Information Bounds and Convergence Rates for Side-Channel Security Evaluators

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

Crypto Group, ICTEAM, UCLouvain, Louvain-la-Neuve, Belgium
firstname.lastname@uclouvain.be

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 (with a given choice of set of models) 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.

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 [8], a broad literature has focused on analyzing physical leakage in order to attack efficiently and to assess 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 [8], it has been shown that profiling the target device (*i.e.*, leveraging an open sample to estimate a leakage model) 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" [17]. As a result, Standaert *et al.* observed that profiled attacks are critical to estimate the worst-case security of an implementation [55]. 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) [62]. Heuser *et al.* then clarified the meaning of an optimal adversary in the information theoretic sense as the one distinguishing thanks to the probability distribution of the leakage conditioned on the targeted secret [31]. These advances consolidated the now standard approach of assessing the security level of cryptographic implementation thanks to information theoretic metrics such as the Mutual Information (MI), which can be used to bound the data complexity of worst-case attacks [23,22].

A second step towards sound side-channel security evaluations is the acknowledgment that even in the profiled setting, performing an optimal attack in the sense of Heuser *et al.*, or equivalently estimating the MI, 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, especially in the presence of countermeasures like masking [16]. This 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 model, possibly biased by estimation and assumption errors [49]. As a result, one can summarize the evaluation problem in two questions:

1. What is the complexity of the best (ideally optimal) online attack?
2. What is the complexity of estimating its model, with profiling?

Here, both complexities are defined in terms of number of measurement traces collected. 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. Durvaux *et al.* therefore formalized the problem of leakage certification as the one of assessing the distance between an evaluator's best attack and the optimal one, or equivalently the distance between the PI and the MI [25]. Bronchain *et al.* showed that the PI is in general a lower bound for the MI and that an upper bound is obtained by estimating the Hypothetical Information (HI), which is the amount of information that would be extractable from a device if the true distribution was the model, *for the empirical distribution* [9]. They additionally showed that the expected value of the resulting empirical Hypothetical Information (eHI) asymptotically converges towards the MI. Unfortunately, the practical impact of these results is limited since the profiling complexity of multivariate attacks quickly becomes unrealistic for the empirical leakage model. 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 such 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) [17,50,27,5,53,54,19,18] and a steadily increasing (and not exhaustive) list of works leveraging machine (and deep) learning [33,32,36,35,37,38,13,15,65,63,64]. 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 [40]. 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 statistical 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; it only is when the assumptions exploited by the parametric model are met by the true leakage distribution (which was the case in the experiments of [9]). 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 eliminate 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 the 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., $\mathsf{LI} - \mathsf{PI}$) 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, it answers the question: how much information can be learned with my model?

Regarding the second question, we initiate a study of the convergence rate of the gap between TI_N and PI for practically-relevant profiling techniques, such as Gaussian template attacks [17] (denoted in this paper as gTA), their variant with *pooled* covariance matrix estimation [19] (denoted by $\mathsf{p\text{-}gTA}$), logistic regression (denoted by LR_1 and LR_k), and deep neural networks such as Multi-Layer Perceptron (MLP) with L layers and W weights to fit. Our results are synthesized in Table 1. Here, Q denotes the number of profiled classes, D denotes the dimensionality of the observed traces, and N denotes the number of traces acquired on the profiling device, *i.e.*, quantifying the *sample complexity* of profiling.

Table 1: Convergence of 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 perfectly matches the true leakage distribution.

Model	Noise assumption	Attack order	Fast regime	General bound
MLP	None	any	$\widetilde{\mathcal{O}}\left(\frac{QWL}{N}\right)$	$\widetilde{\mathcal{O}}\left(\sqrt{\frac{QWL}{N}}\right)$
LR_k	None	any	$\widetilde{\mathcal{O}}\left(\frac{QD^k}{N}\right)$	$\tilde{\mathcal{O}}\left(\sqrt{\frac{Q\cdot D^k}{N}}\right)$
LR_1	Exponential	1^{st}	$\widetilde{\mathcal{O}}\left(\frac{QD}{N}\right)$	$\tilde{\mathcal{O}}\left(\sqrt{\frac{Q\cdot D}{N}}\right)$
Template Attacks	Gaussian	2^{nd} : gTA	$\mathcal{O}\left(\frac{QD^2}{N}\right)$	
		1^{st} : $\mathsf{p\text{-}gTA}$	$\mathcal{O}\left(\frac{QD}{N}\right)$ for $Q=2$	

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 $\sqrt{N}$ for any (even protected) implementation. It also shows that the convergence rates of the investigated models do not depend on physical leakage characteristics and consolidates the general intuition that side-channel security evaluations represent a trade off between the genericity and the efficiency of the profiling. For example, assuming that the true leakage distribution p is in the model hypothesis class $\mathcal{H}$ may allow faster guaranteed convergence rates (with a modeling error inversely proportional to N); and further leveraging a Gaussian assumption allows performing simple TA that do not require solving an optimization problem. The convergence rates also allow an evaluator to anticipate the number of profiling traces needed to guarantee a certain model quality.

On the other hand, it shows that there are statistical tools that are better suited depending on the contexts. For example, the convergence rate of logistic regression generalized to 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 encoding 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 $\mathsf{poly}(D = k)$ [52, Thm. 20.3]. So our results suggest that if an evaluator has to profile higher-order leakages, leveraging MLPs leads to a more efficient profiling than trying to estimate a specific moment of the leakage distribution.

To synthesize, as side-channel evaluators usually operates within a limited time frame, anticipating and optimizing such complexity is therefore crucial, and we hope that our work will help the practitioner to find the good trade-offs in its decisions.

1.1 Related Works

The use of information theoretic metrics to guide/compare profiled attacks dates back to [54]. In a work from Cosade 2021 [46], 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.

In a recent ePrint report, Ito *et al.* show that the direct optimization of security metrics such as the Success Rate (SR) or Guessing Entropy (GE) [55] can slightly improve an optimization guided by information theoretic metrics in some contexts, at the cost of some computational overheads [34]. 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) [56,48]. Yet, since information theoretic metrics are inversely proportional to the asymptotic complexity of a side-channel attack's online part, the concrete impact of such an observation is also limited. For example, the experiments performed in [34] 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, the study of Cristiani *et al.* investigates the so-called *Neural-based* MI *estimation* (MINE) [21]. It leverages the variational formulation of the MI allowing to train an MLP to maximize a lower bound of the MI, similarly to the PI [20, Eq. (8.93)]. At high level, this research follows the observation of Mather *et al.* [41] that an evaluator may choose to estimate the complexity of his 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

2.1 Notations

In the following, we denote random variables (respectively random vectors) by upper-case (respectively bold upper-case) letters X (respectively $\boldsymbol{X}$). We denote by the same calligraphic letter $\mathcal{X}$ the observation domain of the corresponding random variable (respectively random vector). We denote observations of a random variable (respectively random vector) by the corresponding lower-case roman letter x (respectively $\boldsymbol{x}$). 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 $\mathsf{D}_{\mathsf{KL}}(\mathsf{p} \parallel \mathsf{m}) = \mathop{\mathbb{E}}_{X \sim \mathsf{p}} \left[\frac{\mathsf{p}(X)}{\mathsf{m}(X)} \right]$. In this paper, we use the notation $\mathcal{O}(f(n))$ to hide constant factors in n, whereas we use 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.

2.2 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 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 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 two probability distributions or models p and m:

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

where $\mathsf{H}(Y)$ is the entropy of Y. We remark that this distance uses p as a generative model, while it uses m as a discriminative model only. 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 $\mathsf{MI}(Y; \boldsymbol{L})$ [24,22]. 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 [45]. 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 [9] additionally considered the Hypothetical Information (HI):

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

and the empirical Hypothetical Information (eHI) 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:

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

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

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

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

3 Limitations of the HI

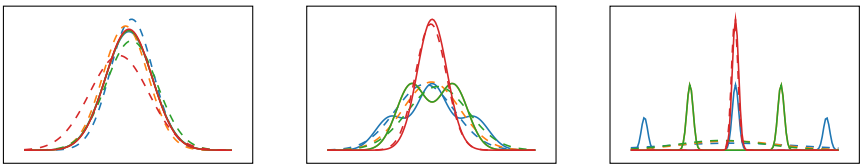
One important question left open by Bronchain *et al.* is whether the properties of the HI generalize to parametric leakage models. This question is important since, as experimentally observed in [9], assessing the security of an implementation with an empirical model (and the corresponding bounds) rapidly becomes too expensive. In this section, we consolidate this HI proposal in two directions. First, we give a counter-example confirming that the HI is in general (*i.e.*, for any model) an upper bound for the PI. It turns out 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 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 phase 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 [9], 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: either they both converge to the same 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 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 [39].

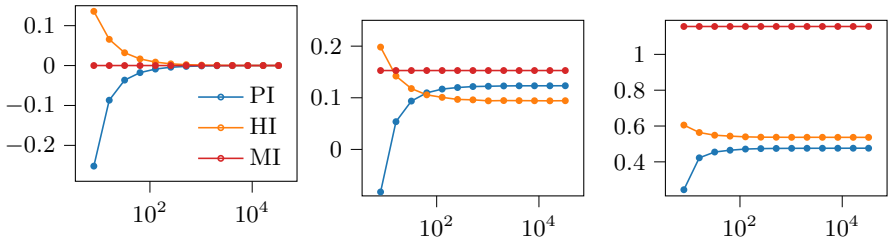


(a) SNR = 0.02 (b) SNR = 2 (c) SNR = 200

Fig. 1: True distributions (continuous lines) and models (dashed lines) trained with 20 samples for each of the 4 classes (*i.e.* $n = 2$ bits).

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).



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

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

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.

Convergence of the Expectation. We first state that the bias of eHI scales exponentially with the dimensionality of the traces D and linearly with $\frac{Q}{N}$, with Q being 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, assumed to be uniformly distributed. Then, the* eHI *satisfies the following inequalities:*

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

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

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

The proof of this statement is directly inspired from Paninski's work [45], 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 [45, 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 [9].[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, and according to different 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.

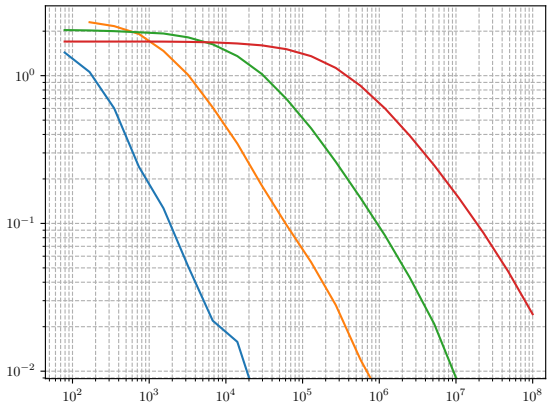


Fig. 3: $\mathsf{eHI} - \mathsf{MI}$ (y-axis) with respect to the number of profiling traces N (x-axis) for $D = 1$ (blue), 2 (orange), 3 (green), and 4 (red). Here, $\omega = 4$ and $Q = 16$.

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.

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

We next eliminate this 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 [9, 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 [45]. 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 $\Delta^{\mathsf{m}}_{\mathsf{p}}$, 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 $\Delta^{\mathsf{m}}_{\tilde{\mathsf{e}}_N}$, 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 [25]. 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 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. 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 following 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*

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

i.e., a mapping 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 profiling 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 value and 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.

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*

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

For short, 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 now possible to derive properties similar to the ones conjectured for the gHI by Bronchain *et al.* [9]. 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. Together, 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*

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

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

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

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

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

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

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

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*

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

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

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

We shall show that the function $\gamma : \tilde{\mathsf{e}}_N \mapsto \Delta^{\widehat{\mathsf{m}}_{\tilde{\mathsf{e}}_N}}_{\tilde{\mathsf{e}}_N}$ is convex. The theorem then follows from Lemma 2 of Bronchain *et al.* [9].

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 $\Delta^{\mathsf{m}}_{\mathsf{e}}$ with respect to e, we have

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

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

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

□

Proposition 1 and Proposition 2 together show that the TI_N satisfies the same monotone convergence of its expectation than the one satisfied by the eHI, as previously shown by Bronchain *et al.* [9]. 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, and 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

In this Section, we show that under some assumptions that we give hereafter, thereafter, the TI_N converges towards the LI for some classes of TI_N-maximizing distinguishers. Furthermore, we provide bounds on the rate of this convergence.

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

$$\sigma(\boldsymbol{x}) = \frac{1}{\sum_{i=1}^{Q} e^{\boldsymbol{x}_i}} \begin{pmatrix} e^{\boldsymbol{x}_1} \\ \vdots \\ e^{\boldsymbol{x}_Q} \end{pmatrix}, \boldsymbol{x} \in \mathbb{R}^Q \ . \tag{16}$$

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

$$\mathsf{m}_{\boldsymbol{\Theta}}(\boldsymbol{l}) = \sigma \begin{pmatrix} f(\boldsymbol{l}; \boldsymbol{\theta}_1) \\ \vdots \\ f(\boldsymbol{l}; \boldsymbol{\theta}_Q) \end{pmatrix} \ , \tag{17}$$

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

Remark 1. It is noticeable that 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

$$\boldsymbol{l}, \boldsymbol{\theta}_i \mapsto f(\boldsymbol{l}; \boldsymbol{\theta}_i) = B_i^{\mathsf{T}} \boldsymbol{l}' \tag{18}$$

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

$$\boldsymbol{l}, \boldsymbol{\theta}_i \mapsto f(\boldsymbol{l}; \boldsymbol{\theta}_i) = \boldsymbol{l}'^{\mathsf{T}} A_i \boldsymbol{l}', \tag{19}$$

where $A_i \in \mathbb{R}^{(D+1)^2}$ is a quadratic form. Here, $\boldsymbol{\theta}_i = A_i$.

In the case of MLP s, the mapping

$$\boldsymbol{l}, \boldsymbol{\theta}_i \mapsto f(\boldsymbol{l}; \boldsymbol{\theta}_i) = \phi_L\left(\cdot; \boldsymbol{\Theta}_i^{(L)}\right) \circ \ldots \circ \phi_1\left(\cdot; \boldsymbol{\Theta}_i^{(1)}\right)(\boldsymbol{l}) \tag{20}$$

is a composition of L *layers* ϕ_i, each being in turn 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 remainder of this paper, we assume that the total number of entries in the weight matrices equals W.

Whereas MLP s are nowadays widely used for profiled side-channel analysis, LR models have never 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 [19]) is included in LR_2 (resp. LR_1). This will be shown in the devoted Section 6. A similar correspondence could be investigated for the inclusion of so-called side-channel attacks of order k [51,43] 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 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 [3, Def. 11.1]). *Let $\mathcal{F}$ be a set of functions mapping from a domain $\mathcal{L}$ to $\mathbb{R}$ and suppose that $\mathcal{S}_N = \{\boldsymbol{l}_1, \ldots, \boldsymbol{l}_N\} \subset \mathcal{L}$ for some positive integer N. Then, $\mathcal{S}_N$ is* pseudo-shattered *by $\mathcal{F}$ if there are real numbers $r_1, \ldots, r_N$ such that for all $\boldsymbol{b} \in \{0,1\}^N$ there is a function $f_{\boldsymbol{b}} \in \mathcal{F}$ such that for all $1 \leq i \leq N$,*

$$f_{\boldsymbol{b}}(\boldsymbol{l}_i) \begin{cases} \leq r_i \text{ if } \boldsymbol{b}_i = 0 \\ > r_i \text{ if } \boldsymbol{b}_i = 1 \end{cases} . \tag{21}$$

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

An example of pseudo-shattering is depicted in Figure 4. 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 21 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 21 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 21.

[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.

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.

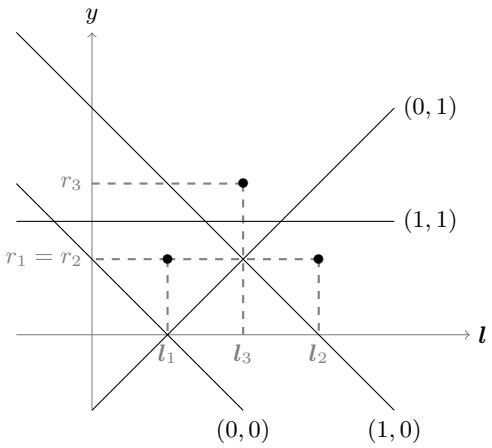


Fig. 4: Illustration of the pseudo-shattering by the set $\mathcal{F}$ of affine functions of $\mathcal{L} = \mathbb{R}$. The tuples denote the different values of $\boldsymbol{b}$. $\{\boldsymbol{l}_1, \boldsymbol{l}_2\}$ is pseudo-shattered by $\mathcal{F}$, while $\{\boldsymbol{l}_1, \boldsymbol{l}_2, \boldsymbol{l}_3\}$ is not.

Definition 8 (Pseudo-dimension [3, 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 $\mathcal{F}$ [3, 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 [3, Thm. 11.8]). *Let $\mathcal{F}$ be the class of all polynomial transformations on $\mathbb{R}^D$ of degree at most k. Then*

$$\mathsf{P}_{\mathsf{dim}}(\mathcal{F}) = \binom{D+k}{k} \ . \tag{22}$$

Theorem 4 (Pseudo-dimension of MLP [6]). *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*

$$cWL\log(W/L) \leq \mathsf{P}_{\mathsf{dim}}(\mathcal{F}) \leq CWL\log(W) \ . \tag{23}$$

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.[3]

5.3 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 $\mathsf{P_{dim}}$. Define the following quantities:*

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

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

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

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

$$\sup_{\mathsf{m}\in\mathcal{H}} \left|\Delta^{\mathsf{m}}_{\widetilde{\mathsf{e}}_N} - \mathsf{PI}(Y;\boldsymbol{L};\mathsf{m})\right| \leq \epsilon_{\mathsf{P_{dim}},Q,V,N,\delta} \tag{24}$$

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

We prove Theorem 5 in Section B. Corollary 1 follows from this result.

Corollary 1. *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 $\mathsf{P_{dim}}$. The following inequalities*

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

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

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 12 (proven in Appendix), while the two last ones follow from Corollary 8. □

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

[3] 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 [60, pp. 159-160].

5.4 Faster Rates for Well-Specified Hypothesis Classes

So far we have emphasized that the convergence rate between TI_N and PI scales with $1/\sqrt{N}$. It is actually possible to emphasize *faster* rates – *i.e.*, asymptotically linear with $1/N$. Until a few years ago, some particular cases were emphasized, where fast rates could indeed hold, provided a few additional assumptions. Those assumptions required the learner to be in a *realizable* case.[4] Whereas in other application fields of machine learning, *e.g.*, in image recognition, this assumption likely holds, such an assumption is not realistic in side-channel analysis. Indeed, when specialized to the case of the information as a loss function, the realizable assumption implies that the MI between the leakage $\boldsymbol{L}$ and the sensitive target variable Y should be equal to its maximum value, namely the entropy of Y. This means that the device under evaluation can be broken in one query with probability one by an optimal adversary. Since the aim of a developer is to increase the security level of the target under evaluation, such attack outcomes are not expected. This drastically reduces the scope of profiling scenarios where the realizability assumption may hold. Hopefully, some recent advances in statistical learning theory have seen the emergence of relaxed alternative assumptions, unified under the name of *central condition* [59]. We present hereafter one such result that we will use to derive fast rates.

Theorem 6 ([42, 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*

$$\mathsf{MI}(Y;\boldsymbol{L}) - \mathsf{PI}(Y;\boldsymbol{L};\widehat{\mathsf{m}}_N) \leq \frac{1}{N} 8B \left(\mathsf{P}\log(16UTN) + \log\left(\frac{1}{\delta}\right)\right) + \frac{1}{N} \ . \quad (25)$$

The condition $\mathsf{p} \in \mathcal{H}$ can be seen as a relaxation of the realizable condition, since it is verified in the realizable case, but no longer requires any assumption on the MI of the true leakage model. Based on Theorem 6, we derive faster rates for the different hypothesis classes we consider in this section.

Corollary 2. *Let LR_1 be a TI_N-maximizer attacker using logistic regression 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 $1 \leq i \leq Q$, $\|\boldsymbol{\theta}_i\|_2 \leq S$, for some $S \in \mathbb{R}$.*

If the true leakage belongs to the hypothesis class of LR_1 and $N \geq 5$, then, denoting $h = \log\left(32QSN\sqrt{R^2+1}\right)$, the regret can be upper bounded by

$$\mathsf{R}\left(\mathsf{LR}_1\right) \leq \frac{8}{N}\left(2\sqrt{R^2+1}S + \log(Q)\right)\left((D+1)Qh + \log\left(\frac{1}{\delta}\right)\right) + \frac{1}{N} \quad (26)$$

[4] The terminology differs from one book to another. The "realizable" terminology is used by Shalev-Shwartz & Ben-David [52], whereas Vapnik uses the term "optimistic case" [60], and Anthony & Bartlett use the term "restricted class" [52].

In other words, the regret of an LR_1 attacker is bounded by $\widetilde{\mathcal{O}}\left(\frac{SDQ}{N}\right)$.

Remark 2. The condition $\mathsf{p} \in \mathcal{H}$ is a sufficient but not necessary condition to establish the central condition, and thereby fast rates of convergence. This suggests that convergence rates asymptotically linear with $\frac{1}{N}$ may be obtained even if $\mathsf{p} \notin \mathcal{H}$. Unfortunately, this may also come at the price of prohibitive constants. As an example, for LR_1 with only $Q = 2$ and without assuming $\mathsf{p} \in \mathcal{H}$, it is not possible to derive tighter upper bounds than $\mathcal{O}\left(\frac{e^{RS}}{N}\right)$ [44].

Corollary 3. *Let LR_2 be a TI_N-maximizer attacker using logistic regression 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 $1 \leq i \leq Q$, $\|\boldsymbol{\theta}_i\|_2 \leq S$, for some $S \in \mathbb{R}$.*

If the true leakage belongs to the hypothesis class of LR_2 and $N \geq 5$, then, denoting $h = \log\big(32QSN(R^2+1)\big)$, the regret can be upper bounded by

$$\mathsf{R}\left(\mathsf{LR}_2\right) \leq \frac{8}{N}\left(2(R^2+1)S + \log(Q)\right)\left((D+1)^2 Q h + \log\left(\frac{1}{\delta}\right)\right) + \frac{1}{N} \tag{27}$$

Corollary 4. *Let $\mathcal{A}$ be a TI_N-maximizer attacker using MLP as defined in Equation 20 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 $S \in \mathbb{R}_{\geq 1}$.*

Suppose that the true leakage belongs to the hypothesis class of MLP. Then, the regret can be upper bounded by

$$\mathsf{R}\left(\mathsf{MLP}\right) \leq \frac{8B}{N}\left(WQ\log(16BN) + \log\left(\frac{1}{\delta}\right)\right) + \frac{1}{N}\ , \tag{28}$$

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

Theorem 6, along with Corollaries 2, 3, and 4, are proven in Section C.

6 Gaussian Templates

The assumption $\mathsf{p} \in \mathcal{H}$, which is key to obtaining the fast convergence rate of the previous section, is actually a fairly common assumption made for side-channel security evaluations. One of the most popular models is the Gaussian template where $\mathcal{F}$ is the set of multivariate Gaussian distributions. 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.

In this section, we compute the convergence rates of gTA, first for the original and most generic template attack [17], then in the particular case where the

covariance matrix is diagonal (*i.e.* the white noise case), and finally for the pooled gTA (*i.e.* the covariance is the same for all values of y) [19].

Formally, we assume in this section 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 by the attacker to mount the key recovery.

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 model $\mathsf{m}_{\boldsymbol{\Theta}}$ coincides with the Gaussian template. Likewise, if we further assume that the covariance matrix is the same for all 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 covered by the hypothesis class LR_1.[5] In other words, despite a gTA (resp. $\mathsf{p\text{-}gTA}$) 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 the convergence rates of both approaches, *e.g.* by comparing their respective regrets. It follows from the Gaussian distribution assumption that $\mathsf{LI} = \mathsf{MI}$. Accordingly, the regret is bounded if and only if the gap between the LI and the PI is bounded. This is the aim of this section.

Remark 3. The Gaussian TA (resp. pooled TA) is actually 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.* [26,28]), and their results are 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 can be applied to the PI.

6.1 General bound

The first theorem[6] presented hereafter uniformly bounds the regret of gTA with the regret induced by an imperfect characterization of the true distribution.

[5] Actually, those inclusions of hypothesis sets are not tight, as argued by Efron who tells that the set LR_1 could coincide with the set of template attacks with exponential family distribution sets, with common nuisance parameter [26].

[6] We prove the claims of this section in Section D.

Theorem 7. *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.*

$$0 \leq \mathsf{R}\left(\mathsf{TA}\right) \leq \frac{1}{Q}\sum_{y} \mathsf{D}_{\mathsf{KL}}\Big(\mathsf{f}_y(\cdot) \,\Big\|\, \widehat{\mathsf{f}}_y(\cdot)\Big) \leq \max_{y} \mathsf{D}_{\mathsf{KL}}\Big(\mathsf{f}_y(\cdot) \,\Big\|\, \widehat{\mathsf{f}}_y(\cdot)\Big) \tag{29}$$

Note that Theorem 7 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 1. *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,*

$$\mathsf{D}_{\mathsf{KL}}(\mathsf{f}_{\boldsymbol{X}}(\cdot) \parallel \mathsf{f}_{\boldsymbol{Y}}(\cdot)) = \mathsf{D}_{\mathsf{KL}}(\mathsf{f}_{A\cdot\boldsymbol{X}+\boldsymbol{b}}(\cdot) \parallel \mathsf{f}_{A\cdot\boldsymbol{Y}+\boldsymbol{b}}(\cdot)) \ . \tag{30}$$

For Gaussian templates, we can therefore reduce the study of the KL divergence of Theorem 7 to the particular case where the true covariance matrix Σ is the identity using Lemma 1. Furthermore, in the case of gTA with $\Sigma = I$, the following lemma gives an algebraic formulation of the upper bound.

Lemma 2. *For a Gaussian distribution with* $\Sigma = I$*, the KL divergence is given by:*

$$2\mathsf{D}_{\mathsf{KL}}\Big(\mathsf{f}(\cdot) \,\Big\|\, \widehat{\mathsf{f}}(\cdot)\Big) = \log\Big(\det\Big(\widehat{\Sigma}\Big)\Big) + \mathrm{Tr}\Big(\widehat{\Sigma}^{-1}\Big) - D \tag{31}$$

$$+ \,(\widehat{\mu}-\mu)^{\intercal}\,\widehat{\Sigma}^{-1}\,(\widehat{\mu}-\mu) \ . \tag{32}$$

We prove Lemma 2 in Section D. We now bound each term of Lemma 2.

Bounding Equation 32. The term (32) is the well known Hotelling's T^2 statistic, as recalled by the following lemma.

Lemma 3 ([2, Thm. 5.2.2]). *For* $N/Q \geq D$*, the quantity*

$$\frac{N}{QD}\frac{N/Q-D}{N/Q-1}\cdot(\widehat{\mu}-\mu)^{\intercal}\widehat{\Sigma}^{-1}(\widehat{\mu}-\mu) \tag{33}$$

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 (32) belongs to $\mathcal{O}\Big(\frac{DQ}{N}\Big)$.

Bounding Equation 31. The terms of Equation 31 are upper bounded in the following theorem, proved in Appendix D.

Theorem 8. *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* $N \geq 4C^2\log\big(\frac{2}{\delta}\big)\,D$ *the second following inequality holds with probability at least* $1-\delta$*:*

$$0 \leq \log\Big(\det\Big(\widehat{\Sigma}\Big)\Big) + \mathrm{Tr}\Big(\widehat{\Sigma}^{-1}\Big) - D \leq 2C\log\left(\frac{2}{\delta}\right)\frac{QD^2}{N} \ . \tag{34}$$

Going back to practical considerations, we summarize Theorem 8 by the following corollary.

Corollary 5. *The regret* $\mathsf{R}\left(\mathit{gTA}\right)$ *of an attacker instantiating a Gaussian template attack is upper-bounded by* $\mathcal{O}\Big(\frac{QD^2}{N}\log\big(\frac{2}{\delta}\big)\Big)$.

Proof. Comparing the bounds in Lemma 3 and Theorem 8, we can see that Hotelling's T^2 statistic can be neglected. □

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. So far, both parameters D and Q were controlled by the evaluator for computational complexity reasons, since the run-time and memory complexity of running a gTA also scales with $\mathcal{O}\big(QD^3\big)$. In this perspective, Corollary 5 shows that beyond the computational complexity, controlling both parameters is also a matter of *profiling* complexity.

6.2 The General Bound is Tight

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 5 essentially comes from the error terms due to the estimation of the empirical covariance matrix, namely $\log\Big(\det\Big(\widehat{\Sigma}\Big)\Big)$ and $\mathrm{Tr}\Big(\widehat{\Sigma}^{-1}\Big)-D$. Hereafter, we show that the sum of both error terms scale with $\Theta\Big(\frac{QD^2}{N}\Big)$ in expectation.

Theorem 9 ([14, 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*

$$\frac{1}{\sqrt{2QD/N}}\left(\log\left(\frac{\det\Big(\widehat{\Sigma}\Big)}{\det(\Sigma)}\right)-QD(D+1)/(2N)\right)\xrightarrow[N\to\infty]{L}\mathcal{N}(0,1)\ . \tag{35}$$

Theorem 9 is an analogue of the Central-Limit Theorem for the log-det term with a $\Theta\Big(\frac{QD^2}{N}\Big)$ positive bias. The following term shows that the bias from the trace of inverse covariance matrix is positive.

Lemma 4. *The trace of the inverse empirical covariance matrix is positively biased:*

$$\mathbb{E}\left[\mathrm{Tr}\Big(\widehat{\Sigma}^{-1}\Big)-D\right]\geq 0\ . \tag{36}$$

Therefore, the latter bias cannot compensate the former one, which proves the tightness of our bounds in the general case. 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.

6.3 The Covariance Matrix is Diagonal: Naive Bayes.

Assuming a Gaussian multivariate distribution with diagonal covariance matrix for the true leakage function reduces the covariance estimation to the estimation of the variance in each dimension. As a result, the convergence is faster.

Theorem 10. *Assume that $\Sigma = I$ and $\widehat{\Sigma}$ is a diagonal matrix. Then, for all $\delta > 0$ the following inequality holds:*

$$0 \leq \log\left(\det\left(\widehat{\Sigma}\right)\right) + \mathrm{Tr}\left(\widehat{\Sigma}^{-1}\right) - D \leq C \log\left(\frac{2}{\delta}\right) \frac{DQ}{N} \quad . \tag{37}$$

Sketch. 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 37 can be reduced to estimate the sum of D error terms, each for one-dimensional covariance matrices. Therefore, using Equation 34 in the particular case where $D = 1$, and multiplying by the true dimensionality D gives the result. □

Corollary 6. *The regret* $\mathsf{R}\,(\mathsf{diag\text{-}gTA})$ *of an attacker instantiating a Gaussian template attack knowing that the covariance matrices are all diagonal is upper-bounded by* $\mathcal{O}\left(\frac{QD}{N}\log\left(\frac{2}{\delta}\right)\right)$.

6.4 Choudary and Kuhn's Pooled Template Attacks.

For gTA-based side-channel attacks, the bottleneck task is the estimation of the covariance matrices. In particular, 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 [19]. 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 [10,37,11,12], 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. Therefore, can we find another explanation to the success of p-gTA? Intuitively, using Q times more traces to estimate the pooled covariance matrix would induce a $\mathcal{O}\left(\frac{D^2}{N}\right)$ bound in Theorem 8, and thereby a $\mathcal{O}\left(\frac{D\max\{D,Q\}}{N}\right)$ bound in Corollary 5 for the ultimate regret of pooled template attacks.

However, we conjecture that the latter upper bound can be even tightened to $\mathcal{O}\left(\frac{QD}{N}\right)$, becoming fully linear in the trace dimensionality, despite the D^2 matrix coefficients to estimate. Our conjecture is grounded on a proof in the particular case where $Q = 2$.

Theorem 11. *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*

$$\boldsymbol{\beta} = \widehat{\Sigma}^{-1}\left(\widehat{\mu}_1 - \widehat{\mu}_0\right) \qquad -\Sigma^{-1}\left(\mu_1 - \mu_0\right) \quad , \tag{38}$$

$$\gamma = -\tfrac{1}{2}\left(\widehat{\mu}_1\widehat{\Sigma}^{-1}\widehat{\mu}_1 - \widehat{\mu}_0\widehat{\Sigma}^{-1}\widehat{\mu}_0\right) + \frac{1}{2}\left(\mu_1\Sigma^{-1}\mu_1 - \mu_0\Sigma^{-1}\mu_0\right) \quad . \tag{39}$$

Then, the regret of p-gTA satisfies

$$\mathsf{R}\left(\textit{p-gTA}\right) \leq \left(\gamma^2 + \|\boldsymbol{\beta}\|_2^2 + |\gamma\boldsymbol{\beta}_1|\right) + \mathcal{O}\left(\left(\gamma^2 + \|\boldsymbol{\beta}\|_2^2\right)^{3/2}\right) \tag{40}$$

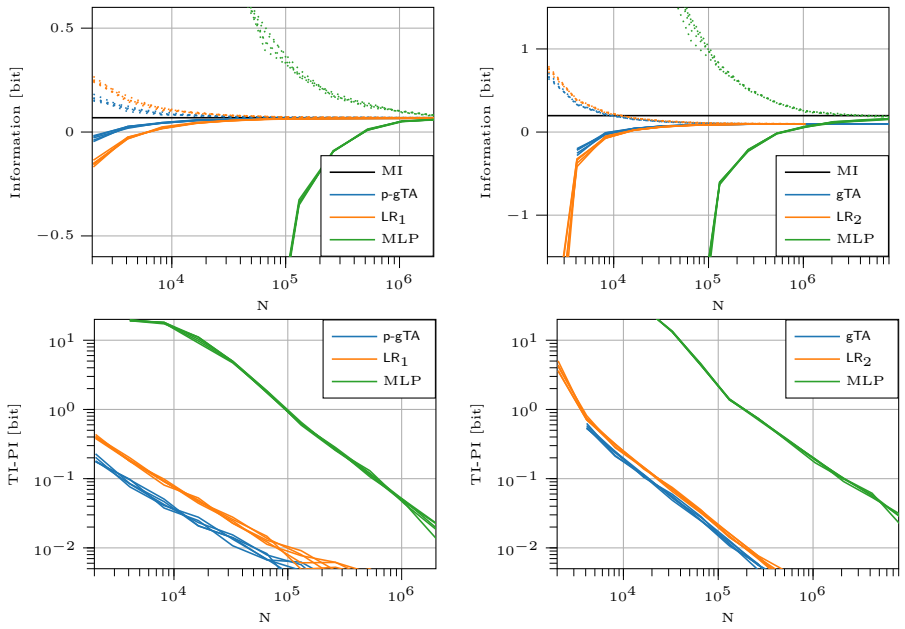
where $\boldsymbol{\beta}_1$ is the first element of $\boldsymbol{\beta}$.

Corollary 7. *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 $\Delta^2 = \left(\mu_1 - \mu_0\right)^{\mathsf{T}}\Sigma^{-1}\left(\mu_1 - \mu_0\right)$ denotes the* Mahalanobis *distance between the two centroids.*

7 Case study and practical use

Let us now illustrate the properties of the TI_N and discuss its practical usage in a side-channel evaluation context. For this purpose, we consider the Hamming Weight leakage of an 8-bit secret in two simulation settings: the first one (HW) corresponds to a typical hardware implementation: no masking and high SNR, while the second one (SW) corresponds to a protected software implementation: 2-shares Boolean masking and high SNR (each share leaking its Hamming Weight independently). In the HW setting, we evaluate the linear models: LR_1 and p-gTA, as well as an MLP (single hidden layer with 100 neurons). In the SW setting, since the leakage model is non-linear, we evaluate the following models: LR_2, gTA and MLP (with the same meta-parameters as in the HW setting). The TI_N and PI of these models for varying number of training traces are shown in Figure 5 (the training is repeated for 5 different training sets). Since the true distribution is known, the MI is also shown.

In the upper part of the figure, we can 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 Theorems 5



(a) HW: not masked, SNR=0.1 (b) SW: masked, SNR=10.

Fig. 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.

and 6, and Corollary 5.[7] Since moreover the expectation of the TI_N is an upper bound for the LI, the TI_N is unlikely to be smaller than the LI, and even if it were to happen, the gap would be practically insignificant. This leads to our first practical conclusion:

At any point in the training, the TI_N *can be considered as an upper bound for the* LI*, it also bounds any* PI *achievable by the considered model.*

As a result, training can be stopped before convergence, as soon as the evaluator is satisfied by the bound provided by the TI_N. The non-tightness of this evaluation method (hereafter named *training gap*, also known as *regret* or *generalization error*) can be quantified: it is bounded by the difference between the TI_N and the PI. Next, we consider the lower part of Figure 5 which depicts the training gap. The slope in the logarithmic plot is close to -1, which means that the gap is inversely proportional to N, as proven in Theorem 6 and Corollary 5.[8] We observe that the slope is close to -1 even when the models do not converge, and over a wide range of training set sizes (more than two decades). This leads to our second practical conclusion:

[7] The hypothesis $\mathsf{p} \in \mathcal{H}$ is not satisfied for the gTA hence Corollary 5 does not apply, but its conclusion seems to nevertheless hold in our case-study.

[8] For the gTA and LR_2, $\mathsf{p} \in \mathcal{H}$ does not hold, but convergence is still in $1/N$.

The training gap is inversely proportional to N. Therefore, after a small and fast training, the evaluator can extrapolate the number of traces required to reach a target training gap.

We finally remark that the MLP model has a higher LI than the LR_2 and gTA models, which means that it is able to better model the true distribution. This increased versatility comes however at a cost: training it requires at least two orders of magnitude more traces than the simpler models (we note that is roughly matches 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 impact for guiding side-channel security evaluators, these results highlight the connections and differences between statistical learning theory and side-channel analysis. For example, in order to obtain convergence rates, we observed that the evaluator's goal, namely maximizing the PI in order 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 results put 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.

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 has been funded in part by the ERC project number 724725 (acronym SWORD).

References

1. N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM*, 44(4):615–631, 1997.
2. T. Anderson. *An Introduction to Multivariate Statistical Analysis.* Wiley Series in Probability and Statistics. Wiley, 2003.
3. M. Anthony and P. L. Bartlett. *Neural Network Learning - Theoretical Foundations.* Cambridge University Press, 2002.
4. A. Antos and I. Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Structures & Algorithms*, 19(3-4):163–193, 2001.

5. C. Archambeau, E. Peeters, F. Standaert, and J. Quisquater. Template attacks in principal subspaces. In *CHES*, volume 4249 of *Lecture Notes in Computer Science*, pages 1–14. Springer, 2006.
6. P. L. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *J. Mach. Learn. Res.*, 20:63:1–63:17, 2019.
7. S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2014.
8. E. Brier, C. Clavier, and F. Olivier. Correlation power analysis with a leakage model. In *CHES*, volume 3156 of *Lecture Notes in Computer Science*, pages 16–29. Springer, 2004.
9. O. Bronchain, J. M. Hendrickx, C. Massart, A. Olshevsky, and F. 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.
10. O. Bronchain and F. Standaert. Side-channel countermeasures' dissection and the limits of closed source security evaluations. *IACR Trans. Cryptogr. Hardw. Embed. Syst.*, 2020(2):1–25, 2020.
11. E. Cagli, C. Dumas, and E. Prouff. Enhancing dimensionality reduction methods for side-channel attacks. In N. Homma and M. 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.
12. E. Cagli, C. Dumas, and E. Prouff. Kernel discriminant analysis for information extraction in the presence of masking. In K. Lemke-Rust and M. 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.
13. E. Cagli, C. Dumas, and E. 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.
14. T. T. Cai, T. Liang, and H. 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.
15. M. Carbone, V. Conin, M. Cornelie, F. Dassance, G. Dufresne, C. Dumas, E. Prouff, and A. Venelli. Deep learning to evaluate secure RSA implementations. *IACR Trans. Cryptogr. Hardw. Embed. Syst.*, 2019(2):132–161, 2019.
16. S. Chari, C. S. Jutla, J. R. Rao, and P. Rohatgi. Towards sound approaches to counteract power-analysis attacks. In *CRYPTO*, volume 1666 of *Lecture Notes in Computer Science*, pages 398–412. Springer, 1999.
17. S. Chari, J. R. Rao, and P. Rohatgi. Template attacks. In *CHES*, volume 2523 of *Lecture Notes in Computer Science*, pages 13–28. Springer, 2002.
18. M. O. Choudary and M. 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.
19. O. Choudary and M. G. Kuhn. Efficient template attacks. In *CARDIS*, volume 8419 of *Lecture Notes in Computer Science*, pages 253–270. Springer, 2013.
20. T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, 2012.

21. V. Cristiani, M. Lecomte, and P. 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.
22. E. de Chérisey, S. Guilley, O. Rioul, and P. 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.
23. A. Duc, S. Faust, and F. 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.
24. A. Duc, S. Faust, and F. 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.
25. F. Durvaux, F. Standaert, and N. 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.
26. B. Efron. The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70(352):892–898, 1975.
27. B. Gierlichs, K. Lemke-Rust, and C. Paar. Templates vs. stochastic methods. In *CHES*, volume 4249 of *Lecture Notes in Computer Science*, pages 15–29. Springer, 2006.
28. T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition*. Springer Series in Statistics. Springer, 2009.
29. D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inf. Comput.*, 100(1):78–150, 1992.
30. D. 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.
31. A. Heuser, O. Rioul, and S. 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.
32. A. Heuser and M. 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.
33. G. Hospodar, B. Gierlichs, E. D. Mulder, I. Verbauwhede, and J. Vandewalle. Machine learning in side-channel analysis: a first study. *J. Cryptogr. Eng.*, 1(4):293–302, 2011.
34. A. Ito, R. Ueno, and N. Homma. Toward optimal deep-learning based side-channel attacks: Probability concentration inequality loss and its usage. *IACR Cryptol. ePrint Arch.*, page 1216, 2021.
35. L. Lerman, G. Bontempi, and O. Markowitch. Power analysis attack: an approach based on machine learning. *Int. J. Appl. Cryptogr.*, 3(2):97–115, 2014.
36. L. Lerman, S. F. Medeiros, G. Bontempi, and O. Markowitch. A machine learning approach against a masked AES. In *CARDIS*, volume 8419 of *Lecture Notes in Computer Science*, pages 61–75. Springer, 2013.
37. L. Lerman, R. Poussier, G. Bontempi, O. Markowitch, and F. 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.

38. H. Maghrebi, T. Portigliatti, and E. Prouff. Breaking cryptographic implementations using deep learning techniques. In *SPACE*, volume 10076 of *Lecture Notes in Computer Science*, pages 3–26. Springer, 2016.
39. S. 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.
40. L. Masure, C. Dumas, and E. Prouff. A comprehensive study of deep learning for side-channel analysis. *IACR Trans. Cryptogr. Hardw. Embed. Syst.*, 2020(1):348–375, 2020.
41. L. Mather, E. Oswald, J. Bandenburg, and M. 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.
42. N. Mehta. Fast rates with high probability in exp-concave statistical learning. In A. Singh and X. J. 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.
43. A. Moradi and F. Standaert. Moments-correlating DPA. In *TISCCS*, pages 5–15. ACM, 2016.
44. J. Mourtada and S. Gaïffas. An improper estimator with optimal excess risk in misspecified density estimation and logistic regression, 2020.
45. L. Paninski. Estimation of entropy and mutual information. *Neural Comput.*, 15(6):1191–1253, 2003.
46. G. Perin, I. Buhan, and S. Picek. Learning when to stop: a mutual information approach to fight overfitting in profiled side-channel analysis. *IACR Cryptol. ePrint Arch.*, page 58, 2020.
47. K. B. Petersen, M. S. Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.
48. S. Picek, A. Heuser, A. Jovic, S. Bhasin, and F. 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.
49. M. Renauld, F. Standaert, N. Veyrat-Charvillon, D. Kamel, and D. 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.
50. W. Schindler, K. Lemke, and C. Paar. A stochastic model for differential side channel cryptanalysis. In *CHES*, volume 3659 of *Lecture Notes in Computer Science*, pages 30–46. Springer, 2005.
51. T. Schneider and A. Moradi. Leakage assessment methodology - extended version. *J. Cryptogr. Eng.*, 6(2):85–99, 2016.
52. S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press, 2014.
53. F. Standaert and C. 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.
54. F. Standaert, F. Koeune, and W. Schindler. How to compare profiled side-channel attacks? In *ACNS*, volume 5536 of *Lecture Notes in Computer Science*, pages 485–498, 2009.

55. F. Standaert, T. Malkin, and M. 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.
56. F. Standaert, E. Peeters, C. Archambeau, and J. Quisquater. Towards security limits in side-channel attacks. In *CHES*, volume 4249 of *Lecture Notes in Computer Science*, pages 30–45. Springer, 2006.
57. C. J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.
58. C. J. Stone. Optimal uniform rate of convergence for nonparametric estimators of a density function or its derivatives. In M. H. Rizvi, J. S. Rustagi, and D. Siegmund, editors, *Recent Advances in Statistics*, pages 393–406. Academic Press, 1983.
59. T. van Erven, P. D. Grünwald, N. A. Mehta, M. D. Reid, and R. C. Williamson. Fast rates in statistical and online learning. *J. Mach. Learn. Res.*, 16:1793–1861, 2015.
60. V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.
61. R. Vershynin. How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 25(3):655–686, 2012.
62. C. Whitnall, E. Oswald, and F. 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.
63. L. Wouters, V. Arribas, B. Gierlichs, and B. Preneel. Revisiting a methodology for efficient CNN architectures in profiling attacks. *IACR Trans. Cryptogr. Hardw. Embed. Syst.*, 2020(3):147–168, 2020.
64. G. Zaid, L. Bossuet, F. Dassance, A. Habrard, and A. Venelli. Ranking loss: Maximizing the success rate in deep learning side-channel analysis. *IACR Trans. Cryptogr. Hardw. Embed. Syst.*, 2021(1):25–55, 2021.
65. G. Zaid, L. Bossuet, A. Habrard, and A. Venelli. Methodology for efficient CNN architectures in profiling attacks. *IACR Trans. Cryptogr. Hardw. Embed. Syst.*, 2020(1):1–36, 2020.

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.* [9, Thm. 5]. Nevertheless, we provide here a simpler alternative proof, by taking inspiration from the work of Paninski [45, 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:

$$\mathsf{eHI}_N(Y;\boldsymbol{L}) = \mathsf{MI}(Y;\boldsymbol{L}) \tag{41}$$

$$+ \sum_{y,\boldsymbol{l}} (\tilde{\mathsf{e}}_N(y,\boldsymbol{l}) - \mathsf{p}(y,\boldsymbol{l})) \log_2(\mathsf{p}(y \mid \boldsymbol{l})) \tag{42}$$

$$+ \sum_{\boldsymbol{l}} \tilde{\mathsf{e}}_N(\boldsymbol{l}) \mathsf{D}_{\mathsf{KL}}(\tilde{\mathsf{e}}_N(\cdot|\boldsymbol{l}) \parallel \mathsf{p}(\cdot \mid \boldsymbol{l})) \ , \tag{43}$$

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 [45, Proposition 1], we use the fact that

$$0 \leq \mathop{\mathbb{E}}_{\tilde{\mathsf{e}}_N} \left[\mathsf{D}_{\mathsf{KL}}(\tilde{\mathsf{e}}_N(\cdot|\boldsymbol{l}) \parallel \mathsf{p}(\cdot \mid \boldsymbol{l})) \right] \leq \log\left(1 + \frac{Q-1}{N}\right) \ . \tag{44}$$

Finally, we have

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

We conclude the proof by observing that $|\mathcal{L}|$ is the number of bins. In addition, Equation 5 is a direct consequence of [45, Thm. 5]. □

Proof of Theorem 2. Notice that

$$\mathsf{eHI}_N = \mathsf{H}(Y) + \widehat{\mathsf{H}(\boldsymbol{L})} - \widehat{\mathsf{H}(Y, \boldsymbol{L})} \ , \tag{45}$$

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

$$\left| \mathsf{eHI}_N - \mathbb{E}\left[\mathsf{eHI}_N\right] \right| \leq \left| \widehat{\mathsf{H}(\boldsymbol{L})} - \mathbb{E}\left[\widehat{\mathsf{H}(\boldsymbol{L})}\right] \right| + \left| \widehat{\mathsf{H}(Y, \boldsymbol{L})} - \mathbb{E}\left[\widehat{\mathsf{H}(Y, \boldsymbol{L})}\right] \right| \ . \tag{46}$$

Now, using McDiarmid's inequality [4, Thm. 1], we have that for all $\epsilon > 0$

$$\Pr\left(\left| \widehat{\mathsf{H}(\boldsymbol{L})} - \mathbb{E}\left[\widehat{\mathsf{H}(\boldsymbol{L})}\right] \right| > \frac{\epsilon}{2} \right) \leq 2 \exp\left(- \frac{\epsilon^2 N}{8 \log_2(N)^2} \right) \ . \tag{47}$$

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 $\epsilon > 0$

$$\Pr\left(\left| \mathsf{eHI}_N - \mathbb{E}\left[\mathsf{eHI}_N\right] \right| > \epsilon \right) \leq 4 \exp\left(- \frac{\epsilon^2 N}{8 \log_2(N)^2} \right) \ . \tag{48}$$

Denoting by δ the right hand-side of Equation 48, 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 [4] (Section 4.1). □

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 [57,58]. 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 5.3

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 5.2.
5. We now have all the ingredients to state the theorem and its corollary.

B.1 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:*

$$\Pr\left(\sup_{\mathsf{m}\in\mathcal{H}}\left|\Delta^{\mathsf{m}}_{\tilde{\mathsf{e}}_N} - \mathsf{PI}(Y;\boldsymbol{L};\mathsf{m})\right| \geq \epsilon\right) \leq \delta \ . \tag{49}$$

Theorem 12 (Uniform Convergence implies Learnability). *With the same notations as in Definition 9, the inequality*

$$\mathsf{LI}(Y;\boldsymbol{L};\mathcal{H}) - \mathsf{PI}(Y;\boldsymbol{L};\widehat{\mathsf{m}}_N) \leq 2\sup_{\mathsf{m}\in\mathcal{H}}\left|\mathsf{PI}(Y;\boldsymbol{L};\mathsf{m}) - \Delta^{\mathsf{m}}_{\tilde{\mathsf{e}}_N}\right| \tag{50}$$

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 $\mathsf{TI}_N(Y; \boldsymbol{L}; \mathcal{A}_{\mathcal{H}}) = \Delta^{\widehat{\mathsf{m}}_N}_{\tilde{\mathsf{e}}_N} \geq \Delta^{\mathsf{m}}_{\tilde{\mathsf{e}}_N}$, therefore

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

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 [1, Thm. 4.2].[9]

Corollary 8. *Let* $\epsilon = \sup_{\mathsf{m} \in \mathcal{H}} \left|\mathsf{PI}(Y; \boldsymbol{L}; \mathsf{m}) - \Delta^{\mathsf{m}}_{\tilde{\mathsf{e}}_N}\right|$, *the following inequalities hold*

$$-3\epsilon \leq \mathsf{TI}_N(Y; \boldsymbol{L}; \mathcal{A}_{\mathcal{H}}) - \mathsf{LI}(Y; \boldsymbol{L}; \mathcal{H}) \leq \epsilon \tag{51}$$

Proof. We first prove the first inequality:

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

where the bound on the first term comes from Theorem 12 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

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

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

[9] For more general learning problem, the uniform convergence may be not necessary (see counter-example in [60, 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 [60, Thm. 3.2].

B.2 Bounding Uniform Convergence with Covering Numbers

We now turn to emphasize uniform bounds, which, thanks to Corollary 8, 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 [52, 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 $\mathcal{A}$.*

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 [3, 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:*

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

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

$$\mathcal{N}_1(\epsilon, \mathcal{H}, N) = \max_{\mathcal{S}_N \in \mathcal{L}^N} N_1(\epsilon, \mathcal{H}_{\mathcal{S}_N}) \quad . \tag{52}$$

Covering numbers are crucial in statistical learning theory. This is formally stated by Theorem 13 hereafter.

Theorem 13 ([29, Thm. 3]). *Let $\mathcal{H}$ be a permissible*[10] *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*

$$\Pr\left(\sup_{\mathsf{m} \in \mathcal{H}} \left|\mathsf{PI}(Y; \boldsymbol{L}; \mathsf{m}) - \Delta^{\mathsf{m}}_{\tilde{\mathsf{e}}_N}\right| > \epsilon\right) \leq 2\,\mathcal{N}_1(\epsilon, \log \circ \mathcal{H}, 2N)\, e^{-\frac{\epsilon^2 N}{64 B^2}} \quad , \tag{53}$$

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

[10] A very loose condition, see [29, Footnote 11].

It now remains to see when Theorem 13 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 53 tends to infinity with $N \to \infty$, if ϵ is small enough. In other words, without further assumption, Theorem 13 is a rather tautological result, and further conditions on $\mathcal{H}$ must be set for sound bounds.

Hopefully, we will see in Section B.4 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.3 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 5 (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:*

$$\|\phi(\boldsymbol{a}) - \phi(\boldsymbol{b})\|_{\mathcal{B}} \leq \rho \|\boldsymbol{a} - \boldsymbol{b}\|_{\mathcal{A}} \ . \tag{54}$$

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

$$N_1(\rho\epsilon, \phi \circ \mathcal{A}) \leq N_1(\epsilon, \mathcal{A}) \tag{55}$$

is valid.

Lemma 5 is inspired by the proof given by Shalev-Shwartz and Ben-David [52, 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:

$$\|\boldsymbol{a} - \boldsymbol{a'}\| \leq \epsilon \ . \tag{56}$$

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

$$\|\phi(\boldsymbol{a}) - \phi(\boldsymbol{a'})\| \leq \rho \|\boldsymbol{a} - \boldsymbol{a'}\| \leq \rho\epsilon \ . \tag{57}$$

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

Corollary 9 (Contraction). *Using the same notations as in Lemma 5, 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:*

$$\mathcal{N}_1(\rho\epsilon, \phi \circ \mathcal{F}, N) \leq \mathcal{N}_1(\epsilon, \mathcal{F}, N) \ . \tag{58}$$

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 5. □

Informally, Corollary 9 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 9 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 6.

Lemma 6. *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 10. *For all $\epsilon > 0$, and for all $N \geq 1$, the following inequality holds:*[11]

$$\mathcal{N}_1(\epsilon, \log \circ \mathcal{H}, N) \leq \mathcal{N}_1\left(\frac{\epsilon}{\sqrt{Q}}, \mathcal{F}^Q, N\right) \quad . \tag{59}$$

Thanks to Corollary 10, 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.

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*

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

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

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

[11] A similar result can be found in [3, Lemma 17.6]

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 7.

Lemma 7 ([29, Lemma 7]). *If $\mathcal{F}_1, \ldots, \mathcal{F}_Q$ are defined as above, then*

$$\mathcal{N}_1\big(\epsilon, \mathcal{F}^Q, N\big) \leq \prod_{i=1}^{k} \mathcal{N}_1\left(\frac{\epsilon}{Q}, \mathcal{F}_i, N\right) \ . \tag{60}$$

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

$$\mathcal{U} = \{(f_1, \ldots, f_Q) : f_i \in \mathcal{U}_i, 1 \leq i \leq Q\} \ . \tag{61}$$

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

$$\|\boldsymbol{g} - \boldsymbol{f}\|_1 = \sum_{i=1}^{Q} \|g_i - f_i\|_1 \leq Q \cdot \frac{\epsilon}{Q} \leq \epsilon \ . \tag{62}$$

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.4 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 14 ([30, 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*

$$\mathcal{N}_1(\epsilon, \mathcal{F}, N) \leq e(\mathsf{P}_{\mathsf{dim}}(\mathcal{F}) + 1) \left(\frac{2e}{\epsilon}\right)^{\mathsf{P}_{\mathsf{dim}}(\mathcal{F})} \tag{63}$$

for all $\epsilon > 0$.

Corollary 11. *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*

$$\mathcal{N}_1(\epsilon, \mathcal{F}, N) \leq e(\mathsf{P}_{\mathsf{dim}}(\mathcal{F}) + 1) \left(\frac{2eB}{\epsilon}\right)^{\mathsf{P}_{\mathsf{dim}}(\mathcal{F})} \tag{64}$$

for all $\epsilon > 0$.

Proof. Straightforward, by applying Corollary 9 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 11 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 13 much more useful now.

B.5 Putting all Together

Now we have characterized every element in the upper bound of Theorem 13 in terms of pseudo-dimension of $\mathcal{F}$, we may gather all those results to come back to a concrete bound. Let us denote $P = \Pr\left(\sup_{\mathsf{m}\in\mathcal{H}} \left|\mathsf{PI}(Y;\boldsymbol{L};\mathsf{m}) - \Delta^{\mathsf{m}}_{\hat{\mathsf{e}}_N}\right| > \epsilon\right)$. Applying Theorem 13, it comes that

$$\begin{aligned}
P \cdot e^{\frac{\epsilon^2 N}{64B^2}} &\overset{(53)}{\leq} 2\,\mathcal{N}_1(2\epsilon, \log\circ\mathcal{H}, 2N) \\
&\overset{(59)}{\leq} 2\,\mathcal{N}_1\left(2\frac{\epsilon}{\sqrt{Q}}, \mathcal{F}^Q, 2N\right) \\
&\overset{(60)}{\leq} 2\,\mathcal{N}_1\left(2\frac{\epsilon}{Q^{3/2}}, \mathcal{F}, 2N\right)^Q \\
&\overset{(64)}{\leq} 2\left((e\,\mathsf{P}_{\mathsf{dim}}(\mathcal{F})+1)\left(\frac{eBQ^{3/2}}{\epsilon}\right)^{\mathsf{P}_{\mathsf{dim}}(\mathcal{F})}\right)^Q \quad .
\end{aligned}$$

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

$$P \leq \exp\left(-\alpha\epsilon^2 - \beta\log\left(\epsilon^2\right) + \gamma\right) \quad . \tag{65}$$

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

$$\alpha\epsilon^2 + \beta\log\left(\epsilon^2\right) \geq \gamma + \log\left(\frac{1}{\delta}\right) \quad . \tag{66}$$

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 66 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 66.

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

$$0 \leq -\log(\sigma(\boldsymbol{x})) \leq 2V + \log(Q) \quad . \tag{67}$$

Proof.

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

□

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 65, and expliciting the expression of ϵ gives Theorem 5.

C Proofs of fast rate

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

Lemma 9. *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 6, we get that for all $(y, \boldsymbol{l})$,

$$\left|\log\Big(\sigma(\mathsf{m}_{\boldsymbol{\Theta}}(\boldsymbol{l}))_y\Big) - \log\Big(\sigma(\mathsf{m}_{\boldsymbol{\Theta}'}(\boldsymbol{l}))_y\Big)\right| \leq \sqrt{Q}\sqrt{\sum_{i=1}^{Q}\left(\mathsf{m}_{\boldsymbol{\Theta}}(\boldsymbol{l})_i - \mathsf{m}_{\boldsymbol{\Theta}'}(\boldsymbol{l})_i\right)^2} \quad . \tag{68}$$

Since m is an LR_2 model, $\mathsf{m}_{\boldsymbol{\Theta}}(\boldsymbol{l})_i = \boldsymbol{l}'^{\intercal} A_i \boldsymbol{l}'$ where $\boldsymbol{l}' = (\boldsymbol{l}, 1)$. Therefore, using Cauchy-Schwartz' inequality, we get

$$\begin{aligned}
|\mathsf{m}_{\boldsymbol{\Theta}}(\boldsymbol{l})_i - \mathsf{m}_{\boldsymbol{\Theta}'}(\boldsymbol{l})_i| &= |\boldsymbol{l}'^{\intercal}(A_i - A'_i)\boldsymbol{l}'| \\
&\leq \|\boldsymbol{l}'\|_2^2 \, \|A_i - A'_i\|_* \\
&\leq (R^2+1) \, \|A_i - A'_i\|_F \\
&= (R^2+1) \, \|\boldsymbol{\theta}_i - \boldsymbol{\theta}'_i\|_2
\end{aligned}$$

Injecting this bound into Equation 68 gives the desired result. □

Lemma 10. *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 68 concludes the proof. □

Restatement of Theorem 6. The original version of Mehta's theorem [42, Thm. 1] required the loss function to be *exp-concave*,[12] 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 $\mathcal{H}$ [59, 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 2. This is a direct application of Theorem 6, 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 8 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 10, we get that the Lipschitz constant L is upper bounded by $\sqrt{Q(R^2+1)}$. Putting all together into Equation 25 gives the desired result. □

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

Proof of Corollary 3. This is a direct application of Theorem 6, 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 9, we get that $L \leq \sqrt{Q}(R^2+1)$. Putting all together into Equation 25 gives the desired result. □

Proof of Corollary 4. This is a direct application of Theorem 6, by properly setting the parameters of the theorem to fit the different assumptions.

First, recal 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 $\boldsymbol{\theta}, \boldsymbol{\theta}'$,

$$\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| \leq T = 2SQL \ . \tag{69}$$

Third, we show the Lipschitzness of MLPs. Using Lemma 6, we get that for all $(y, \boldsymbol{l})$,

$$\left|\log\Big(\sigma(\mathsf{m}_{\boldsymbol{\theta}}(\boldsymbol{l}))_y\Big) - \log\Big(\sigma(\mathsf{m}_{\boldsymbol{\theta}'}(\boldsymbol{l}))_y\Big)\right| \leq \sqrt{Q}\sqrt{\sum_{i=1}^{Q} \left(\mathsf{m}_{\boldsymbol{\theta}}(\boldsymbol{l})_i - \mathsf{m}_{\boldsymbol{\theta}'}(\boldsymbol{l})_i\right)^2} \ . \tag{70}$$

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 $\phi\Big(\boldsymbol{x}^{(j)}, \boldsymbol{\Theta}_i^{(j)}\Big)$ is $\left\|\boldsymbol{x}^{(j)}\right\|$-Lipschitz (resp. $\left\|\boldsymbol{\Theta}_i^{(j)}\right\|$-Lipschitz) in its input $\boldsymbol{\Theta}_i^{(j)}$ (resp. $\left\|\boldsymbol{x}^{(j)}\right\|$), hence

$$\left\|\phi\Big(\boldsymbol{x}^{(j)}, \boldsymbol{\Theta}_i^{(j)}\Big) - \phi\Big(\boldsymbol{x}'^{(j)}, \boldsymbol{\Theta}_i'^{(j)}\Big)\right\| \leq \left\|\boldsymbol{\Theta}_i'^{(j)}\right\| \left\|\boldsymbol{x}^{(j)} - \boldsymbol{x}'^{(j)}\right\| + \left\|\boldsymbol{x}^{(j)}\right\| \left\|\boldsymbol{\Theta}_i^{(j)} - \boldsymbol{\Theta}_i'^{(j)}\right\| . \tag{71}$$

Let us now prove by induction that

$$\left\|\boldsymbol{x}^{(j)} - \boldsymbol{x}'^{(j)}\right\| \leq RS^j \sum_{k=0}^{j} \left\|\boldsymbol{\Theta}_i^{(k)} - \boldsymbol{\Theta}_i'^{(k)}\right\| \ , \tag{72}$$

where $\boldsymbol{x}^{(j+1)} = \phi\Big(\boldsymbol{x}^{(j)}, \boldsymbol{\Theta}_i^{(j)}\Big)$, $\boldsymbol{x}'^{(j+1)} = \phi\Big(\boldsymbol{x}'^{(j)}, \boldsymbol{\Theta}_i'^{(j)}\Big)$ and $\boldsymbol{x}^{(0)} = \boldsymbol{x}'^{(0)} = \boldsymbol{l}$. The base case $j = 1$ is a direct consequence of Equation 71, since $\|\boldsymbol{l}\| \leq R$ and $S \geq 1$. For $j \neq 1$, we observe that $\left\|\boldsymbol{x}^{(j+1)}\right\| \leq \left\|\boldsymbol{\Theta}_i^{(j)}\right\| \left\|\boldsymbol{x}^{(j)}\right\| \leq S^j \|\boldsymbol{l}\| \leq S^j R$. Then, injecting this observation in the second term of Equation 71 and using the induction hypothesis in the first term gives the desired result. Finally, we apply Equation 72 to the full MLP, giving

$$|\mathsf{m}_{\boldsymbol{\theta}}(\boldsymbol{l})_i - \mathsf{m}_{\boldsymbol{\theta}'}(\boldsymbol{l})_i| \leq R \cdot S^L \|\boldsymbol{\theta}_i' - \boldsymbol{\theta}_i\| \ . \tag{73}$$

Injecting the right hand-side of Equation 73 into the one of Equation 70, 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 69, Equation 70, Equation 73 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 25 gives the desired result. □

D Proofs of Section 6

Proof of Theorem 7. By definition, the MI is expressed computed as follows:

$$\mathsf{MI}(Y;\boldsymbol{L}) = \mathsf{H}(Y) + \frac{1}{Q}\sum_{y} \mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_y}\left[\log\left(\frac{\mathsf{f}_y(\boldsymbol{L})}{\sum_{y'}\mathsf{f}_{y'}(\boldsymbol{L})}\right)\right] \ .$$

Likewise, the PI is similar to the MI, by turning true pdfs into estimated pdfs inside the log:

$$\mathsf{PI}(Y;\boldsymbol{L};\mathsf{gTA}) = \mathsf{H}(Y) + \frac{1}{Q}\sum_{y} \mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_y}\left[\log\left(\frac{\widehat{\mathsf{f}}_y(\boldsymbol{L})}{\sum_{y'}\widehat{\mathsf{f}}_{y'}(\boldsymbol{L})}\right)\right] \ .$$

So the regret is expessed as follows:

$$\mathsf{R}\left(\mathsf{gTA}\right) = \frac{1}{Q}\sum_{y} \mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_y}\left[\log\left(\frac{\mathsf{f}_y(\boldsymbol{L})}{\widehat{\mathsf{f}}_y(\boldsymbol{L})}\right) - \log\left(\frac{\sum_{y'}\mathsf{f}_{y'}(\boldsymbol{L})}{\sum_{y'}\widehat{\mathsf{f}}_{y'}(\boldsymbol{L})}\right)\right] \ .$$

Remark that

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

and that by linearity of the expectation,

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

Thus, we end up with the following equality

$$\mathsf{R}\left(\mathsf{gTA}\right) = \frac{1}{Q}\sum_{y} \mathsf{D}_{\mathsf{KL}}\Big(\mathsf{f}_y(\cdot) \,\Big\|\, \widehat{\mathsf{f}}_y(\cdot)\Big) - \mathsf{D}_{\mathsf{KL}}\left(\frac{\sum_y \mathsf{f}_y(\cdot)}{Q} \,\middle\|\, \frac{\sum_y \widehat{\mathsf{f}}_y(\cdot)}{Q}\right) \ .$$

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

Proof of Lemma 1. Let $\boldsymbol{X'} = A\boldsymbol{X} + \boldsymbol{b}$, then the pdf of $\boldsymbol{X'}$ is

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

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 30. □

Proof of Lemma 2. By definition,

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

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

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

Using [47, 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. □

D.1 Proof of Theorem 8

This paragraph is devoted to prove Theorem 8. We begin by recalling a few technical lemmas useful for the proofs in this section.

Lemma 11. *Let $A, B \in \mathbb{R}^{D\times D}$ be symmetric matrices with real coefficients. 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*

$$0 \leq x - \log(1+x) \leq \frac{x^2}{1+x} \ . \tag{74}$$

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

$$0 \leq x - \log(1+x) \leq x\left(1 - \frac{1}{1+x}\right) = \frac{x^2}{1+x} \ .$$

□

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 15 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 $\left\|\widehat{\Sigma} - \mathbf{I}\right\|_* \leq 1/2$,*

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

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,

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

Besides, using Lemma 11,

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

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

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

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

$$\log\Big(\det\Big(\widehat{\Sigma}\Big)\Big) + \mathrm{Tr}\Big(\widehat{\Sigma}^{-1}\Big) - D \leq \sum_{i=1}^{D} \lambda_i \left(\frac{1}{\lambda_i} - 1\right)^2 = \sum_{i=1}^{D} \frac{(\lambda_i - 1)^2}{\lambda_i} \ . \tag{75}$$

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

$$0 \leq \frac{(\lambda_i - 1)^2}{\lambda_i} \leq \frac{\left\| \widehat{\Sigma} - \mathbf{I} \right\|_*^2}{1 - \left\| \widehat{\Sigma} - \mathbf{I} \right\|_*} \leq 2 \left\| \widehat{\Sigma} - \mathbf{I} \right\|_*^2 \ . \tag{76}$$

Finally, combining Equation 76 with Equation 75 gives the result. □

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

Theorem 15 (Prop. 2.1 [61]). *For all Σ, there exists a constant C such that for all $\delta > 0$, the inequality*

$$\left\| \widehat{\Sigma} - \Sigma \right\|_* \leq C \left\| \Sigma \right\|_* \cdot \sqrt{\log\left(\frac{2}{\delta}\right) \frac{D}{N}} \tag{77}$$

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

By combining Theorem 15 and Lemma 13, we conclude the proof of Theorem 8.

D.2 Proof of Tightness

Proof of Lemma 4. For any symmetric positive matrix such as $\widehat{\Sigma}$, the mapping $\widehat{\Sigma} \mapsto \mathrm{Tr}\left(\widehat{\Sigma}^{-1}\right)$ is convex [7, Ex. 3.18]. Using Jensen's inequality, we get

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

Hence, the left hand-side of Equation 36 is non-negative. □

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 $\boldsymbol{e_1} = (1, 0, \ldots, 0)$). 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 14. *The regret for two classes can be rephrased as follows*

$$\mathsf{2R}\left(\textit{p-gTA}\right) = \mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_0}\left[\log\left(1+e^{\widehat{\lambda}(\boldsymbol{L})}\right)\right] + \mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_1}\left[\log\left(1+e^{-\widehat{\lambda}(\boldsymbol{L})}\right)\right] - 2\mathop{\mathbb{E}}_{\boldsymbol{L}\sim\mathsf{f}_0}\left[\log\left(1+e^{\Delta\boldsymbol{e}_1^\intercal\boldsymbol{L}}\right)\right] \quad , \quad (78)$$

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

Proof. First, denoting $\boldsymbol{l}_1 = \boldsymbol{e_1^\intercal l}$ (and $\boldsymbol{L}_1 = \boldsymbol{e_1^\intercal L}$), 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

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

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. □

Lemma 15. *The regret satisfies the following inequality:*

$$\mathsf{R}\left(\boldsymbol{\beta},\gamma\right) \leq \gamma^2 + \|\boldsymbol{\beta}\|_2^2 + |\gamma\boldsymbol{\beta}_1| + \mathcal{O}\left(\left(\gamma^2+\|\boldsymbol{\beta}\|^2\right)^{3/2}\right) \quad . \quad (79)$$

Proof. Using the expression of the regret given in Lemma 14 and taking the Taylor expansion, we have

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

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 14 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

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

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 $i \neq 1$,

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

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

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

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

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

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

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

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

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

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

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

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 $\frac{\partial^2}{\partial\gamma\partial\boldsymbol{\beta}_1}\mathsf{R}(\mathbf{0},0) \leq 0$. Let us now look for a lower bound:

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

Finally, we have that

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

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

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

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

□

Next, we use the following lemma to prove Corollary 7.

Lemma 16 ([26, Lemma 2]). *The estimation error of $\boldsymbol{\beta},\gamma$ satisfies the following convergence in law:*

$$\sqrt{N}\begin{pmatrix}\gamma\\ \boldsymbol{\beta}\end{pmatrix} \underset{N\to\infty}{\overset{L}{\longrightarrow}} \mathcal{N}(\boldsymbol{O},\Sigma) \ , \tag{86}$$

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 7. 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

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

Considering a constant δ gives the final result. □