\boldsymbol{l}))_y\Big)\right| \leq \sqrt{Q}\sqrt{\sum_{i=1}^{Q} \left(\mathsf{m}_{\boldsymbol{\theta}}(\boldsymbol{l})_i - \mathsf{m}_{\boldsymbol{\theta}'}(\boldsymbol{l})_i\right)^2} \ . \tag{58}$$

[14] A function φ is said to be η-exp-concave if the mapping $z \mapsto e^{-\eta f(z)}$ is concave.

We are now reduced to bound the Lipschitz constant of each entry model $\mathsf{m}_{\boldsymbol{\theta}}(\boldsymbol{l})_i$ of the softmax. Then, we may notice that since the ReLU activation function is 1-Lipschitz, each layer $\phi\left(\boldsymbol{x}^{(j)}, \boldsymbol{\Theta}_i^{(j)}\right)$ is $\|\boldsymbol{x}^{(j)}\|$-Lipschitz (resp. $\left\|\boldsymbol{\Theta}_i^{(j)}\right\|$-Lipschitz) in its input $\boldsymbol{\Theta}_i^{(j)}$ (resp. $\|\boldsymbol{x}^{(j)}\|$), hence

$$\left\|\phi\left(\boldsymbol{x}^{(j)}, \boldsymbol{\Theta}_i^{(j)}\right) - \phi\left(\boldsymbol{x}'^{(j)}, \boldsymbol{\Theta}_i'^{(j)}\right)\right\| \leq \left\|\boldsymbol{\Theta}_i'^{(j)}\right\| \left\|\boldsymbol{x}^{(j)} - \boldsymbol{x}'^{(j)}\right\| + \left\|\boldsymbol{x}^{(j)}\right\| \left\|\boldsymbol{\Theta}_i^{(j)} - \boldsymbol{\Theta}_i'^{(j)}\right\| . \tag{59}$$

Let us now prove by induction that

$$\left\|\boldsymbol{x}^{(j)} - \boldsymbol{x}'^{(j)}\right\| \leq RS^j \sum_{k=0}^{j} \left\|\boldsymbol{\Theta}_i^{(k)} - \boldsymbol{\Theta}_i'^{(k)}\right\| \quad , \tag{60}$$

where $\boldsymbol{x}^{(j+1)} = \phi\left(\boldsymbol{x}^{(j)}, \boldsymbol{\Theta}_i^{(j)}\right)$, $\boldsymbol{x}'^{(j+1)} = \phi\left(\boldsymbol{x}'^{(j)}, \boldsymbol{\Theta}_i'^{(j)}\right)$ and $\boldsymbol{x}^{(0)} = \boldsymbol{x}'^{(0)} = \boldsymbol{l}$. The base case $j = 1$ is a direct consequence of Equation 59, since $\|\boldsymbol{l}\| \leq R$ and $S \geq 1$. For $j \neq 1$, we observe that $\|\boldsymbol{x}^{(j+1)}\| \leq \left\|\boldsymbol{\Theta}_i^{(j)}\right\| \|\boldsymbol{x}^{(j)}\| \leq S^j \|\boldsymbol{l}\| \leq S^j R$. Then, injecting this observation in the second term of Equation 59 and using the induction hypothesis in the first term gives the desired result. Finally, we apply Equation 60 to the full MLP, giving

$$|\mathsf{m}_{\boldsymbol{\theta}}(\boldsymbol{l})_i - \mathsf{m}_{\boldsymbol{\theta}'}(\boldsymbol{l})_i| \leq R \cdot S^L \|\boldsymbol{\theta}_i' - \boldsymbol{\theta}_i\| \quad . \tag{61}$$

Injecting the right hand-side of Equation 61 into the one of Equation 58, we get that the Lipschitz constant is upper bounded by $U = \sqrt{Q}RS^L$. Finally, since $\mathsf{p} \in \mathcal{H}^\intercal$, we may combine Equation 57, Equation 58, Equation 61 to get that $\left|\log\left(\frac{\mathsf{m}(y|\boldsymbol{l})}{\mathsf{p}(y|\boldsymbol{l})}\right)\right| \leq B = 2Q^{3/2}RLS^{L+1}$. Putting all together into Equation 55 gives the desired result. □

C.2 Convergence of the TI_N

Proposition 3. *Let $\mathcal{H}$ be a finite hypothesis class such that any model $\mathsf{m} \in \mathcal{H}$ returns a probability distribution such that for any secret hypothesis y, $-\log \mathsf{m}[y] \leq B$, for some positive B. Assume that the true model p belongs to $\mathcal{H}$. Then*

$$\mathop{\mathbb{E}}_{\mathcal{S}_N}\left[\mathsf{TG}_N(Y; \boldsymbol{L}; \mathcal{A}_\mathcal{H})\right] \leq \frac{2B \cdot (\log|\mathcal{H}| + 1)}{N}$$

and

$$\mathop{\mathbb{E}}_{\mathcal{S}_N}\left[\left|\mathsf{TG}_N(Y; \boldsymbol{L}; \mathcal{A}_\mathcal{H}) - \mathop{\mathbb{E}}_{\mathcal{S}_N}\left[\mathsf{TG}_N(Y; \boldsymbol{L}; \mathcal{A}_\mathcal{H})\right]\right|\right] \in \mathcal{O}\left(\frac{B\log|\mathcal{H}|}{N} + \frac{1}{\sqrt{N}}\right) \quad .$$

Proof. Let $\Gamma = \mathsf{TG}_N(Y; \boldsymbol{L}; \mathcal{A}_\mathcal{H}) - \left(\Delta^{\mathsf{p}}_{\mathcal{S}_N} - \mathsf{LI}(Y; \boldsymbol{L}; \mathcal{H})\right)$. Notice that by definition, $\mathop{\mathbb{E}}_{\mathcal{S}_N}\left[\Delta^{\mathsf{p}}_{\mathcal{S}_N}\right] = \mathsf{MI}(Y; \boldsymbol{L})$ and since by assumption $\mathsf{p} \in \mathcal{H}$, we have $\mathsf{MI}(Y; \boldsymbol{L}) = \mathsf{LI}(Y; \boldsymbol{L}; \mathcal{H})$. As a result,

$$\mathop{\mathbb{E}}_{\mathcal{S}_N}[\Gamma] = \mathop{\mathbb{E}}_{\mathcal{S}_N}\left[\mathsf{TG}_N(Y; \boldsymbol{L}; \mathcal{A}_\mathcal{H})\right] \quad .$$

Moreover, since $\mathsf{TI}_N(Y; \boldsymbol{L}; \mathcal{A}_\mathcal{H}) \geq \Delta^{\mathsf{p}}_{\mathcal{S}_N}$ and $\mathsf{PI}(Y; \boldsymbol{L}; \mathcal{A}_\mathcal{H}) \leq \mathsf{LI}(Y; \boldsymbol{L}; \mathcal{H})$, $\Gamma \geq 0$. We are then reduced to bound the expected value of Γ. To this end, as recalled in Lemma 1, the assumption $\mathsf{p} \in \mathcal{H}$ implies that the central condition is verified. Van Erven *et al.* show that this implies that the so-called *Bernstein's condition* is verified [vEGM+15, p. 1829]. Bernstein's condition in turn implies that

$$\Gamma \leq 2 \cdot B \cdot \frac{\log(|\mathcal{H}|/\delta)}{N} \tag{62}$$

with probability at least $1 - \delta$ [Ler, p. 16]. We then use the well-known identity for positive random variables $\underset{\mathcal{S}_N}{\mathbb{E}}[\Gamma] = \int_0^\infty \Pr(\Gamma \geq \epsilon)\, d\epsilon$. Using Equation 62, we have that $\Pr(\Gamma \geq \epsilon) \leq |\mathcal{H}| \cdot e^{-\epsilon \frac{N}{2B}}$. This inequality is non-trivial for $\epsilon \geq \frac{2B \cdot \log|\mathcal{H}|}{N}$, otherwise we can anyway bound the probability by one. Hence,

$$\underset{\mathcal{S}_N}{\mathbb{E}}[\Gamma] = \int_0^\infty \Pr(\Gamma \geq \epsilon)\, d\epsilon \leq \int_0^{\frac{2B \cdot \log|\mathcal{H}|}{N}} d\epsilon + \int_{\frac{2B \cdot \log|\mathcal{H}|}{N}}^\infty |\mathcal{H}| \cdot e^{-\epsilon \frac{N}{2B}}\, d\epsilon = \frac{2B \cdot (\log|\mathcal{H}| + 1)}{N} \quad .$$

Next, we analyze the variance. We bound $\mathbb{V}[\Gamma] \leq \mathbb{E}\left[\Gamma^2\right]$ as

$$\begin{aligned}\underset{\mathcal{S}_N}{\mathbb{V}}[\Gamma] &= \int_0^\infty \Pr\left(\Gamma^2 \geq \epsilon\right) d\epsilon \leq \int_0^{\left(\frac{2B \cdot \log|\mathcal{H}|}{N}\right)^2} d\epsilon + \int_{\left(\frac{2B \cdot \log|\mathcal{H}|}{N}\right)^2}^\infty |\mathcal{H}| \cdot e^{-\sqrt{\epsilon}\frac{N}{2B}}\, d\epsilon \quad , \\ &= \left(\frac{2B \cdot (\log|\mathcal{H}| + 1)}{N}\right)^2 \quad .\end{aligned}$$

Then, from the definition of Γ, can bound the following standard deviation:

$$\begin{aligned}\sqrt{\underset{\mathcal{S}_N}{\mathbb{V}}[\mathsf{TG}_N(Y; \boldsymbol{L}; \mathcal{A}_{\mathcal{H}})]} &\leq \sqrt{\underset{\mathcal{S}_N}{\mathbb{V}}[\Gamma]} + \sqrt{\underset{\mathcal{S}_N}{\mathbb{V}}\left[\Delta^{\mathsf{p}}_{\mathcal{S}_N} - \mathsf{LI}(Y; \boldsymbol{L}; \mathcal{H})\right]} \\ &\leq \frac{2B \cdot (\log|\mathcal{H}| + 1)}{N} + \sqrt{\underset{\mathcal{S}_N}{\mathbb{V}}\left[\Delta^{\mathsf{p}}_{\mathcal{S}_N}\right]} \quad ,\end{aligned}$$

and, by Jensen's inequality, we can bound the average absolute deviation

$$\left(\underset{\mathcal{S}_N}{\mathbb{E}}\left[\left|\mathsf{TG}_N(Y; \boldsymbol{L}; \mathcal{A}_{\mathcal{H}}) - \underset{\mathcal{S}_N}{\mathbb{E}}[\mathsf{TG}_N(Y; \boldsymbol{L}; \mathcal{A}_{\mathcal{H}})]\right|\right]\right)^2 \leq \underset{\mathcal{S}_N}{\mathbb{V}}[\mathsf{TG}_N(Y; \boldsymbol{L}; \mathcal{A}_{\mathcal{H}})] \quad .$$

Finally, by the central limit theorem, the standard deviation of $\Delta^{\mathsf{p}}_{\mathcal{S}_N}$ belongs to $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$. □

Strictly speaking, Proposition 3 only holds for finite hypothesis classes. Nevertheless, we argue that the result extends to the TI_N-maximizers considered in this paper. Indeed, it is possible to make a reduction from infinite to finite hypothesis classes, similarly to what is stated in Theorem 7 in Appendix B. The log-cardinal of the finite hypothesis class would eventually be replaced by the *pseudo-dimension* introduced by Definition 8 in Appendix B, and that scales similarly to the constants in Corollaries 1 and 2 (see Theorems 3 and 4 in Appendix B).

D Proofs of Section 6

D.1 Proofs for the gTA bound

The first theorem presented hereafter uniformly bounds the regret of gTA with the regret induced by an imperfect characterization of the true distribution.

Theorem 10. *Let TA be an adversary with templates, i.e. with generative models* $\widehat{\mathsf{f}}_y(\cdot)\,, y \in \mathcal{Y}$ *of the distribution. Then, the following inequalities hold true.*

$$0 \leq \mathsf{R}(\mathit{TA}) \leq \frac{1}{Q}\sum_y \mathsf{D}_{\mathsf{KL}}\left(\mathsf{f}_y(\cdot) \,\middle\|\, \widehat{\mathsf{f}}_y(\cdot)\right) \leq \max_y \mathsf{D}_{\mathsf{KL}}\left(\mathsf{f}_y(\cdot) \,\middle\|\, \widehat{\mathsf{f}}_y(\cdot)\right) \tag{63}$$

Proof. Using the successively the definitons of the PI and the MI, and the linearity of the expectation, we get

$$\begin{aligned}
\mathsf{R}\left(\mathsf{gTA}\right) &= \mathsf{MI}(Y;\boldsymbol{L}) - \mathsf{PI}(Y;\boldsymbol{L};\mathsf{gTA}) \\
&= \frac{1}{Q}\sum_{y} \mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_y}\left[\log\left(\frac{\mathsf{f}_y(\boldsymbol{L})}{\sum_{y'}\mathsf{f}_{y'}(\boldsymbol{L})}\right) - \log\left(\frac{\widehat{\mathsf{f}}_y(\boldsymbol{L})}{\sum_{y'}\widehat{\mathsf{f}}_{y'}(\boldsymbol{L})}\right)\right] \\
&= \frac{1}{Q}\sum_{y} \mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_y}\left[\log\left(\frac{\mathsf{f}_y(\boldsymbol{L})}{\widehat{\mathsf{f}}_y(\boldsymbol{L})}\right)\right] - \mathop{\mathbb{E}}_{\boldsymbol{L}\sim\frac{1}{Q}\sum_y \mathsf{f}_y}\left[\log\left(\frac{\sum_{y'}\mathsf{f}_{y'}(\boldsymbol{L})}{\sum_{y'}\widehat{\mathsf{f}}_{y'}(\boldsymbol{L})}\right)\right] \\
&= \frac{1}{Q}\sum_{y} \mathsf{D}_{\mathsf{KL}}\left(\mathsf{f}_y(\cdot)\,\middle\|\,\widehat{\mathsf{f}}_y(\cdot)\right) - \mathsf{D}_{\mathsf{KL}}\left(\frac{\sum_y \mathsf{f}_y(\cdot)}{Q}\,\middle\|\,\frac{\sum_y \widehat{\mathsf{f}}_y(\cdot)}{Q}\right).
\end{aligned}$$

Since the KL divergence is always non-negative, we get the desired result. □

Note that Theorem 10 is not particular to Gaussian templates, and may be applied to any generative model. Next, we remark that the KL divergence remains invariant by affine transformation, as stated hereafter.

Lemma 8. *Let $A \in \mathbb{R}^{D\times D}$ be invertible, let $\boldsymbol{b} \in \mathbb{R}^D$, and $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^D$ be two random vectors, of pdf respectively $\mathsf{f}_{\boldsymbol{X}}(\cdot), \mathsf{f}_{\boldsymbol{Y}}(\cdot)$. Then,*

$$\mathsf{D}_{\mathsf{KL}}(\mathsf{f}_{\boldsymbol{X}}(\cdot) \parallel \mathsf{f}_{\boldsymbol{Y}}(\cdot)) = \mathsf{D}_{\mathsf{KL}}(\mathsf{f}_{A\cdot\boldsymbol{X}+\boldsymbol{b}}(\cdot) \parallel \mathsf{f}_{A\cdot\boldsymbol{Y}+\boldsymbol{b}}(\cdot)) \ . \tag{64}$$

Proof. Let $\boldsymbol{X'} = A\boldsymbol{X} + \boldsymbol{b}$, then the pdf of $\boldsymbol{X'}$ is

$$\mathsf{f}_{\boldsymbol{X'}}(\boldsymbol{x}) = |A|^{-1}\,\mathsf{f}_{\boldsymbol{X}}\left(A^{-1}\boldsymbol{x} - \boldsymbol{b}\right) \ .$$

By applying the change of variable $\boldsymbol{x'} = A\boldsymbol{x} + \boldsymbol{b}$ in the definition of KL divergence, it follows that

$$\begin{aligned}
\mathsf{D}_{\mathsf{KL}}(\mathsf{f}_{\boldsymbol{X}}(\cdot) \parallel \mathsf{f}_{\boldsymbol{Y}}(\cdot)) &= \mathop{\mathbb{E}}_{\boldsymbol{X}\sim|A|^{-1}\,\mathsf{f}_{\boldsymbol{X}}(A^{-1}(\cdot-\boldsymbol{b}))}\left[\log\left(\frac{|A|^{-1}\,\mathsf{f}_{\boldsymbol{X}}\left(A^{-1}(\boldsymbol{X}-\boldsymbol{b})\right)}{|A|^{-1}\,\mathsf{f}_{\boldsymbol{Y}}(A^{-1}(\boldsymbol{X}-\boldsymbol{b}))}\right)\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{X'}\sim\mathsf{f}_{\boldsymbol{X'}}}\left[\log\left(\frac{\mathsf{f}_{\boldsymbol{X'}}(\boldsymbol{X'})}{\mathsf{f}_{\boldsymbol{Y'}}(\boldsymbol{X'})}\right)\right]
\end{aligned}$$

Hence, we identify the right hand-side of Equation 64. □

For Gaussian templates, we can therefore reduce the study of the KL divergence of Theorem 10 to the particular case where the true covariance matrix Σ is the identity using Lemma 8. Furthermore, in the case of gTA with $\Sigma = I$, the following lemma gives an algebraic formulation of the upper bound.

Lemma 9. *For a Gaussian distribution with $\Sigma = I$, the KL divergence is given by:*

$$\begin{aligned}
2\mathsf{D}_{\mathsf{KL}}\left(\mathsf{f}(\cdot)\,\middle\|\,\widehat{\mathsf{f}}(\cdot)\right) &= \log\left(\det\left(\widehat{\Sigma}\right)\right) + \mathrm{Tr}\left(\widehat{\Sigma}^{-1}\right) - D \qquad (65)\\
&+ \left(\widehat{\mu} - \mu\right)^{\mathsf{T}}\,\widehat{\Sigma}^{-1}\left(\widehat{\mu} - \mu\right) \ . \qquad (66)
\end{aligned}$$

Proof. By definition,

$$\mathsf{D}_{\mathsf{KL}}\left(\mathsf{f}(\cdot)\,\middle\|\,\widehat{\mathsf{f}}(\cdot)\right) = \mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}}\left[\log\left(\frac{\mathsf{f}(\boldsymbol{L})}{\widehat{\mathsf{f}}(\boldsymbol{L})}\right)\right] \ .$$

Substituting both $\mathsf{f}(\cdot)$ and $\widehat{\mathsf{f}}(\cdot)$ with their respective density, it follows that

$$2\mathsf{D}_{\mathsf{KL}}\left(\mathsf{f}(\cdot) \,\middle\|\, \widehat{\mathsf{f}}(\cdot)\right) = \log\left(\frac{\det\left(\widehat{\Sigma}\right)}{\det(\Sigma)}\right) + \underset{\boldsymbol{L}\sim\mathsf{f}}{\mathbb{E}}\left[(\boldsymbol{L} - \widehat{\mu})^{\intercal}\,\widehat{\Sigma}^{-1}\,(\boldsymbol{L} - \widehat{\mu}) - (\boldsymbol{L} - \mu)^{\intercal}\,\Sigma^{-1}\,(\boldsymbol{L} - \mu)\right]$$

Using [PP$^+$08, Lemma 8.2.2], it follows that the second term inside the brackets has D as expected value, whereas the first term inside the brackets has $(\mu - \widehat{\mu})^{\intercal}\,\widehat{\Sigma}^{-1}\,(\mu - \widehat{\mu}) + \mathrm{Tr}\left(\widehat{\Sigma}^{-1}\Sigma\right)$ as expected value, hence the result. □

We now bound each term of Lemma 9. First, we bound Equation 66. The term (66) is the well known Hotelling's T^2 statistic, as recalled by the following lemma.

Lemma 10 ([And03, Thm. 5.2.2])**.** *For $N/Q \geq D$, the quantity*

$$\frac{N}{QD}\,\frac{N/Q - D}{N/Q - 1} \cdot (\widehat{\mu} - \mu)^{\intercal}\widehat{\Sigma}^{-1}(\widehat{\mu} - \mu) \tag{67}$$

follows a Fisher-Snedecor law of parameters $(D, N/Q - D)$.

Accordingly, as the Fisher distribution converges towards a χ^2 distribution with D degrees of freedom, it follows that the quantity (66) belongs to $\mathcal{O}\left(\frac{DQ}{N}\right)$.

Second, we bound Equation 65. The terms of Equation 65 are upper bounded in the following theorem.

Theorem 11. *Suppose that the leakage follows a Gaussian distribution with $\Sigma = I$, and that $\left\|\widehat{\Sigma} - I\right\|_* \leq 1/2$. Then the first following inequality always holds true and there exists a constant C such that for all $\delta > 0$ and for all $N \geq 4C^2 \log\left(\frac{2}{\delta}\right) D$ the second following inequality holds with probability at least $1 - \delta$:*

$$0 \leq \log\left(\det\left(\widehat{\Sigma}\right)\right) + \mathrm{Tr}\left(\widehat{\Sigma}^{-1}\right) - D \leq 2C \log\left(\frac{2}{\delta}\right) \frac{QD^2}{N} \quad . \tag{68}$$

The proof of this theorem relies on the following thechnical lemmas.

Lemma 11 (Basic linear algebra)**.** *Let $A, B \in \mathbb{R}^{D\times D}$ be symmetric matrices. Then,*

- $\det(AB) = \det(A)\det(B)$,
- $\det(A) = \prod_{i=1}^{D} \lambda_i$, *where $\lambda_1, \ldots, \lambda_D$ are its eigenvalues,*
- $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$,
- $\mathrm{Tr}(A) = \sum_{i=1}^{D} \lambda_i$,
- *If λ is an eigenvalue of A, then $\frac{1}{\lambda}$ is an eigenvalue of A^{-1}.*

Lemma 12. *For all $x \in (-1, 1)$, we have*

$$0 \leq x - \log(1 + x) \leq \frac{x^2}{1 + x} \quad . \tag{69}$$

Proof. It is widely known that $\frac{x}{1+x} \leq \log(1 + x) \leq x$. Multiplying by -1 and adding x, we get the result. □

We are now ready to demonstrate the desired result. The whole proof comes into two parts. First, in Lemma 13 we upper bound the quantity of interest in terms of spectral norms of the estimation error of the covariance matrix. Then, we invoke Theorem 12 to upper bound the latter spectral norm in terms of the parameters $N/Q, D$ of our problem.

Lemma 13. *Let $\widehat{\Sigma}$ be an empirical covariance matrix estimated from samples following the D-dimensional normal distribution with zero mean and the identity $\mathbf{I}$ as a covariance matrix. Then, if $\left\|\widehat{\Sigma}-\mathbf{I}\right\|_* \leq 1/2$,*

$$0 \leq \log\left(\det\left(\widehat{\Sigma}\right)\right) + \operatorname{Tr}\left(\widehat{\Sigma}^{-1}\right) - D \leq 2D\left\|\widehat{\Sigma}-\mathbf{I}\right\|_*^2 \ .$$

Proof. First, we rephrase the first two terms of the KL divergence in terms of eigenvalues $\lambda_1 \geq \ldots \geq \lambda_D$ of $\widehat{\Sigma}$. Since $\widehat{\Sigma}$ is a positive symmetric matrix, we know that λ_D is non-negative. Moreover, by assuming that $N/Q \geq D$, we know that $\lambda_D > 0$ with high probability. Furthermore,

$$\log\left(\det\left(\widehat{\Sigma}\right)\right) = \log\left(\prod_{i=1}^{D}\lambda_i\right) = \sum_{i=1}^{D}\log(\lambda_i) \ .$$

Besides, using Lemma 11,

$$\operatorname{Tr}\left(\widehat{\Sigma}^{-1}\right) - D = \sum_{i=1}^{D}\left(\frac{1}{\lambda_i} - 1\right) \ .$$

Hence, we may rephrase the quantity to upper bound as follows:

$$\log\left(\det\left(\widehat{\Sigma}\right)\right) + \operatorname{Tr}\left(\widehat{\Sigma}^{-1}\right) - D = \sum_{i=1}^{D}\left(\frac{1}{\lambda_i} - 1 - \log\left(\frac{1}{\lambda_i}\right)\right)$$

Using Lemma 12, the right hand-side of the latter equation is upper-bounded as follows:

$$\log\left(\det\left(\widehat{\Sigma}\right)\right) + \operatorname{Tr}\left(\widehat{\Sigma}^{-1}\right) - D \leq \sum_{i=1}^{D}\lambda_i\left(\frac{1}{\lambda_i} - 1\right)^2 = \sum_{i=1}^{D}\frac{(\lambda_i-1)^2}{\lambda_i} \ . \tag{70}$$

We then remark that if λ_i is an eigenvalue of $\widehat{\Sigma}$, then $\lambda_i - 1$ is an eigenvalue of $\widehat{\Sigma} - \mathbf{I}$, where $\mathbf{I} \in \mathbb{R}^{D\times D}$ denotes the identity matrix. As a consequence, for all $1 \leq i \leq D$,

$$|\lambda_i - 1| \leq \max_i|\lambda_i - 1| = \left\|\widehat{\Sigma}-\mathbf{I}\right\|_* \ .$$

Therefore, since by assumption $\left\|\widehat{\Sigma}-\mathbf{I}\right\|_* \leq 1/2$ we have for all i

$$0 \leq \frac{(\lambda_i-1)^2}{\lambda_i} \leq \frac{\left\|\widehat{\Sigma}-\mathbf{I}\right\|_*^2}{1-\left\|\widehat{\Sigma}-\mathbf{I}\right\|_*} \leq 2\left\|\widehat{\Sigma}-\mathbf{I}\right\|_*^2 \ . \tag{71}$$

Finally, combining Equation 71 with Equation 70 gives the result. □

We are now reduced to bound $\left\|\widehat{\Sigma}-\mathbf{I}\right\|_*$, which is the purpose of the following theorem.

Theorem 12 (Prop. 2.1 [Ver12])**.** *For all Σ, there exists a constant C such that for all $\delta > 0$, the inequality*

$$\left\|\widehat{\Sigma}-\Sigma\right\|_* \leq C\left\|\Sigma\right\|_* \cdot \sqrt{\log\left(\frac{2}{\delta}\right)\frac{D}{N}} \tag{72}$$

holds with probability at least $1-\delta$.

Proof of Theorem 11. The theorm is a direct combination of the bounds of Theorem 12 and Lemma 13. □

Proof of Corollary 3. Let us now combine all the previous results.

Proof. Starting from the KL divergence of Theorem 10, we restrict ourselves to the case $\Sigma = I$ using Lemma 8. Then, we get a bound on the KL divergence with Lemma 9, whose term are themselve bounded in Lemma 10 and Theorem 11. Finally, we can see that Hotelling's T^2 statistic can be neglected. □

D.1.1 Proof of Tightness

Theorem 13 ([CLZ15, Cor. 1])**.** *For all $\Sigma \in \mathbb{R}^{D \times D}$, the log determinant of $\widehat{\Sigma}$, estimated for N samples drawn from a multivariate Gaussian distribution of covariance matrix Σ, satisfies*

$$\frac{1}{\sqrt{2QD/N}} \left(\log\left(\frac{\det\left(\widehat{\Sigma}\right)}{\det(\Sigma)} \right) - QD(D+1)/(2N) \right) \xrightarrow[N \to \infty]{L} \mathcal{N}(0,1) \quad . \tag{73}$$

Theorem 13 is an analogue of the Central-Limit Theorem for the log-det term with a $\Theta\left(\frac{QD^2}{N}\right)$ positive bias. The following term shows that the bias from the trace of inverse covariance matrix is positive.

Lemma 14. *The trace of the inverse empirical covariance matrix is positively biased:*

$$\mathbb{E}\left[\mathrm{Tr}\left(\widehat{\Sigma}^{-1}\right) - D\right] \geq 0 \quad . \tag{74}$$

Proof. For any symmetric positive matrix such as $\widehat{\Sigma}$, the mapping $\widehat{\Sigma} \mapsto \mathrm{Tr}\left(\widehat{\Sigma}^{-1}\right)$ is convex [BV14, Ex. 3.18]. Using Jensen's inequality, we get

$$\mathbb{E}\left[\mathrm{Tr}\left(\widehat{\Sigma}^{-1}\right)\right] \geq \mathrm{Tr}\left(\mathbb{E}\left[\widehat{\Sigma}\right]^{-1}\right) \geq \mathrm{Tr}(I_D) = D \quad .$$

Hence, the left hand-side of Equation 74 is non-negative. □

Therefore, the latter bias cannot compensate the former one, which proves the tightness of our KL divergence bound (Lemma 9) in the general case.

D.2 Proofs for the Naive Bayes bound

Theorem 14. *Assume that $\Sigma = I$ and $\widehat{\Sigma}$ is a diagonal matrix. Then, for all $\delta > 0$ the following inequality holds:*

$$0 \leq \log\left(\det\left(\widehat{\Sigma}\right)\right) + \mathrm{Tr}\left(\widehat{\Sigma}^{-1}\right) - D \leq C \log\left(\frac{2}{\delta}\right) \frac{DQ}{N} \quad . \tag{75}$$

Proof. Since $\widehat{\Sigma}$ is diagonal then $\log\det\left(\widehat{\Sigma}\right)$ exactly coincides with the sum of the empirical log-variances estimated for each of the D time samples of the traces. Likewise, $\mathrm{Tr}\left(\widehat{\Sigma}^{-1}\right)$ coincides with the sum of inverse empirical variances. Estimating the error term in Equation 75 can be reduced to estimate the sum of D error terms, each for one-dimensional covariance matrices. Therefore, using Equation 68 in the particular case where $D = 1$, and multiplying by the true dimensionality D gives the result. □

Proof of Corollary 4. The proof is almost identical to the proof of Corollary 3, using Theorem 14 instead of Theorem 11. Finally, we can see that Hotelling's T^2 statistic has the same convergence rate as Theorem 14. □

D.3 Proof for the p-gTA

For two classes, we may use a change of variable such that the true covariance matrix is the identity, and the two true centroids are situated respectively at $\mp\frac{\Delta}{2}\boldsymbol{e_1}$ (where $\boldsymbol{e_1} = (1, 0, \dots, 0)$). In that case, $\boldsymbol{\beta} = \widehat{\Sigma}^{-1}(\widehat{\mu}_1 - \widehat{\mu}_0) - \Delta\boldsymbol{e_1}$ and $\gamma = \frac{1}{2}\left(\widehat{\mu}_0^\intercal\widehat{\Sigma}^{-1}\widehat{\mu}_0 - \widehat{\mu}_1^\intercal\widehat{\Sigma}^{-1}\widehat{\mu}_1\right)$. It also follows:

Lemma 15. *The regret for two classes can be rephrased as follows*

$$2\mathsf{R}\left(\textit{p-gTA}\right) = \mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_0}\left[\log\left(1 + e^{\widehat{\lambda}(\boldsymbol{L})}\right)\right] + \mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_1}\left[\log\left(1 + e^{-\widehat{\lambda}(\boldsymbol{L})}\right)\right] - 2\mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_0}\left[\log\left(1 + e^{\Delta\boldsymbol{e}_1^\intercal\boldsymbol{L}}\right)\right] \quad , \tag{76}$$

where $\widehat{\lambda}(\boldsymbol{L}) = (\Delta\boldsymbol{e_1} + \boldsymbol{\beta})^\intercal\boldsymbol{L} + \gamma$.

Proof. First, denoting $\boldsymbol{l}_1 = \boldsymbol{e}_1^\intercal\boldsymbol{l}$ (and $\boldsymbol{L}_1 = \boldsymbol{e}_1^\intercal\boldsymbol{L}$), we observe that

$$\mathsf{p}(0 \mid \boldsymbol{l}) = \frac{\mathsf{f}_0(\boldsymbol{l})}{\mathsf{f}_0(\boldsymbol{l}) + \mathsf{f}_1(\boldsymbol{l})} = \frac{e^{-\frac{1}{2}(\boldsymbol{l}_1 + \frac{\Delta}{2})^2}}{e^{-\frac{1}{2}(\boldsymbol{l}_1 + \frac{\Delta}{2})^2} + e^{-\frac{1}{2}(\boldsymbol{l}_1 - \frac{\Delta}{2})^2}} = \frac{1}{1 + e^{\Delta\boldsymbol{l}_1}}$$

and, since $\mathsf{f}_1(-\boldsymbol{l}) = \mathsf{f}_0(\boldsymbol{l})$, we have $\mathsf{p}(1 \mid \boldsymbol{l}) = \mathsf{p}(0 \mid -\boldsymbol{l})$. Furthermore,

$$\mathsf{m}(0 \mid \boldsymbol{l}) = \frac{e^{-\frac{1}{2}(\boldsymbol{l} - \widehat{\mu}_0)^\intercal\widehat{\Sigma}^{-1}(\boldsymbol{l} - \widehat{\mu}_0)}}{e^{-\frac{1}{2}(\boldsymbol{l} - \widehat{\mu}_0)^\intercal\widehat{\Sigma}^{-1}(\boldsymbol{l} - \widehat{\mu}_0)} + e^{-\frac{1}{2}(\boldsymbol{l} - \widehat{\mu}_1)^\intercal\widehat{\Sigma}^{-1}(\boldsymbol{l} - \widehat{\mu}_1)}} = \frac{1}{1 + e^{\widehat{\lambda}(\boldsymbol{l})}} \quad ,$$
$$\mathsf{m}(1 \mid \boldsymbol{l}) = \frac{e^{-\frac{1}{2}(\boldsymbol{l} - \widehat{\mu}_1)^\intercal\widehat{\Sigma}^{-1}(\boldsymbol{l} - \widehat{\mu}_1)}}{e^{-\frac{1}{2}(\boldsymbol{l} - \widehat{\mu}_0)^\intercal\widehat{\Sigma}^{-1}(\boldsymbol{l} - \widehat{\mu}_0)} + e^{-\frac{1}{2}(\boldsymbol{l} - \widehat{\mu}_1)^\intercal\widehat{\Sigma}^{-1}(\boldsymbol{l} - \widehat{\mu}_1)}} = \frac{1}{1 + e^{-\widehat{\lambda}(\boldsymbol{l})}} \quad .$$

Then, we have

$$\begin{aligned} 2\mathsf{R}\left(\mathsf{p\text{-}gTA}\right) &= \mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_0}\left[\log\left(\frac{\mathsf{p}(0 \mid \boldsymbol{L})}{\mathsf{m}(0 \mid \boldsymbol{L})}\right)\right] + \mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_1}\left[\log\left(\frac{\mathsf{p}(1 \mid \boldsymbol{L})}{\mathsf{m}(1 \mid \boldsymbol{L})}\right)\right] \\ &= \mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_0}\left[\log\left(\frac{1}{\mathsf{m}(0 \mid \boldsymbol{L})}\right)\right] + \mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_1}\left[\log\left(\frac{1}{\mathsf{m}(1 \mid \boldsymbol{L})}\right)\right] \\ &\quad - \left(\mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_0}\left[\log\left(\frac{1}{\mathsf{p}(0 \mid \boldsymbol{L})}\right)\right] + \mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_1}\left[\log\left(\frac{1}{\mathsf{p}(1 \mid \boldsymbol{L})}\right)\right]\right) \end{aligned}$$

and, by making the change of variable $\boldsymbol{L}' = -\boldsymbol{L}$ in the last term, we remark that the two last terms are equal. Finally, injecting our values of $\mathsf{p}(0 \mid \boldsymbol{l})$, $\mathsf{m}(0 \mid \boldsymbol{l})$ and $\mathsf{m}(1 \mid \boldsymbol{l})$ into this expression gives the expected result. □

Theorem 15. *Let* μ_0, μ_1, Σ *be respectively the* D*-dimensional centroids of the two classes, and the pooled covariance matrix. Let* p-gTA *be an attacker outputting estimates* $\widehat{\mu}_0, \widehat{\mu}_1, \widehat{\Sigma}$ *from the profiling phase. Let*

$$\boldsymbol{\beta} = \quad \widehat{\Sigma}^{-1}\left(\widehat{\mu}_1 - \widehat{\mu}_0\right) \quad -\Sigma^{-1}\left(\mu_1 - \mu_0\right) \quad , \tag{77}$$

$$\gamma = \quad -\tfrac{1}{2}\left(\widehat{\mu}_1\widehat{\Sigma}^{-1}\widehat{\mu}_1 - \widehat{\mu}_0\widehat{\Sigma}^{-1}\widehat{\mu}_0\right) \quad +\frac{1}{2}\left(\mu_1\Sigma^{-1}\mu_1 - \mu_0\Sigma^{-1}\mu_0\right) \quad . \tag{78}$$

Then, the regret of p-gTA *satisfies*

$$\mathsf{R}\left(\textit{p-gTA}\right) \leq \left(\gamma^2 + \|\boldsymbol{\beta}\|_2^2 + |\gamma\boldsymbol{\beta}_1|\right) + \mathcal{O}\left(\left(\gamma^2 + \|\boldsymbol{\beta}\|_2^2\right)^{3/2}\right) \tag{79}$$

where $\boldsymbol{\beta}_1$ *is the first element of* $\boldsymbol{\beta}$.

Proof. Using the expression of the regret given in Lemma 15 and taking the Taylor expansion (with the notation $\mathsf{R}(\boldsymbol{\beta},\gamma) = \mathsf{R}(\mathsf{p\text{-}gTA})$), we have

$$\begin{aligned}\mathsf{R}(\boldsymbol{\beta},\gamma) = {} & \mathsf{R}(\mathbf{0},0) + \frac{\partial}{\partial\gamma}\mathsf{R}(\mathbf{0},0)\cdot\gamma + \nabla_{\boldsymbol{\beta}}\mathsf{R}(\mathbf{0},0)^{\intercal}\boldsymbol{\beta} \\ & + \frac{1}{2}\left(\frac{\partial^2}{\partial\gamma^2}\mathsf{R}(\mathbf{0},0)\gamma^2 + \frac{\partial}{\partial\gamma}\nabla_{\boldsymbol{\beta}}\mathsf{R}(\mathbf{0},0)^{\intercal}\boldsymbol{\beta}\cdot\gamma + \boldsymbol{\beta}^{\intercal}\nabla^2_{\boldsymbol{\beta}}\mathsf{R}(\mathbf{0},0)\boldsymbol{\beta}\right) \\ & + \mathcal{O}\left(\left(\gamma^2 + \|\boldsymbol{\beta}\|^2\right)^{3/2}\right) \quad .\end{aligned} \tag{80}$$

We shall prove that

1. All zero-th and first-order terms are zero and,
2. The second-order terms are bounded by constant independent of D.

All first-order terms are zero. First, observe that for $\boldsymbol{\beta} = \mathbf{0}, \gamma = 0$, the model corresponds to the true distribution: $\mathsf{m}(y \mid \boldsymbol{l}) = \mathsf{p}(y \mid \boldsymbol{l})$ and thus $\mathsf{R}(\mathbf{0},0) = 0$. Second, let us express $\frac{\partial}{\partial\gamma}\mathsf{R}(\mathbf{0},0)$:

$$\begin{aligned}\frac{\partial}{\partial\gamma}\mathsf{R}(\mathbf{0},0) &= \frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_0}\left[\frac{\partial}{\partial\gamma}\left\{\log\left(1+e^{\widehat{\lambda}(\boldsymbol{L})}\right)\right\}(\mathbf{0},0)\right] + \frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_1}\left[\frac{\partial}{\partial\gamma}\left\{\log\left(1+e^{-\widehat{\lambda}(\boldsymbol{L})}\right)\right\}(\mathbf{0},0)\right] \\ &= \frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_0}\left[\frac{\frac{\partial}{\partial\gamma}\left\{e^{\widehat{\lambda}(\boldsymbol{L})}\right\}(\mathbf{0},0)}{1+e^{\Delta\boldsymbol{e}_1^{\intercal}\boldsymbol{L}}}\right] + \frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_1}\left[\frac{\frac{\partial}{\partial\gamma}\left\{e^{-\widehat{\lambda}(\boldsymbol{L})}\right\}(\mathbf{0},0)}{1+e^{-\Delta\boldsymbol{e}_1^{\intercal}\boldsymbol{L}}}\right] \\ &= \frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_0}\left[\frac{e^{\Delta\boldsymbol{e}_1^{\intercal}\boldsymbol{L}}}{1+e^{\Delta\boldsymbol{e}_1^{\intercal}\boldsymbol{L}}}\right] - \frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_1}\left[\frac{e^{-\Delta\boldsymbol{e}_1^{\intercal}\boldsymbol{L}}}{1+e^{-\Delta\boldsymbol{e}_1^{\intercal}\boldsymbol{L}}}\right] \\ &= \frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_0}\left[\frac{e^{\Delta\boldsymbol{e}_1^{\intercal}\boldsymbol{L}}}{1+e^{\Delta\boldsymbol{e}_1^{\intercal}\boldsymbol{L}}}\right] - \frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_0}\left[\frac{e^{\Delta\boldsymbol{e}_1^{\intercal}\boldsymbol{L}}}{1+e^{\Delta\boldsymbol{e}_1^{\intercal}\boldsymbol{L}}}\right] = 0,\end{aligned}$$

where we used the same change of variable as in the proof of Lemma 15 in the last line.

Now, let us express $\nabla_{\boldsymbol{\beta}}\mathsf{R}(\mathbf{0},0)^{\intercal}\boldsymbol{\beta}$. Similarly to the derivation of $\frac{\partial}{\partial\gamma}\mathsf{R}(\mathbf{0},0)$, we get that

$$\frac{\partial}{\partial\boldsymbol{\beta}_i}\mathsf{R}(\mathbf{0},0) = \frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_0}\left[\frac{\boldsymbol{L}_i e^{\Delta\boldsymbol{e}_1^{\intercal}\boldsymbol{L}}}{1+e^{\Delta\boldsymbol{e}_1^{\intercal}\boldsymbol{L}}}\right] + \frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_1}\left[\frac{-\boldsymbol{L}_i e^{-\Delta\boldsymbol{e}_1^{\intercal}\boldsymbol{L}}}{1+e^{-\Delta\boldsymbol{e}_1^{\intercal}\boldsymbol{L}}}\right] \tag{81}$$

Applying the same change of variable as previously, we get that

$$\frac{\partial}{\partial\boldsymbol{\beta}_i}\mathsf{R}(\mathbf{0},0) = \mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_0}\left[\frac{\boldsymbol{L}_i e^{\Delta\boldsymbol{e}_1^{\intercal}\boldsymbol{L}}}{1+e^{\Delta\boldsymbol{e}_1^{\intercal}\boldsymbol{L}}}\right] \quad .$$

Since f_0 is a multivariate Gaussian with diagonal covariance matrix, $\boldsymbol{L}_i$ is independent of $\boldsymbol{L}_1$ for all $1 < i \leq D$, and furthermore the mean of $\boldsymbol{L}_i$ is zero. Therefore, for such $i \neq 1$,

$$\frac{\partial}{\partial\boldsymbol{\beta}_i}\mathsf{R}(\mathbf{0},0) = \mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_0}\left[\boldsymbol{L}_i\right]\mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_0}\left[\frac{e^{\Delta\boldsymbol{e}_1^{\intercal}\boldsymbol{L}}}{1+e^{\Delta\boldsymbol{e}_1^{\intercal}\boldsymbol{L}}}\right] = 0 \quad .$$

For the remaining case where $i = 1$, observe that

$$\mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_0}\left[\boldsymbol{L}_1\frac{e^{\Delta\boldsymbol{L}_1}}{1+e^{\Delta\boldsymbol{L}_1}}\right] = K\int_{-\infty}^{\infty}\frac{e^{-\frac{1}{2}(x+\Delta/2)^2}x}{1+e^{-\Delta x}}\mathrm{d}x = Ke^{-\Delta^2/8}\int_{-\infty}^{\infty}x\frac{e^{-x^2/2}}{e^{\Delta x/2}+e^{-\Delta x/2}}\mathrm{d}x$$

for some constant K. Since the latter integrand is an even function of $\mathbb{R}$, the integral equals 0.

Bounds for Second-Order Terms. Finally, it remains to bound the second-order terms. For $1 \leq i, j, \leq D$, the (i, j)-coefficient of the Hessian matrix of $\log\left(1 + e^{\widehat{\lambda}(\boldsymbol{L})}\right)$ is given by

$$\frac{\partial^2}{\partial \boldsymbol{\beta}_i \partial \boldsymbol{\beta}_j} \log\left(1 + e^{\widehat{\lambda}(\boldsymbol{L})}\right) = \frac{\partial}{\partial \boldsymbol{\beta}_i} \left\{ \boldsymbol{L}_j \frac{e^{\widehat{\lambda}(\boldsymbol{L})}}{1 + e^{\widehat{\lambda}(\boldsymbol{L})}} \right\} = \boldsymbol{L}_i \boldsymbol{L}_j \frac{e^{\widehat{\lambda}(\boldsymbol{L})}}{\left(1 + e^{\widehat{\lambda}(\boldsymbol{L})}\right)^2} \ .$$

Likewise, we have

$$\frac{\partial^2}{\partial \boldsymbol{\beta}_i \partial \boldsymbol{\beta}_j} \log\left(1 + e^{-\widehat{\lambda}(\boldsymbol{L})}\right) = -\frac{\partial}{\partial \boldsymbol{\beta}_i} \left\{ \boldsymbol{L}_j \frac{e^{-\widehat{\lambda}(\boldsymbol{L})}}{1 + e^{-\widehat{\lambda}(\boldsymbol{L})}} \right\} = \boldsymbol{L}_i \boldsymbol{L}_j \frac{e^{\widehat{\lambda}(\boldsymbol{L})}}{\left(1 + e^{\widehat{\lambda}(\boldsymbol{L})}\right)^2} \ .$$

Using the change of variable $\mathsf{f}_1(\boldsymbol{l}) = \mathsf{f}_0(-\boldsymbol{l})$, this gives

$$\frac{\partial^2}{\partial \boldsymbol{\beta}_i \partial \boldsymbol{\beta}_j} \mathsf{R}\left(\mathbf{0}, 0\right) = \frac{1}{2} \left(\mathop{\mathbb{E}}_{\boldsymbol{L} \sim \mathsf{f}_0} \left[\frac{\boldsymbol{L}_i \boldsymbol{L}_j e^{\Delta \boldsymbol{L}_1}}{(1 + e^{\Delta \boldsymbol{L}_1})^2} \right] + \mathop{\mathbb{E}}_{\boldsymbol{L} \sim \mathsf{f}_1} \left[\frac{\boldsymbol{L}_i \boldsymbol{L}_j e^{\Delta \boldsymbol{L}_1}}{(1 + e^{\Delta \boldsymbol{L}_1})^2} \right] \right) = \mathop{\mathbb{E}}_{\boldsymbol{L} \sim \mathsf{f}_0} \left[\boldsymbol{L}_i \boldsymbol{L}_j \frac{e^{\Delta \boldsymbol{L}_1}}{(1 + e^{\Delta \boldsymbol{L}_1})^2} \right] . \tag{82}$$

For $1 \leq i < j \leq D$, the right hand-side of Equation 82 is zero since $\boldsymbol{L}_j$ is independent of $\boldsymbol{L}_i$ and $\boldsymbol{L}_1$, and furthermore the mean of $\boldsymbol{L}_j$ is zero. For $1 < i = j \leq D$ the right hand-side is positive and can be upper bounded by $\mathop{\mathbb{E}}_{\boldsymbol{L} \sim \mathsf{f}_0} \left[\boldsymbol{L}_i^2 \right] = 1$. In the last case $i = j = 1$, the second derivative of the regret is also positive and reduces to

$$\int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}(\boldsymbol{L}_1 + \Delta/2)^2}}{\sqrt{2\pi}} \boldsymbol{L}_1^2 \frac{e^{\Delta \boldsymbol{L}_1}}{(1 + e^{\Delta \boldsymbol{L}_1})^2} \mathrm{d}\boldsymbol{L}_1 = \int_{-\infty}^{\infty} \frac{e^{-\frac{\Delta^2}{8}}}{(1 + e^{\Delta \boldsymbol{L}_1})^2} \boldsymbol{L}_1^2 \frac{e^{-\frac{\boldsymbol{L}_1^2}{2}}}{\sqrt{2\pi}} \mathrm{d}\boldsymbol{L}_1 \leq \int_{-\infty}^{\infty} \boldsymbol{L}_1^2 \frac{e^{-\frac{1}{2}\boldsymbol{L}_1^2}}{\sqrt{2\pi}} \mathrm{d}\boldsymbol{L}_1 \tag{83}$$

where the last integral is equal to 1 (it is the variance of a standard normal distribution). Therefore, the following bounds hold:

$$0 \leq \boldsymbol{\beta}^{\intercal} \nabla_{\boldsymbol{\beta}}^2 \mathsf{R}\left(\mathbf{0}, 0\right) \boldsymbol{\beta} \leq \|\boldsymbol{\beta}\|_2^2 \ .$$

Similarly to Equation 82, it can be shown that for $1 \leq j \leq D$ we have

$$\frac{\partial^2}{\partial \gamma \partial \boldsymbol{\beta}_j} \mathsf{R}\left(\mathbf{0}, 0\right) = \mathop{\mathbb{E}}_{\boldsymbol{L} \sim \mathsf{f}_0} \left[\frac{\boldsymbol{L}_j e^{\Delta \boldsymbol{L}_1}}{(1 + e^{\Delta \boldsymbol{L}_1})^2} \right] \ . \tag{84}$$

For $j > 1$ the latter partial derivative equals zero since $\boldsymbol{L}_j$ is independent of $\boldsymbol{L}_1$ and has zero mean. For $j = 1$, using a reasonning similar to Equation 83, we get that $\frac{\partial^2}{\partial \gamma \partial \boldsymbol{\beta}_1} \mathsf{R}\left(\mathbf{0}, 0\right) \leq 0$. Let us now look for a lower bound:

$$\begin{aligned}
\frac{\partial^2}{\partial \gamma \partial \boldsymbol{\beta}_1} \mathsf{R}\left(\mathbf{0}, 0\right) &= \int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}(\boldsymbol{L}_1 + \Delta/2)^2}}{\sqrt{2\pi}} \boldsymbol{L}_1 \frac{e^{\Delta \boldsymbol{L}_1}}{(1 + e^{\Delta \boldsymbol{L}_1})^2} \mathrm{d}\boldsymbol{L}_1 = \int_{-\infty}^{\infty} \frac{e^{-\frac{\Delta^2}{8}}}{(1 + e^{\Delta \boldsymbol{L}_1})^2} \boldsymbol{L}_1 \frac{e^{-\frac{\boldsymbol{L}_1^2}{2}}}{\sqrt{2\pi}} \mathrm{d}\boldsymbol{L}_1 \\
&\geq \int_{-\infty}^{0} \frac{e^{-\frac{\Delta^2}{8}}}{(1 + e^{\Delta \boldsymbol{L}_1})^2} \boldsymbol{L}_1 \frac{e^{-\frac{\boldsymbol{L}_1^2}{2}}}{\sqrt{2\pi}} \mathrm{d}\boldsymbol{L}_1 = -\int_{0}^{\infty} \frac{e^{-\frac{\Delta^2}{8}}}{(1 + e^{-\Delta \boldsymbol{L}_1})^2} \boldsymbol{L}_1 \frac{e^{-\frac{\boldsymbol{L}_1^2}{2}}}{\sqrt{2\pi}} \mathrm{d}\boldsymbol{L}_1 \\
&\geq -\int_{0}^{\infty} \boldsymbol{L}_1 e^{-\frac{1}{2}\boldsymbol{L}_1^2} \mathrm{d}\boldsymbol{L}_1 = -1 \ .
\end{aligned}$$

Finally, we have that

$$\frac{\partial^2}{\partial \gamma^2} \mathsf{R}\left(\mathbf{0}, 0\right) = \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{L} \sim \mathsf{f}_0} \left[\frac{e^{\Delta \boldsymbol{L}_1}}{(1 + e^{\Delta \boldsymbol{L}_1})^2} \right] + \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{L} \sim \mathsf{f}_1} \left[\frac{e^{-\Delta \boldsymbol{L}_1}}{(1 + e^{-\Delta \boldsymbol{L}_1})^2} \right] \ . \tag{85}$$

We deduce from Equation 85 that $\frac{\partial^2}{\partial \gamma^2} \mathsf{R}\left(\mathbf{0}, 0\right) \leq 1$.

Putting All Together. Going back to Equation 80, we may now bound the regret as follows:

$$\mathsf{R}\left(\boldsymbol{\beta}, \gamma\right) \leq \gamma^2 + \|\boldsymbol{\beta}\|_2^2 + |\gamma \boldsymbol{\beta}_1| + \mathcal{O}\left(\left(\gamma^2 + \|\boldsymbol{\beta}\|^2\right)^{3/2}\right) \quad .$$

□

Next, we use the following lemma to prove Corollary 5.

Lemma 16 ([Efr75, Lemma 2])**.** *The estimation error of $\boldsymbol{\beta}, \gamma$ satisfies the following convergence in law:*

$$\sqrt{N} \begin{pmatrix} \gamma \\ \boldsymbol{\beta} \end{pmatrix} \xrightarrow[N \to \infty]{L} \mathcal{N}(\boldsymbol{O}, \Sigma) \quad , \tag{86}$$

where $\mathcal{N}$ denotes the normal distribution centered in the origin, and a diagonal covariance matrix with coefficients $\left(1 + \frac{\Delta^2}{4}, 1 + \frac{\Delta^2}{2}, 1 + \frac{\Delta^2}{4}, \ldots, 1 + \frac{\Delta^2}{4}\right)$.

Proof of Corollary 5. Using Lemma 16, we know that for any δ such that $0 < \delta < 1$, there exists $\alpha_\delta > 0$ and N_δ such that

$$\Pr\left(\forall 0 \leq i \leq D : |\boldsymbol{\beta}_i| \leq \alpha_\delta \sqrt{\frac{\Delta^2 + 1}{N}}\right) \geq \delta$$

for all $N \geq N_\delta$ and for any true distribution parameters μ_0, μ_1 and Σ. It follows that, with probability at least δ,

$$\gamma^2 + \|\boldsymbol{\beta}\|_2^2 + |\gamma \boldsymbol{\beta}_1| \leq \alpha_\delta^2 \left(\Delta^2 + 1\right) \frac{D+1}{N}$$

and

$$\mathcal{O}\left(\left(\gamma^2 + \|\boldsymbol{\beta}\|_2^2\right)^{3/2}\right) \subset \mathcal{O}\left(\alpha_\delta^2 \left(1 + \frac{\Delta^2}{4}\right) \frac{D+1}{N}\right) \quad .$$

Using Theorem 15 and considering a constant δ gives the final result. □

References

[AB02] Martin Anthony and Peter L. Bartlett. *Neural Network Learning - Theoretical Foundations.* Cambridge University Press, 2002.

[ABB+20] Melissa Azouaoui, Davide Bellizia, Ileana Buhan, Nicolas Debande, Sébastien Duval, Christophe Giraud, Éliane Jaulmes, François Koeune, Elisabeth Oswald, François-Xavier Standaert, and Carolyn Whitnall. A systematic appraisal of side channel evaluation strategies. *IACR Cryptol. ePrint Arch.*, page 1347, 2020.

[ABCH97] Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM*, 44(4):615–631, 1997.

[AK01] András Antos and Ioannis Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Structures & Algorithms*, 19(3-4):163–193, 2001.

[And03] T.W. Anderson. *An Introduction to Multivariate Statistical Analysis.* Wiley Series in Probability and Statistics. Wiley, 2003.

[APSQ06] Cédric Archambeau, Eric Peeters, François-Xavier Standaert, and Jean-Jacques Quisquater. Template attacks in principal subspaces. In *CHES*, volume 4249 of *Lecture Notes in Computer Science*, pages 1–14. Springer, 2006.

[BCG$^+$23] Julien Béguinot, Wei Cheng, Sylvain Guilley, Yi Liu, Loïc Masure, Olivier Rioul, and François-Xavier Standaert. Removing the field size loss from duc et al.'s conjectured bound for masked encodings. In *COSADE*, volume 13979 of *Lecture Notes in Computer Science*, pages 86–104. Springer, 2023.

[BCO04] Eric Brier, Christophe Clavier, and Francis Olivier. Correlation power analysis with a leakage model. In *CHES*, volume 3156 of *Lecture Notes in Computer Science*, pages 16–29. Springer, 2004.

[BHLM19] Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *J. Mach. Learn. Res.*, 20:63:1–63:17, 2019.

[BHM$^+$19] Olivier Bronchain, Julien M. Hendrickx, Clément Massart, Alex Olshevsky, and François-Xavier Standaert. Leakage certification revisited: Bounding model errors in side-channel security evaluations. In *CRYPTO (1)*, volume 11692 of *Lecture Notes in Computer Science*, pages 713–737. Springer, 2019.

[BJP20] Shivam Bhasin, Dirmanto Jap, and Stjepan Picek. AES HD dataset - 500 000 traces. AISyLab repository, 2020. https://github.com/AISyLab/AES_HD_2.

[BS20] Olivier Bronchain and François-Xavier Standaert. Side-channel countermeasures' dissection and the limits of closed source security evaluations. *IACR Trans. Cryptogr. Hardw. Embed. Syst.*, 2020(2):1–25, 2020.

[BV14] Stephen P. Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2014.

[CCC$^+$19] Mathieu Carbone, Vincent Conin, Marie-Angela Cornelie, François Dassance, Guillaume Dufresne, Cécile Dumas, Emmanuel Prouff, and Alexandre Venelli. Deep learning to evaluate secure RSA implementations. *IACR Trans. Cryptogr. Hardw. Embed. Syst.*, 2019(2):132–161, 2019.

[CDP15] Eleonora Cagli, Cécile Dumas, and Emmanuel Prouff. Enhancing dimensionality reduction methods for side-channel attacks. In Naofumi Homma and Marcel Medwed, editors, *Smart Card Research and Advanced Applications - 14th International Conference, CARDIS 2015, Bochum, Germany, November 4-6, 2015. Revised Selected Papers*, volume 9514 of *Lecture Notes in Computer Science*, pages 15–33. Springer, 2015.

[CDP16] Eleonora Cagli, Cécile Dumas, and Emmanuel Prouff. Kernel discriminant analysis for information extraction in the presence of masking. In Kerstin Lemke-Rust and Michael Tunstall, editors, *Smart Card Research and Advanced Applications - 15th International Conference, CARDIS 2016, Cannes, France, November 7-9, 2016, Revised Selected Papers*, volume 10146 of *Lecture Notes in Computer Science*, pages 1–22. Springer, 2016.

[CDP17] Eleonora Cagli, Cécile Dumas, and Emmanuel Prouff. Convolutional neural networks with data augmentation against jitter-based countermeasures - profiling attacks without pre-processing. In *CHES*, volume 10529 of *Lecture Notes in Computer Science*, pages 45–68. Springer, 2017.

[CJRR99] Suresh Chari, Charanjit S. Jutla, Josyula R. Rao, and Pankaj Rohatgi. Towards sound approaches to counteract power-analysis attacks. In *CRYPTO*, volume 1666 of *Lecture Notes in Computer Science*, pages 398–412. Springer, 1999.

[CK13] Omar Choudary and Markus G. Kuhn. Efficient template attacks. In *CARDIS*, volume 8419 of *Lecture Notes in Computer Science*, pages 253–270. Springer, 2013.

[CK14] Marios O. Choudary and Markus G. Kuhn. Efficient stochastic methods: Profiled attacks beyond 8 bits. In *CARDIS*, volume 8968 of *Lecture Notes in Computer Science*, pages 85–103. Springer, 2014.

[CLM20] Valence Cristiani, Maxime Lecomte, and Philippe Maurine. Leakage assessment through neural estimation of the mutual information. In *ACNS Workshops*, volume 12418 of *Lecture Notes in Computer Science*, pages 144–162. Springer, 2020.

[CLZ15] T. Tony Cai, Tengyuan Liang, and Harrison H. Zhou. Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional gaussian distributions. *J. Multivar. Anal.*, 137:161–172, 2015.

[CRR02] Suresh Chari, Josyula R. Rao, and Pankaj Rohatgi. Template attacks. In *CHES*, volume 2523 of *Lecture Notes in Computer Science*, pages 13–28. Springer, 2002.

[CT12] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, 2012.

[dCGRP19] Eloi de Chérisey, Sylvain Guilley, Olivier Rioul, and Pablo Piantanida. Best Information is Most Successful. Mutual Information and Success Rate in Side-Channel Analysis. *IACR Trans. Cryptogr. Hardw. Embed. Syst.*, 2019(2):49–79, 2019.

[DFS15] Alexandre Duc, Sebastian Faust, and François-Xavier Standaert. Making masking security proofs concrete - or how to evaluate the security of any leaking device. In *EUROCRYPT (1)*, volume 9056 of *Lecture Notes in Computer Science*, pages 401–429. Springer, 2015.

[DFS19] Alexandre Duc, Sebastian Faust, and François-Xavier Standaert. Making masking security proofs concrete (or how to evaluate the security of any leaking device), extended version. *J. Cryptol.*, 32(4):1263–1297, 2019.

[DSV14] François Durvaux, François-Xavier Standaert, and Nicolas Veyrat-Charvillon. How to certify the leakage of a chip? In *EUROCRYPT*, volume 8441 of *Lecture Notes in Computer Science*, pages 459–476. Springer, 2014.

[Efr75] Bradley Efron. The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70(352):892–898, 1975.

[GLP06] Benedikt Gierlichs, Kerstin Lemke-Rust, and Christof Paar. Templates vs. stochastic methods. In *CHES*, volume 4249 of *Lecture Notes in Computer Science*, pages 15–29. Springer, 2006.

[Hau92] David Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inf. Comput.*, 100(1):78–150, 1992.

[Hau95] David Haussler. Sphere Packing Numbers for Subsets of the Boolean n-Cube with Bounded Vapnik-Chervonenkis Dimension. *J. Comb. Theory, Ser. A*, 69(2):217–232, 1995.

[HGM$^+$11] Gabriel Hospodar, Benedikt Gierlichs, Elke De Mulder, Ingrid Verbauwhede, and Joos Vandewalle. Machine learning in side-channel analysis: a first study. *J. Cryptogr. Eng.*, 1(4):293–302, 2011.

[HRG14] Annelie Heuser, Olivier Rioul, and Sylvain Guilley. Good is not good enough - deriving optimal distinguishers from communication theory. In *CHES*, volume 8731 of *Lecture Notes in Computer Science*, pages 55–74. Springer, 2014.

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition.* Springer Series in Statistics. Springer, 2009.

[HZ12] Annelie Heuser and Michael Zohner. Intelligent machine homicide - breaking cryptographic devices using support vector machines. In *COSADE*, volume 7275 of *Lecture Notes in Computer Science*, pages 249–264. Springer, 2012.

[IUH22] Akira Ito, Rei Ueno, and Naofumi Homma. Perceived information revisited new metrics to evaluate success rate of side-channel attacks. *IACR Trans. Cryptogr. Hardw. Embed. Syst.*, 2022(4):228–254, 2022.

[KB15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[KJJ99] Paul C. Kocher, Joshua Jaffe, and Benjamin Jun. Differential power analysis. In *CRYPTO*, volume 1666 of *Lecture Notes in Computer Science*, pages 388–397. Springer, 1999.

[LBM14] Liran Lerman, Gianluca Bontempi, and Olivier Markowitch. Power analysis attack: an approach based on machine learning. *Int. J. Appl. Cryptogr.*, 3(2):97–115, 2014.

[Ler] Matthieu Lerasle. Lecture Notes - Learning theory: Part I Empirical risk minimization and related fields. https://lerasle.perso.math.cnrs.fr/docs/LectureNotes3.pdf.

[LMBM13] Liran Lerman, Stephane Fernandes Medeiros, Gianluca Bontempi, and Olivier Markowitch. A machine learning approach against a masked AES. In *CARDIS*, volume 8419 of *Lecture Notes in Computer Science*, pages 61–75. Springer, 2013.

[LP07] Kerstin Lemke-Rust and Christof Paar. Gaussian mixture models for higher-order side channel analysis. In *CHES*, volume 4727 of *Lecture Notes in Computer Science*, pages 14–27. Springer, 2007.

[LPB$^+$15] Liran Lerman, Romain Poussier, Gianluca Bontempi, Olivier Markowitch, and François-Xavier Standaert. Template attacks vs. machine learning revisited (and the curse of dimensionality in side-channel analysis). In *COSADE*, volume 9064 of *Lecture Notes in Computer Science*, pages 20–33. Springer, 2015.

[Man04] Stefan Mangard. Hardware countermeasures against DPA ? A statistical analysis of their effectiveness. In *CT-RSA*, volume 2964 of *Lecture Notes in Computer Science*, pages 222–235. Springer, 2004.

[MDP20] Loïc Masure, Cécile Dumas, and Emmanuel Prouff. A comprehensive study of deep learning for side-channel analysis. *IACR Trans. Cryptogr. Hardw. Embed. Syst.*, 2020(1):348–375, 2020.

[Meh17] Nishant Mehta. Fast rates with high probability in exp-concave statistical learning. In Aarti Singh and Xiaojin (Jerry) Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pages 1085–1093. PMLR, 2017.

[MG22] Jaouad Mourtada and Stéphane Gaïffas. An improper estimator with optimal excess risk in misspecified density estimation and logistic regression. *J. Mach. Learn. Res.*, 23:31:1–31:49, 2022.

[MOBW13] Luke Mather, Elisabeth Oswald, Joe Bandenburg, and Marcin Wójcik. Does my device leak information? An a priori statistical power analysis of leakage detection tests. In *ASIACRYPT (1)*, volume 8269 of *Lecture Notes in Computer Science*, pages 486–505. Springer, 2013.

[MPP16] Houssem Maghrebi, Thibault Portigliatti, and Emmanuel Prouff. Breaking cryptographic implementations using deep learning techniques. In *SPACE*, volume 10076 of *Lecture Notes in Computer Science*, pages 3–26. Springer, 2016.

[MRS22] Loïc Masure, Olivier Rioul, and François-Xavier Standaert. A nearly tight proof of duc et al.'s conjectured security bound for masked implementations. In *CARDIS*, volume 13820 of *Lecture Notes in Computer Science*, pages 69–81. Springer, 2022.

[MS16] Amir Moradi and François-Xavier Standaert. Moments-correlating DPA. In *TISCCS*, pages 5–15. ACM, 2016.

[Pan03] Liam Paninski. Estimation of entropy and mutual information. *Neural Comput.*, 15(6):1191–1253, 2003.

[PBP21] Guilherme Perin, Ileana Buhan, and Stjepan Picek. Learning when to stop: A mutual information approach to prevent overfitting in profiled side-channel analysis. In Shivam Bhasin and Fabrizio De Santis, editors, *Constructive Side-Channel Analysis and Secure Design - 12th International Workshop, COSADE 2021, Lugano, Switzerland, October 25-27, 2021, Proceedings*, volume 12910 of *Lecture Notes in Computer Science*, pages 53–81. Springer, 2021.

[PGM$^+$19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 8024–8035, 2019.

[PHG17] Stjepan Picek, Annelie Heuser, and Sylvain Guilley. Template attack versus bayes classifier. *J. Cryptogr. Eng.*, 7(4):343–351, 2017.

[PHJ$^+$19] Stjepan Picek, Annelie Heuser, Alan Jovic, Shivam Bhasin, and Francesco Regazzoni. The curse of class imbalance and conflicting metrics with machine learning for side-channel evaluations. *IACR Trans. Cryptogr. Hardw. Embed. Syst.*, 2019(1):209–237, 2019.

[PP$^+$08] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

[PSK$^+$18] Stjepan Picek, Ioannis Petros Samiotis, Jaehun Kim, Annelie Heuser, Shivam Bhasin, and Axel Legay. On the performance of convolutional neural networks for side-channel analysis. In *SPACE*, volume 11348 of *Lecture Notes in Computer Science*, pages 157–176. Springer, 2018.

[RSV$^+$11] Mathieu Renauld, François-Xavier Standaert, Nicolas Veyrat-Charvillon, Dina Kamel, and Denis Flandre. A formal study of power variability issues and side-channel attacks for nanoscale devices. In *EUROCRYPT*, volume 6632 of *Lecture Notes in Computer Science*, pages 109–128. Springer, 2011.

[SA08] François-Xavier Standaert and Cédric Archambeau. Using subspace-based template attacks to compare and combine power and electromagnetic information leakages. In *CHES*, volume 5154 of *Lecture Notes in Computer Science*, pages 411–425. Springer, 2008.

[SB14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning - From Theory to Algorithms.* Cambridge University Press, 2014.

[SKS09] François-Xavier Standaert, François Koeune, and Werner Schindler. How to compare profiled side-channel attacks? In *ACNS*, volume 5536 of *Lecture Notes in Computer Science*, pages 485–498, 2009.

[SLP05] Werner Schindler, Kerstin Lemke, and Christof Paar. A stochastic model for differential side channel cryptanalysis. In *CHES*, volume 3659 of *Lecture Notes in Computer Science*, pages 30–46. Springer, 2005.

[SM16] Tobias Schneider and Amir Moradi. Leakage assessment methodology - extended version. *J. Cryptogr. Eng.*, 6(2):85–99, 2016.

[SMY09] François-Xavier Standaert, Tal Malkin, and Moti Yung. A unified framework for the analysis of side-channel key recovery attacks. In *EUROCRYPT*, volume 5479 of *Lecture Notes in Computer Science*, pages 443–461. Springer, 2009.

[SPAQ06] François-Xavier Standaert, Eric Peeters, Cédric Archambeau, and Jean-Jacques Quisquater. Towards security limits in side-channel attacks. In *CHES*, volume 4249 of *Lecture Notes in Computer Science*, pages 30–45. Springer, 2006.

[Sto82] Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.

[Sto83] Charles J. Stone. Optimal uniform rate of convergence for nonparametric estimators of a density function or its derivatives. In M. Haseeb Rizvi, Jagdish S. Rustagi, and David Siegmund, editors, *Recent Advances in Statistics*, pages 393–406. Academic Press, 1983.

[Vap98] Vladimir Vapnik. *Statistical Learning Theory.* Wiley, 1998.

[vEGM$^+$15] Tim van Erven, Peter D. Grünwald, Nishant A. Mehta, Mark D. Reid, and Robert C. Williamson. Fast rates in statistical and online learning. *J. Mach. Learn. Res.*, 16:1793–1861, 2015.

[Ver12] Roman Vershynin. How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 25(3):655–686, 2012.

[WAGP20] Lennert Wouters, Victor Arribas, Benedikt Gierlichs, and Bart Preneel. Revisiting a methodology for efficient CNN architectures in profiling attacks. *IACR Trans. Cryptogr. Hardw. Embed. Syst.*, 2020(3):147–168, 2020.

[WOS14] Carolyn Whitnall, Elisabeth Oswald, and François-Xavier Standaert. The myth of generic dpa...and the magic of learning. In *CT-RSA*, volume 8366 of *Lecture Notes in Computer Science*, pages 183–205. Springer, 2014.

[ZBD$^+$21] Gabriel Zaid, Lilian Bossuet, François Dassance, Amaury Habrard, and Alexandre Venelli. Ranking loss: Maximizing the success rate in deep learning side-channel analysis. *IACR Trans. Cryptogr. Hardw. Embed. Syst.*, 2021(1):25–55, 2021.

[ZBHV20] Gabriel Zaid, Lilian Bossuet, Amaury Habrard, and Alexandre Venelli. Methodology for efficient CNN architectures in profiling attacks. *IACR Trans. Cryptogr. Hardw. Embed. Syst.*, 2020(1):1–36, 2020.