Best Information is Most Successful

Mutual Information and Success Rate in Side-Channel Analysis

Eloi de Chérisey[1], Sylvain Guilley[1,3,4], Olivier Rioul[1] and Pablo Piantanida[2]

[1] LTCI, Télécom ParisTech, 46 rue Barrault, 75013 Paris, France
[2] LSS, Centrale-Supélec 8-10 rue Joliot-Curie, 91190 Gif-sur-Yvette, France
[3] Secure-IC S.A.S., 15 Rue Claude Chappe, Bât. B, ZAC des Champs Blancs, 35 510 Cesson-Sévigné, France
[4] École Normale Supérieure, 45 rue d'Ulm, 75005 Paris, France

Abstract. Using information-theoretic tools, this paper establishes a mathematical link between the probability of success of a side-channel attack and the minimum number of queries to reach a given success rate, valid for *any* possible distinguishing rule and with the best possible knowledge on the attacker's side. This link is a lower bound on the number of queries highly depends on Shannon's mutual information between the traces and the secret key. This leads us to derive upper bounds on the mutual information that are as tight as possible and can be easily calculated. It turns out that, in the case of an additive white Gaussian noise, the bound on the probability of success of any attack is directly related to the signal to noise ratio. This leads to very easy computations and predictions of the success rate in any leakage model.

Keywords: Side-Channel Analysis · Information Theory · Guessing Entropy · Success Rate

1 Introduction

> As a general rule, the most successful man in life
> is a man who has the best information.
> — Benjamin Disraeli.

Side-channel analysis is renown as an effective "eavesdropping" attack technique to extract sensitive secrets from cryptographic chips. In recent literature, many exploits have been put forward. Starting from the seminal timing attack of Kocher [Koc96], various biases of different kinds have been exhibited. Vertical attacks such as power analysis [KJJ99] have been shown to be highly efficient. However, from a designer's viewpoint, the exact details of the various attacks are irrelevant. Instead, defenders aim at estimating a security risk in general, e.g., the chance that a major security breach occurs. It is thus highly desired to protect designs against all kinds of SCA attacks in a provable way. When implementing a secure design, the natural question which arises is the quantification of its security, with respect to its architecture and its operational environment. In [DSV14], the authors present several metrics that can help the designers to secure cryptographic chips. Shannon's mutual information (MI) between measured traces and guessed models has been considered, but is often thought of as theoretical (too far from practical evaluations) and impracticable (too computationally inefficient). Mutual Information as a metric to quantify the security of a chip has been proposed by [vW01]. In [SPAQ06], the authors explain the relative importance of MI and probability of success, but in a separate way. Our aim is to join the two concepts and to show how the knowledge of MI allows to derive an upper bound on the success rate.

IACR Transactions on Cryptographic Hardware and Embedded Systems ISSN 2569-2925, Vol. 2019, No. 2, pp. 49–79
DOI:10.13154/tches.v2019.i2.49-79

We wish to estimate the success rate with very few assumptions, based on simple and easy-to-compute tools, such as the signal to noise ratio (SNR). The calculation of the SNR can be made without the knowledge of the leakage model as the SNR is the ratio between the power of the useful signal and the power of the noise. The power of the noise is easily measured as is is the measurement noise, and as the power of the useful signal is the difference between the power of the measured signal and the power of the noise, the SNR is obtained.

Related Work As our main goal is to find an estimation of the success rate of an attack that can be as accurate as possible. Using Information theoretic tools, [HRG14] extracted the best possible distinguishing rule. However, this does not give any clue to estimate the success rate of an attack. In practice, the success rate is estimated by repeating a sufficient number of simulations. Moreover, this is dependent of the knowledge of the leakage model. In practice, it is difficult to know exactly this model. Indeed, the estimation may be biased, the learning phase of the model, may be too short, the model, may be too complicated, etc. This is why, we wish to use general information theoretic tools in order to be as generic as possible, and to give bounds that are true whatever the attacker may do or may know.

In [Man04], a link between the success rate and the number of traces to succeed in a correlation power analysis [BCO04] has been studied, and an analytical formula has been derived. However, this results is untrustworthy in practice because of the assumption that incorrect key guesses lead to independent distinguishers, which is not true. Subsequent work on this topic therefore consider the joint distribution of all values of the distinguisher (correct key and all remaining incorrect key guesses).

In [Riv08, LPR+14, GHR15], the authors propose an estimation of the success rate of specific distinguishers. Namely, Rivain [Riv08] studies the distribution of two examples of distinguishers (correlation and template) in the presence of normal noise. Lomné et al. [LPR+14] extend this work for masked implementations, while however still focusing on correlation and template attacks. Guilley et al. [GHR15] extend the approach from additive to some non-additive distinguishers (such as the mutual information analysis), but through the approximation that the number of traces tends to the infinity. To summarize, all three papers [Riv08, LPR+14, GHR15] have in common that the knowledge of the leakage model, or at least an estimation via a learning phase with templates, is needed to predict the success rate. In addition, this estimation, in the three cases, is based on the central limit theorem, meaning that it is relevant for a large number of traces and only for additive distinguishers. We wish a bound valid for any distinguisher, for any number of traces (even small).

A bound on the Mutual Information is proposed in [PR13]. The MI involved is based on one trace, supposing that every leakage is independent from each other. We show in this paper that this is not the case in practice. In this paper, the bound is valid for MI with only *one measurement*. We will see in our paper that calculating MI with the probability functions of *all* the traces is crucial.

In [DFS15, Theorem 2], the authors proposed a link between success rate and the number of measurements. This bound is based on the the link between MI and random probing. Therefore, it is valid only for leakages with very low SNR and the bound is very loose. For instance (see Figure 8), with $\mathsf{SNR} > 10^{-4}$, the bound of Duc et al. [DFS15] is trivial (the success rate is smaller than one), and for $\mathsf{SNR} = 10^{-5}$, it predicts a number of traces 4, which is much smaller than our result of 1.3×10^6 (where the best attack using Maximum Likelihood (ML) predicts 1.5×10^6, which is in the order of magnitude of our prediction). In fact, the main contribution of the bound of Duc et al. [DFS15] is to show that the masking order of an attack has an exponential impact on the success rate, but not to yield an accurate link between number of traces and success rate.

MIA [GBTP08, BGP+11] is a distinguisher, which consists in measuring the dependence

between a key-related model and the leakage measurements, pre-processed to be represented as probability density (or mass) functions. Still, such model can be inaccurate in practical situations, in particular because the model must be non-injective for MIA to be sound. Therefore, MIA is not necessarily the most efficient distinguisher. In this paper, we aim at determining a conservative minimum trace count performance with respect to the key extraction while being independent of any adversarial strategy to exploit leaks. We determine a minimum amount of traces for any attack to recover the key, be it with MIA or other side-channel distinguishers. For this reason, we focus on conservative (albeit tight) metrics to estimate the worst case (for the defender) attack complexity in terms of number of traces. We do study mutual information because it provides a theoretical mathematical tool that measures the amount of information that leaks independently of any attack and allows us to determine the desired conservative minimum trace count.

In the field of information theory, Arimoto [Ari73] proved a lower bound of the error rate (hence an upper-bound of the success rate) in terms of a so-called Gallager coefficient. However, not only requires intensive computations, but also the model assumes a freely chosen input distribution. In our case, that input distribution is set by the leakage model and therefore, cannot be freely chosen. Arimoto's main result (Equation 24 of [Ari73]) remains true because it represents the best possible case for an attacker for all possible input distributions; but the resulting bound is very loose in our side-channel context. Equation 9 of [Ari73] could be used instead but depends on a parameter β. With our notations (presented in section 2), Arimoto's Equation 9 becomes:

$$\forall \beta > 0, \qquad \mathrm{P}_e \leq 1 - 2^{n(\beta-1)} \sum_{\mathbf{t} \in \mathcal{T}^q} \mathbb{P}(\mathbf{t}) \sum_{\mathbf{x} \in \mathcal{X}^q} \left[\sum_{k=0}^{2^n-1} \mathbb{P}(k) \mathbb{P}(\mathbf{x} \mid k, \mathbf{t})^{1/\beta} \right]^{\beta}.$$

The minimization of the r.h.s is practical untraceable for $q > 1$. Indeed, it consists in sums over $|\mathcal{X}|^q$ elements; the complexity is even worse when the output is continuous.

Overall, we can sum up the related work with the following table 1. The table classifies the state-of-the-art according various criteria, such as the way the results are derived and whether or not the mutual information is involved in the estimation of the success rate. The last two columns show whether a closed form bound exists and whether it is generic in the attack method. Our method provided an analytic expression for the lower bound (Theorem 1) and is agnostic in the attack method.

Table 1: Summary of the related work

Related work	**Link with information theory**	**Usage of MI**	**Closed form bound on SR**	**Generic**
[HRG14]	Yes	No	No	No
[Riv08] [LPR+14] [GHR15]	No	No	Yes (but asymptotic)	No
[PR13]	No	Yes	No	Yes
[DFS15]	No	Yes	Yes (but very loose)	Yes
[Ari73]	Yes	No	Computationally too difficult	Yes
This paper	Yes	Yes	Yes (Theorem 1)	Yes

Contributions In this article, we derive bounds on the success rate of any attack, irrespective to the exact attack. Thus we can consider our bounds as *universal*. To do so, we address this problem using rigorous information theoretic tools. This is why we revisit the use of MI as a conservative security metric. Our main contribution is to give a clear

relationship between MI and probability of success. More precisely, we seek a lower bound on the number of available traces where a given success level can be reached, based only on theoretical assumptions on the channel. The actual value of MI is important to estimate and such an estimation is not immediate because random vectors of very high dimensions are involved in its expression. Therefore, we propose several ways to simply estimate the MI by mathematically proved upper bounds and by numerical estimations. Our results are applied to the most common type of noise, namely the additive white Gaussian noise. We show that, in the case of additive Gaussian noise, the only calculation of the SNR is sufficient enough predict accurately the security of a device. Last, the main result on success rate is translated in terms of guessing entropy, another informative criterion in side-channel analysis.

Organization This paper is organized as follows. Section 2 describes the side-channel and shows how a leakage can be modeled with a Markov chain. Section 3 provides our main result and three different ways to exploit it. An application to leakages with additive Gaussian noise is carried out in Section 4, where we show at the end that the SNR is enough to predict the security of a device. In Section 5, we apply our results on practical experiments. The link to the guessing entropy is done in Section 6. Section 7 concludes. Technical computations involved in proofs are in Appendix.

Notations Throughout this paper we use the following notations. Calligraphic letters (e.g. $\mathcal{X}$) denote sets. Uppercase letters (e.g. X) denote random variables taking their values in the corresponding set (e.g. $\mathcal{X}$). Lowercase letters (e.g. x) denote realizations of this random variable. Vectors are written in bold characters. By default, the length of a vector is $q \in \mathbb{N}$. Thus, a random vector is denoted with a bold capital letter (e.g. $\mathbf{X} = (X_1, X_2, \ldots, X_q)$) and a vector of realizations on this random vector is denoted with a small bold letter (e.g. $\mathbf{x} = (x_1, x_2, \ldots, x_q)$). Given the random variable X taking its values in $\mathcal{X}$ and $x \in \mathcal{X}$, the probability that X equals x is noted $\mathbb{P}(X = x)$ or simply $\mathbb{P}(x)$.

We also define some information theoretic tools. The entropy of a random vector $\mathbf{X}$ of length q is defined by:

$$H(\mathbf{X}) = -\sum_{\mathbf{x} \in \mathcal{X}^q} \mathbb{P}(\mathbf{x}) \log_2 \mathbb{P}(\mathbf{x}).$$

The conditional entropy of a random vector $\mathbf{X}$ knowing vector $\mathbf{Y}$ is defined by:

$$\begin{aligned} H(\mathbf{X} \mid \mathbf{Y}) &= -\sum_{\mathbf{y} \in \mathcal{Y}^q} \mathbb{P}(\mathbf{y}) H(\mathbf{X} \mid \mathbf{Y} = \mathbf{y}) \\ &= -\sum_{\mathbf{y} \in \mathcal{Y}^q} \mathbb{P}(\mathbf{y}) \sum_{\mathbf{x} \in \mathcal{X}^q} \mathbb{P}(\mathbf{x} \mid \mathbf{y}) \log_2 \mathbb{P}(\mathbf{x} \mid \mathbf{y}). \end{aligned}$$

The Mutual Information between two random vectors $\mathbf{X}$ and $\mathbf{Y}$ is defined as $I(\mathbf{X}; \mathbf{Y}) = H(\mathbf{X}) - H(\mathbf{X} \mid \mathbf{Y})$. The conditional Mutual Information $I(\mathbf{X}; \mathbf{Y} \mid \mathbf{T})$ where $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{T}$ are random vectors is defined as $I(\mathbf{X}; \mathbf{Y} \mid \mathbf{T}) = H(\mathbf{X} \mid \mathbf{T}) - H(\mathbf{X} \mid \mathbf{Y}, \mathbf{T})$. Last, the Kullback-Leibler divergence between two distributions $\mathbb{P}$ and $\mathbb{Q}$ over the same set $\mathcal{X}$ is defined as:

$$\mathrm{D}(\mathbb{P} \| \mathbb{Q}) = \sum_{x \in \mathcal{X}} \mathbb{P}(x) \log_2 \frac{\mathbb{P}(x)}{\mathbb{Q}(x)}.$$

2 Side-Channel Seen as a Communication Channel

The link between side-channel analysis and information theory has been proposed by [SMY09] and later exploited by [HRG14] to derive the optimal distinguisher. In this section, we review how the side-channel can be seen as a communication channel. The secret key byte that the attacker wants to recover is denoted as k^* and is n bits long (typically $n = 8$). We assume that the attacker inputs q text bytes $\mathbf{t} = (t_1, t_2, \ldots, t_q)$ and receives that many traces in a vector $\mathbf{x} = (x_1, x_2, \ldots, x_q)$, with the following *leakage model*:

$$x_i = f(t_i \oplus k^*) + n_i \qquad (i = 1, 2, \ldots, q) \tag{1}$$

where $\mathbf{n} = (n_1, n_2, \ldots, n_q)$ is an additive noise independent of $\mathbf{x}$ and $f(.)$ is some leakage function. We assume that f is deterministic but not necessary known to the attacker. This assumption will make our calculations generic and therefore true for any type of attack. This is the worst possible case for the security designers. Define the *sensitive variable* $\mathbf{y}(k) = \mathbf{y_t}(k)$ as

$$\mathbf{y_t}(k) = f(\mathbf{t} \oplus k) = (f(t_1 \oplus k), \ldots, f(t_q \oplus k)) \tag{2}$$

so that the leakage can be written in compact form as

$$\mathbf{x} = \mathbf{y_t}(k^*) + \mathbf{n}.$$

Such vectors $\mathbf{t}, \mathbf{y}$ and $\mathbf{x}$ are realizations of random vectors noted $\mathbf{T}, \mathbf{Y}$ and $\mathbf{X}$. In the case of one particular sample, t, y and x are realizations of random variables T, Y and X. We assume that the channel is *memoryless*, which means that each trace x_i depends on the input $\mathbf{y}$ only from y_i. In particular x_i and y_j are independent for all if $i \neq j$. We also make the natural assumption that the secret key is independent from all text bytes: the secret key random variable K is independent from $\mathbf{T}$. In other words, the text bytes do not give any information about the secret key (at least in a design which adheres to Kerckhoffs's principle).

Following [HRG14] we make the following hypotheses:

- K is uniformly distributed over $\mathcal{K} = \{0, \ldots, 2^n - 1\}$. K is a scalar (there is one key-byte to break), and is therefore not written in bold font.
- T is uniformly distributed over $\mathcal{T} = \{0, \ldots, 2^n - 1\}$. Moreover, we suppose that vector $\mathbf{T}$ is *balanced*, meaning that the number of occurrences of each symbol in the vector is the same.
- As seen above, the random variable Y is such that $Y = f(T \oplus K)$, with f a known deterministic function.
- As q textbytes are sent and therefore q traces are received, we consider the random vectors $\mathbf{T}, \mathbf{Y}$ and $\mathbf{X}$.

Thus from (1), we can write

$$\begin{aligned} \mathbf{X} &= f(\mathbf{T} \oplus K) + \mathbf{N} \\ &= \mathbf{Y} + \mathbf{N}. \end{aligned}$$

Considering only scalars, this writes for random variables

$$\begin{aligned} X &= f(T \oplus K) + N \\ &= Y + N. \end{aligned}$$

After acquiring q traces, the attacker applies a function called *distinguisher* $\mathcal{D}$ to obtain an estimate $\hat{K} = \mathcal{D}(\mathbf{X}, \mathbf{T})$ of the secret key from $\mathbf{X}$ and $\mathbf{T}$. This allows us to define the communication channel as depicted in Figure 1:

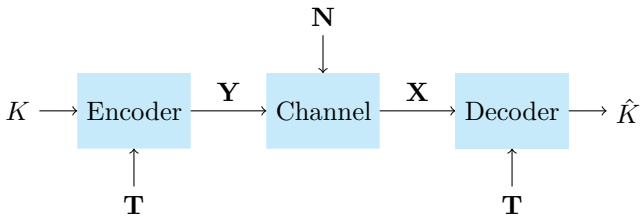


Figure 1: Representation of Side-Channel

- the "encoder" models the leakage from the device: not only the composition of the algorithm which mixes the unknown key K with the known text $\mathbf{T}$ into a sensitive variable, but also the way the device leaks the sensitive variable (function f);
- the (side) channel consists in noise addition, arising from the untargeted parts of the design and from the measurement setup; and
- the "decoder" implements the *distinguishing rule* with allows the attacker to get a key guess $\hat{K}$ from the measured leakage $\mathbf{X}$ and the knowledge of public text bytes $\mathbf{T}$. The realizations $\mathbf{t}$ of the random vector $\mathbf{T}$ are known by the attacker.

From the model we can deduce Lemma 1 dealing with Markov chains. We recall that a Markov chain is *a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event.*

Lemma 1. *The communication channel just described admits the following Markov chains:*

$$\begin{aligned} &(K, \mathbf{T}) \longrightarrow (\mathbf{Y}, \mathbf{T}) \longrightarrow (\mathbf{X}, \mathbf{T}) \longrightarrow \hat{K} \\ &K \longrightarrow \mathbf{Y} \longrightarrow \mathbf{X} \longrightarrow \hat{K} \\ &(K, \mathbf{T}) \longrightarrow \mathbf{Y} \longrightarrow \mathbf{X}. \end{aligned} \tag{3}$$

Proof. The first case is easily seen by re-drawing Figure 1 into the different constitutive blocks as shown in Figure 2, where all the variables pass through different blocks corresponding to the Markov Chain. The two other cases are proved similarly. □

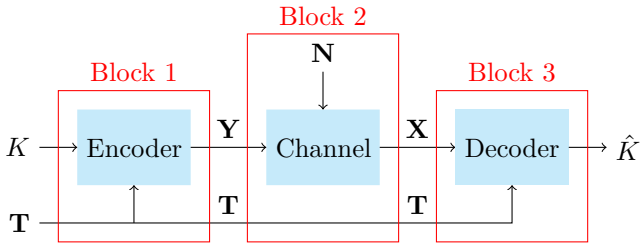


Figure 2: The Markov chain $(K, \mathbf{T}) \longrightarrow (\mathbf{Y}, \mathbf{T}) \longrightarrow (\mathbf{X}, \mathbf{T}) \longrightarrow \hat{K}$.

3 Theoretical Bounds on Mutual Information

One of the important properties of a Markov chain is the *data processing inequality* [CT06], which is used to prove the following theorem in this section, which is our main result.

3.1 Main Result

Let $\mathrm{P}_s = \mathbb{P}(\hat{K} = K)$ be the probability of success of an attack and $H_2(\mathrm{P}_s)$ its binary entropy[1] [CT06]:

$$H_2(\mathrm{P}_s) = -\mathrm{P}_s \log_2(\mathrm{P}_s) - (1 - \mathrm{P}_s) \log_2(1 - \mathrm{P}_s).$$

The following theorem is fundamental because it provides a trade-off for any possible type of attack.

Theorem 1. *The following inequality is always true for any distinguishing rule:*

$$H(K) - (1 - \mathrm{P}_s) \log_2(2^n - 1) - H_2(\mathrm{P}_s) \leq q \cdot I(X; Y \mid T). \tag{4}$$

The probability of success of an attack also follows the following inequality:

$$\begin{aligned} &H(K) - (1 - \mathrm{P}_s) \log_2(2^n - 1) - H_2(\mathrm{P}_s) \\ &\leq \mathbb{E}_{\mathbf{T}} \mathbb{E}_{K_1} \log_2 \mathbb{E}_{K_2} \exp\left(-\mathrm{D}(\mathbb{P}_{\mathbf{X}|K1,\mathbf{T}} \| \mathbb{P}_{\mathbf{X}|K_2,\mathbf{T}})\right); \end{aligned} \tag{5}$$

where $\mathrm{D}(\mathbb{P}\|\mathbb{P}')$ *is the Kullback-Leibler divergence [CT06] and* K_1, K_2 *are identically distributed as* K*.*

Merging these two equations we can write:

$$\begin{aligned} &H(K) - (1 - \mathrm{P}_s) \log_2(2^n - 1) - H_2(\mathrm{P}_s) \\ &\leq \min(\mathbb{E}_{\mathbf{T}} \mathbb{E}_{K_1} \log_2 \mathbb{E}_{K_2} \exp\left(-\mathrm{D}(\mathbb{P}_{\mathbf{X}|K1,\mathbf{T}} \| \mathbb{P}_{\mathbf{X}|K_2,\mathbf{T}})\right), qI(X; Y \mid T)). \end{aligned} \tag{6}$$

Notice that function $\mathrm{P}_s \in [2^{-n}, 1] \mapsto H(K) + (\mathrm{P}_s - 1) \log_2(2^n - 1) - H_2(\mathrm{P}_s) \in [0, n]$ is strictly increasing. The reason to start P_s from 2^{-n} is that this is the baseline for the probability to guess a correct key out of 2^n candidates. Therefore, P_s can be derived by inversing this bijective function.

This theorem shows that the success rate of an attack is directly linked to the Mutual Information between the leakage and the model. Furthermore, as we consider generic attacks, this inequality remains true whatever the attacker does with the traces. In the next subsections we prove both inequalities and we show that (4) is more interesting for low values of q while (5) is a better approximation for high values of q.

To do so, we first demonstrate a preliminary lemma in Section 3.2 that will be useful for both Equation (4) and (5).

3.2 A Fundamental Lower Bound on Mutual Information $I(\mathbf{X}; \mathbf{Y} \mid \mathbf{T})$

The first step of the demonstration of Theorem 1 is the following lemma that links the Mutual Information between the random vectors $\mathbf{X}$ and $\mathbf{Y}$ with the probability of success.

Lemma 2. *With the notations of Theorem 1, we have:*

$$H(K) - (1 - \mathrm{P}_s) \log_2(2^n - 1) - H_2(\mathrm{P}_s) \leq I(\mathbf{X}; \mathbf{Y} \mid \mathbf{T}). \tag{7}$$

Proof. Using the Markov Chain (3) we compare two MI values thanks to the *data processing inequality* [BR12]. Indeed, this is a direct consequence of Lemma 1. This inequality states that the further two random variables are in a Markov Chain, the less MI between these variables. Here we have

$$I((K, \mathbf{T}); (\mathbf{X}, \mathbf{T})) \leq I((\mathbf{Y}, \mathbf{T}); (\mathbf{X}, \mathbf{T})). \tag{8}$$

[1] The binary entropy is the entropy of a binary random variable with probabilities p and $1 - p$.

Let us expand both sides of this inequality. In the l.h.s., since the channel is memoryless and K and $\mathbf{T}$ are independent, we have:

$$\begin{aligned} I((K,\mathbf{T});(\mathbf{X},\mathbf{T})) &= H(K,\mathbf{T}) - H((K,\mathbf{T}) \mid (\mathbf{X},\mathbf{T})) \\ &= H(K) + H(\mathbf{T}) - H(K \mid \mathbf{T},\mathbf{X}). \end{aligned}$$

As $\hat{K}$ is a deterministic function of $\mathbf{T}$ and $\mathbf{X}$, adding the knowledge of $\hat{K}$ does not change the entropy:

$$\begin{aligned} I((K,\mathbf{T});(\mathbf{X},\mathbf{T})) &= H(K) + H(\mathbf{T}) - H(K \mid \mathbf{T},\mathbf{X},\hat{K}); \\ &\geq H(K) + H(\mathbf{T}) - H(K \mid \hat{K}). \end{aligned}$$

The latter inequality holds since conditioning reduces entropy [CT06]. Now by Fano's inequality[2] [CT06, Page 43],

$$H(K \mid \hat{K}) \leq H_2(\mathrm{P}_e) + \mathrm{P}_e \log_2(|\mathcal{K}| - 1)$$

where P_e is the probability of error $\mathrm{P}_e = \mathbb{P}(K \neq \hat{K})$. Since $\mathrm{P}_s = 1 - \mathrm{P}_e$ and $H_2(\mathrm{P}_e) = H_2(\mathrm{P}_s) = -\mathrm{P}_e \log_2(\mathrm{P}_e) - \mathrm{P}_s \log_2(\mathrm{P}_s)$, this is rewritten as

$$H(K \mid \hat{K}) \leq H_2(\mathrm{P}_s) + (1 - \mathrm{P}_s) \log_2(2^n - 1).$$

Plugging this inequality into the previous one gives

$$I((K,\mathbf{T});(\hat{K},\mathbf{T})) \geq H(K) + qH(T) - H_2(\mathrm{P}_s) - (1 - \mathrm{P}_s) \log_2(2^n - 1). \tag{9}$$

On the other hand, the r.h.s. of the data processing inequality (8) is:

$$\begin{aligned} I((\mathbf{Y},\mathbf{T});(\mathbf{X},\mathbf{T})) &= H(\mathbf{X},\mathbf{T}) - H(\mathbf{X},\mathbf{T} \mid \mathbf{Y},\mathbf{T}); \\ &= H(\mathbf{X},\mathbf{T}) - H(\mathbf{X} \mid \mathbf{Y},\mathbf{T}); \\ &= I(\mathbf{X};\mathbf{Y} \mid \mathbf{T}) + H(\mathbf{T}). \end{aligned} \tag{10}$$

Combining Equations (9) and (10), we obtain the following fundamental inequality:

$$H(K) - H_2(\mathrm{P}_s) - (1 - \mathrm{P}_s) \log_2(2^n - 1) \leq I(\mathbf{X};\mathbf{Y} \mid \mathbf{T}), \tag{11}$$

And proving Lemma 2. □ □

The same l.h.s. of (11) will be used to prove for both inequalities (4) and (5), the difference being the way that $I(\mathbf{X};\mathbf{Y} \mid \mathbf{T})$ is evaluated. Indeed, the next part of the proofs for Equations (4) and (5) is about finding an upper-bound for $I(\mathbf{X};\mathbf{Y} \mid \mathbf{T})$. We have to do so because there is no analytic expression for this conditional Mutual Information computed with vectors of q dimensions.

Remark 1. A quick analysis of the value $n + (\mathrm{P}_s - 1) \log_2(2^n - 1) - H_2(\mathrm{P}_s)$ reveals that it is always non-negative for any P_s in the range $(0, 1)$ and vanishes if and only if $\mathrm{P}_s = 1/2^n$.

Therefore, when there are no traces, $I(\mathbf{X};\mathbf{Y} \mid \mathbf{T}) = 0$, the only probability that can respect inequality (11) is $\mathrm{P}_s = 1/2^n$, meaning that without information, that attacker can not have a better success rate than $1/2^n$ obtained with an equiprobable random guess, as expected. Every trace will bring additional information and therefore increase the probability of success.

[2] Fano's inequality is an important information-theoretic result about the uncertainty of the transmission of a message, which is due to the error probability and the number of possible errors.

3.3 First Upper Bound on $I(\mathbf{X};\mathbf{Y} \mid \mathbf{T})$: Proof of Inequality (4)

Thanks to Lemma 2, the l.h.s. of Theorem 1 is given. Inequality (4) is a straightforward consequence of the following lemma.

Lemma 3. *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be two random vectors with joint distribution* $\mathbb{P}_{\mathbf{X},\mathbf{Y}}$, $\mathbb{P}_{\mathbf{X}}$ *be the marginal distribution of* $\mathbf{X}$, *and* $\mathbb{P}_X$ *be the marginal of one element* X *of vector* $\mathbf{X}$. *Define the distribution* $\tilde{\mathbb{P}}_{\mathbf{X}} = \prod_{i=1}^{q} \mathbb{P}_{X_i}$. *We have*

$$\begin{aligned} I(\mathbf{X};\mathbf{Y}) &= qI(X;Y) - \mathrm{D}\left(\mathbb{P}_{\mathbf{X}} \| \tilde{\mathbb{P}}_{\mathbf{X}}\right); \\ &\leq qI(X;Y). \end{aligned}$$

This Lemma means that the Mutual Information of two random vectors made of identically distributed random variables is lower than q times the Mutual Information of the marginal distribution of these random vectors.

Proof. From the memoryless assumption of the channel, one has $\mathbb{P}_{\mathbf{X}|\mathbf{Y}} = \prod_{i=1}^{q} \mathbb{P}_{X_i|Y_i}$. Thus

$$\begin{aligned} I(\mathbf{X};\mathbf{Y}) &= \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\log_2 \frac{\mathbb{P}_{\mathbf{X}|\mathbf{Y}}(\mathbf{X} \mid \mathbf{Y})}{\mathbb{P}_{\mathbf{X}}(\mathbf{X})}\right] \\ &= \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\log_2 \frac{\mathbb{P}_{\mathbf{X}|\mathbf{Y}}(\mathbf{X} \mid \mathbf{Y})}{\tilde{\mathbb{P}}_{\mathbf{X}}(\mathbf{X})}\right] + \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\log_2 \frac{\tilde{\mathbb{P}}_{\mathbf{X}}(\mathbf{X})}{\mathbb{P}_{\mathbf{X}}(\mathbf{X})}\right] \\ &= \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\log_2 \frac{\prod_i \mathbb{P}_{X|Y}(X_i \mid Y_i)}{\prod_i \tilde{\mathbb{P}}_X(X_i)}\right] - \mathrm{D}\left(\mathbb{P}_{\mathbf{X}} \| \tilde{\mathbb{P}}_{\mathbf{X}}\right) \\ &= \sum_i \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\log_2 \frac{\mathbb{P}_{X|Y}(X_i \mid Y_i)}{\tilde{\mathbb{P}}_X(X_i)}\right] - \mathrm{D}\left(\mathbb{P}_{\mathbf{X}} \| \tilde{\mathbb{P}}_{\mathbf{X}}\right) \\ &= qI(X;Y) - \mathrm{D}\left(\mathbb{P}_{\mathbf{X}} \| \tilde{\mathbb{P}}_{\mathbf{X}}\right). \end{aligned}$$

The inequality follows since the divergence is always non-negative. □

This upper bound on MI is easily derived but is linear in q, and, therefore, will not converge to a finite value as the number of measurements increases ($q \to \infty$). This will be in contradiction with Lemma 4. Therefore, it is interesting to propose another bound that converges to a finite value. This will be made in the next section.

3.4 Second Upper Bound on $I(\mathbf{X};\mathbf{Y} \mid \mathbf{T})$ - Proof of Inequality (5)

Before proving (5) we first notice that in our side-channel model, as there is a finite number of keys, the MI is always bounded by $H(K)$.

Lemma 4.

$$I(\mathbf{X};\mathbf{Y} \mid \mathbf{T}) = I(K;\mathbf{X} \mid \mathbf{T}) \leq H(K)$$

Proof. We use the Markov chain defined in Equation (3). Notice that, adding the knowledge of $\mathbf{T}, K$ when $\mathbf{T}, \mathbf{Y}$ are already known does not change the entropy of $\mathbf{X}$. Therefore,

$$\begin{aligned} H(\mathbf{X} \mid \mathbf{T},\mathbf{Y}) &= H(\mathbf{X} \mid \mathbf{T},\mathbf{Y},K,\mathbf{T}); \\ &= H(\mathbf{X} \mid \mathbf{T},\mathbf{Y},K). \end{aligned}$$

As $\mathbf{Y}$ is a deterministic function of K and $\mathbf{T}$, it can be removed, so we get:

$$H(\mathbf{X} \mid \mathbf{T},\mathbf{Y}) = H(\mathbf{X} \mid \mathbf{T},K).$$

Therefore, we obtain $I(\mathbf{X};\mathbf{Y} \mid \mathbf{T}) = I(\mathbf{X};K \mid \mathbf{T})$. Since $I(\mathbf{X};K \mid \mathbf{T}) = H(K) - H(K \mid \mathbf{T},\mathbf{X})$ in follows that $I(\mathbf{X};K \mid \mathbf{T}) \leq H(K)$. □

Here $H(K)$ is a constant that depends only on the distribution of K; it reaches its maximum value for a uniform distribution: $H(K) = n$ bit. As a consequence, since $I(\mathbf{X}; \mathbf{Y} \mid \mathbf{T})$ increases with q, it must converge to a finite value when $q \to \infty$. This explains why the upper-bound given by (4) is poor when $q \to \infty$.

Therefore, we provide another bound that is more accurate for large values of q because it converges to a finite value when K is finite. First we need the following

Lemma 5. *For any random variables X and Y and real-valued function $(x, y) \mapsto f(x, y)$,*

$$-\mathbb{E}_Y \log_2 \mathbb{E}_X[\exp(f(X,Y))] \leq -\log_2 \mathbb{E}_X[\exp(\mathbb{E}_Y f(X,Y))].$$

Proof. See Appendix B. □

Corollary 1. *For any random variables X and Y and positive function $(x, y) \mapsto g(x, y)$,*

$$\exp \mathbb{E}_Y \log_2 \mathbb{E}_X[g(X,Y)] \geq \mathbb{E}_X[\exp(\mathbb{E}_Y \log g(X,Y))]$$

Proof. See Appendix C. □

Equipped with Lemma 5, we compute MI as follows:

$$\begin{aligned} I(\mathbf{X}; K \mid \mathbf{T}) &= \mathbb{E}_{\mathbf{T}} \mathbb{E}_{\mathbf{X},K|\mathbf{T}} \log_2 \frac{\mathbb{P}(\mathbf{X} \mid K\mathbf{T})}{\mathbb{P}(\mathbf{X} \mid \mathbf{T})}; \\ &= \mathbb{E}_{\mathbf{T}} \mathbb{E}_K \mathbb{E}_{\mathbf{X}|K,\mathbf{T}} \log_2 \frac{\mathbb{P}(\mathbf{X} \mid K\mathbf{T})}{\mathbb{P}(\mathbf{X} \mid \mathbf{T})}; \end{aligned}$$

We introduce here K_2, a random variable following the same distribution as K.

$$\begin{aligned} I(\mathbf{X}; K \mid \mathbf{T}) &= \mathbb{E}_{\mathbf{T}} \mathbb{E}_K \mathbb{E}_{\mathbf{X}|K,\mathbf{T}} \log_2 \frac{\mathbb{P}(\mathbf{X} \mid K, \mathbf{T})}{\mathbb{E}_{K_2} \mathbb{P}(\mathbf{X} \mid K_2, \mathbf{T})}; \\ &= -\mathbb{E}_{\mathbf{T}} \mathbb{E}_K \mathbb{E}_{\mathbf{X}|K,\mathbf{T}} \log_2 \mathbb{E}_{K_2} \frac{\mathbb{P}(\mathbf{X} \mid K_2, \mathbf{T})}{\mathbb{P}(\mathbf{X} \mid K, \mathbf{T})}; \\ &= -\mathbb{E}_{\mathbf{T}} \mathbb{E}_K \mathbb{E}_{\mathbf{X}|K,\mathbf{T}} \log_2 \mathbb{E}_{K_2} \exp\left[\log_2 \frac{\mathbb{P}(\mathbf{X} \mid K_2, \mathbf{T})}{\mathbb{P}(\mathbf{X} \mid K, \mathbf{T})}\right]. \end{aligned}$$

By Lemma 5 we obtain

$$\begin{aligned} I(\mathbf{X}; K \mid \mathbf{T}) &\leq -\mathbb{E}_{\mathbf{T}} \mathbb{E}_K \log_2 \mathbb{E}_{K_2} \exp\left[\mathbb{E}_{\mathbf{X}|K,\mathbf{T}} \log_2 \frac{\mathbb{P}(\mathbf{X} \mid K_2, \mathbf{T})}{\mathbb{P}(\mathbf{X} \mid K, \mathbf{T})}\right]; \\ &= -\mathbb{E}_{\mathbf{T}} \mathbb{E}_K \log_2 \mathbb{E}_{K_2} \exp\left[-\mathrm{D}(\mathbb{P}_{\mathbf{X}|K,\mathbf{T}} \,||\, \mathbb{P}_{\mathbf{X}|K_2,\mathbf{T}})\right]. \end{aligned}$$

This proves inequality (5) and Theorem 1 [3].

3.5 Numerical Estimation of $I(\mathbf{X}; \mathbf{Y} \mid \mathbf{T})$

Theorem 1 gave analytic bounds to the success rate. However, one may need to obtain a precise value of $I(\mathbf{X}; \mathbf{Y}|\mathbf{T})$ making the bound tighter. In this section, we propose numerical tools to obtain an accurate value of the Mutual Information as a function of the number of queries q. A full estimation of $I(\mathbf{X}; \mathbf{Y} \mid \mathbf{T})$ by numerical integration becomes impossible for q-dimensional distributions, and we have recourse to simplifying approximations of MI. Since

$$\begin{aligned} I(\mathbf{X}; \mathbf{Y} \mid \mathbf{T}) &= H(\mathbf{X} \mid \mathbf{T}) - H(\mathbf{X} \mid \mathbf{Y}, \mathbf{T}) \\ &= H(\mathbf{X} \mid \mathbf{T}) - H(\mathbf{X} \mid \mathbf{Y}) \end{aligned}$$

[3] An alternative proof of inequality (5), which resorts only on convexity arguments, is given in Appendix D.

we can estimate only the entropy $H(\mathbf{X} \mid \mathbf{T})$ because $H(\mathbf{X} \mid \mathbf{Y}) = qH(X \mid Y)$ is easily computable with classical numerical tools.

One possible approximation is from the law of large numbers [CT06, Chapter 3]:

$$H(\mathbf{X} \mid \mathbf{T}) = \lim_{J \to \infty} -\frac{1}{J} \sum_{\mathbf{t} \in \mathcal{T}^q} \sum_{j=1}^{J} \mathbb{P}(\mathbf{t}) \log_2 \mathbb{P}(\mathbf{x}_j \mid \mathbf{t}). \tag{12}$$

Unfortunately, such a computation is not tractable since it involves the sum over all balanced vectors $\mathbf{t}$, which represents $q!$ possibilities. However, we can obtain a good approximation of $H(\mathbf{X} \mid \mathbf{T})$ with only one vector $\mathbf{t}$ form the following

Lemma 6 (A Symmetry Property)**.** *Let $\mathbf{t} = (t_1, \ldots, t_q) \in \mathcal{T}$ and τ be a permutation in $\{1, \ldots, q\}$. Noting $\tau(\mathbf{t}) = (t_{\tau(1)}, \ldots, t_{\tau(q)})$, we have:*

$$H(\mathbf{X} \mid \mathbf{T} = \mathbf{t}) = H(\mathbf{X} \mid \mathbf{T} = \tau(\mathbf{t})). \tag{13}$$

Proof. See Appendix A. □

As a consequence of the symmetry of Lemma 6, one needs only one balanced vector $\mathbf{t}$ to estimate $H(\mathbf{X} \mid \mathbf{T})$. Therefore, by the law of large numbers,

$$H(\mathbf{X} \mid \mathbf{T}) \approx \lim_{J \to \infty} -\frac{1}{J} \sum_{j=1}^{J} \log_2 \mathbb{P}(\mathbf{x}_j \mid \mathbf{t}). \tag{14}$$

This leads to Algorithm 1 to evaluate the entropy $H(\mathbf{X} \mid \mathbf{T})$.

input : A balanced vector $\mathbf{t}$
An integer J
The probability distribution $\mathbb{P}(\mathbf{x} \mid \mathbf{t})$
output : An approximation of $H(\mathbf{X} \mid \mathbf{T})$

```
Hxt ← 0 ;
Generate a secret key byte k* ;
for j ← 0 to J do
    Generate the traces x with the model ;
    Hxt ← Hxt − (1/J) log2 P(x | t);
end
return Hxt
```

Algorithm 1: Computation of the entropy using the law of large numbers.

When the leakage models are not perfectly known (e.g. template attacks), a possible way to estimate Mutual Information is to approximate numerically the distributions. An example is given in [GS18].

Other estimation methods can be used, depending on the distribution of the noise. As an example, for Gaussian noise, we may consider Gaussian mixtures as discussed in [KDOP15].

Such numerical estimations are all the more accurate as J is taken large, which means that they make take a tremendous amount of time to compute. Having $I(\mathbf{X}; \mathbf{T}|\mathbf{T})$ as a function of q, even numerically estimated, is very useful as we have the link between the success rate and the minimum number of traces to reach such probability of success.

3.6 Graphical Comparison

In order to visualize the difference between the two upper bounds given above, we have plotted the mutual information $I(\mathbf{X}; \mathbf{Y} \mid \mathbf{T} = \mathbf{t})$, where $\mathbf{t}$ is a fixed balanced vector. The

leakage model chosen is given by the equation

$$y(k, t_i) = \mathrm{H_w}(\mathrm{S_{box}}(t_i \oplus k)) \qquad (i = 1, 2, \ldots, q)$$

where $\mathrm{H_w}(\cdot)$ is the Hamming weight (of the value written in binary), and $\mathrm{S_{box}}(\cdot)$ is the AES substitution box [NIS01]. We suppose that the zero-mean additive white Gaussian noise (AWGN) has standard deviation $\sigma = 4$. This gives a signal-noise ratio $\mathsf{SNR} = 1/8$.

Figure 3 shows the results on $I(\mathbf{X}; \mathbf{Y} \mid \mathbf{T} = \mathbf{t})$ obtained by Monte-Carlo simulation. We notice that

- as expected in Subsection 3.3, the first upper bound (4) is linear in q;
- as expected in Subsection 3.4, the second upper bound (5) converges to $H(K) = n = 8$.

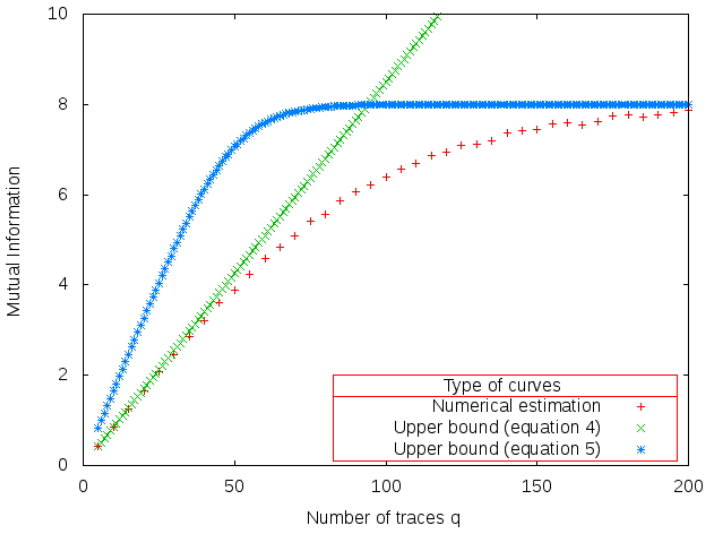


Figure 3: Comparison of the two upper bounds (4) and (5).

4 Application to Additive White Gaussian Noise

In this section, we develop the results of Theorem 1 for leakages with additional white Gaussian noise. Indeed, this is the most common case for attacks such as DPA, where the noise comes from the measurement tools.

With this model, we can link the success rate to Shannon's capacity $C = \frac{1}{2}\log(1+\mathsf{SNR})$, and therefore, to the SNR, where $\mathsf{SNR} = \frac{\mathrm{VAR}(Y)}{\sigma^2}$. Moreover, at the end of this section, we will extract a parametric estimation of the Mutual Information where the only parameter to know is the SNR.

Remark 2. With additive white Gaussian noise, the SNR of the traces can also been written as:

$$\mathsf{SNR} = \frac{\mathrm{Var}(Y)}{\sigma^2},$$

where σ is the standard deviation of the noise.

4.1 Shannon's Channel Capacity

Under the additive white Gaussian noise (AWGN) assumption, it is easily seen that the scalar mutual information $I(X;Y \mid T)$ does not exceed Shannon's capacity. Indeed, we have:

$$\begin{aligned} I(X;Y \mid T) &= \mathbb{E}_T I(X;Y \mid T=t); \\ &= \mathbb{E}_T \left[H(X \mid T=t) - H(X \mid Y, T=t) \right]; \\ &= \mathbb{E}_T \left[H(f(T \oplus K) + N \mid T=t) \right] - H(X \mid Y); \\ &= \mathbb{E}_T \left[H(f(t \oplus K) + N) \right] - H(X \mid Y); \\ &= H(f(K) + N) - H(X \mid Y); \\ &\leq \frac{1}{2} \log_2(2\pi e(\mathrm{Var}_K(f(K)) + \mathrm{Var}(N))) - H(X \mid Y); \\ &= \frac{1}{2} \log_2(1 + \mathrm{SNR}). \end{aligned}$$

Combining this with inequality (4) yields a lower bound on the number of traces to reach a given probability of success:

$$q \geq \frac{n + (\mathrm{P}_s - 1)\log_2(2^n - 1) - H_2(\mathrm{P}_s)}{\frac{1}{2}\log_2(1 + \mathrm{SNR})} \tag{15}$$

Remark 3. The number of traces q to be sure to recover the key is lower-bounded by:

$$\lim_{\mathrm{P}_s \to 1} q \geq \frac{n}{\frac{1}{2}\log_2(1 + \mathrm{SNR})}. \tag{16}$$

However, since as we have seen the MI can never be higher than $H(K)$, the above constant bound is not accurate for real attacks. The next subsection provides a much more accurate estimation.

4.2 Evaluation of the Kullback-Leibler Divergence

Inequality (5) gives an upper bound with a divergence term that depends on $\mathbb{P}_{\mathbf{X}|K_i,\mathbf{T}}$ $(i = 1, 2)$. In the AWGN model, $\mathbb{P}_{\mathbf{X}|K_i,\mathbf{T}}$ follows a multivariate normal distribution $\mathcal{N}(\mathbf{y}(K_i, \mathbf{T}), \sigma^2 I_q)$. For such distributions, the divergence is very easy to compute as the covariance matrix is diagonal. It is easily found that

$$\mathrm{D}(\mathbb{P}_{\mathbf{X}|K,\mathbf{T}} \| \mathbb{P}_{\mathbf{X}|K_2,\mathbf{T}}) = \frac{\|\mathbf{y}(K, \mathbf{T}) - \mathbf{y}(K_2, \mathbf{T})\|_2^2}{2\sigma^2}.$$

Inequality (5), when applied to the AWGN model, becomes

$$n + (\mathrm{P}_s - 1)\log_2(2^n - 1) - H_2(\mathrm{P}_s) \leq -\mathbb{E}_{\mathbf{T}} \mathbb{E}_K \log_2 \mathbb{E}_{K_2} \exp\left(-\frac{\|\mathbf{y}(K, \mathbf{T}) - \mathbf{y}(K_2, \mathbf{T})\|_2^2}{2\sigma^2} \right).$$

In order to make a precise evaluation of the r.h.s., we need several lemmas.

Lemma 7. *Let* $\mathbf{t} = (t_1, \ldots, t_q) \in \mathcal{T}^q$ *and* $(k_1 \neq k_2) \in \mathcal{K}^2$. *One has*

$$\lim_{q \to \infty} \|\mathbf{y}(k_1, \mathbf{t}) - \mathbf{y}(k_2, \mathbf{t})\|_2^2 = +\infty \tag{17}$$

and more precisely:

$$\|\mathbf{y}(k_1, \mathbf{t}) - \mathbf{y}(k_2, \mathbf{t})\|_2^2 \underset{q \to \infty}{\sim} q.\alpha(k_1, k_2), \tag{18}$$

where $\alpha(k_1, k_2) = \frac{1}{2^n} \sum_{t=0}^{2^n - 1} (y(k_1, t) - y(k_2, t))^2$.

Proof. We make use of the assumption made in Section 2 that $\mathbf{T}$ is *balanced*. For $k_1 \neq k_2$, we have

$$\begin{aligned}\|\mathbf{y}(k_1,\mathbf{t}) - \mathbf{y}(k_2,\mathbf{t})\|_2^2 &= \sum_{i=1}^{q} (\mathbf{y}(k_1,t_i) - \mathbf{y}(k_2,t_i))^2; \\ &= q \sum_{i=1}^{q} \frac{(\mathbf{y}(k_1,t_i) - \mathbf{y}(k_2,t_i))^2}{q}; \\ &= q \sum_{t\in\mathcal{T}} \frac{n_t(\mathbf{y}(k_1,t_i) - \mathbf{y}(k_2,t_i))^2}{q};\end{aligned}$$

where n_t is the number of times that a particular $t \in \mathcal{T}$ appears in vector $\mathbf{t}$. As $\mathbf{t}$ is balanced, $\frac{n_t}{q} \to \frac{1}{|\mathcal{T}|}$ and therefore:

$$\|\mathbf{y}(k_1,\mathbf{t}) - \mathbf{y}(k_2,\mathbf{t})\|_2^2 \underset{q\to\infty}{\sim} q \sum_{t\in\mathcal{T}} \frac{(\mathbf{y}(k_1,t_i) - \mathbf{y}(k_2,t_i))^2}{|\mathcal{T}|}.$$

□

Lemma 8. *Let $\mathbf{t} \in \mathcal{T}^q$ be fixed and $k \in \mathcal{K}$ be a fixed key. We have*

$$\lim_{q\to\infty} -\log_2 \mathbb{E}_{K_2} \exp\left(-\frac{\|\mathbf{y}(k,\mathbf{t}) - \mathbf{y}(K_2,\mathbf{t})\|_2^2}{2\sigma^2}\right) = n \tag{19}$$

$$-\log_2 \mathbb{E}_{K_2} \exp\left(-\frac{\|\mathbf{y}(k,\mathbf{t}) - \mathbf{y}(K_2,\mathbf{t})\|_2^2}{2\sigma^2}\right) \underset{q\to\infty}{\sim} -\log_2 \mathbb{E}_{K_2} \exp\left(-\frac{q.\alpha(k,K_2)}{2\sigma^2}\right). \tag{20}$$

Proof. One has

$$-\log_2 \mathbb{E}_{K_2} \exp\left(-\frac{\|\mathbf{y}(k,\mathbf{t}) - \mathbf{y}(K_2,\mathbf{t})\|_2^2}{2\sigma^2}\right) = -\log_2 \left[\sum_{k_2} \frac{1}{2^n} \exp\left(-\frac{\|\mathbf{y}(k,\mathbf{t}) - \mathbf{y}(k_2,\mathbf{t})\|_2^2}{2\sigma^2}\right)\right]$$

When q is a multiple of 2^n we have exactly

$$\|\mathbf{y}(\mathbf{t},k_1) - \mathbf{y}(\mathbf{t},k_2)\|_2^2 = q.\alpha(k_1,k_2)$$

and the proof of Equation (20) is trivial. Otherwise, for $k \neq k_2$ we have $\exp(-q\frac{\alpha(k,k_2)}{2\sigma^2}) \to 0$ as $q \to \infty$; and for $k = k_2$ we have $\exp(-q\frac{\alpha(k,k_2)}{2\sigma^2}) = 1$. Therefore

$$-\log_2 \left[\sum_{k_2} \frac{1}{2^n} \exp\left(-\frac{\|\mathbf{y}(k,\mathbf{t}) - \mathbf{y}(k_2,\mathbf{t})\|_2^2}{2\sigma^2}\right)\right] \longrightarrow n.$$

□

Lemma 9. *With the assumptions made in Section 2, we have as $q \to \infty$:*

$$\mathbb{E}_{\mathbf{T}}\mathbb{E}_K \log_2 \mathbb{E}_{K_2} \exp\left(-\mathrm{D}(\mathbb{P}_{\mathbf{X}|K}||\mathbb{P}_{\mathbf{X}|K_2})\right) \underset{q\to\infty}{\sim} n - \frac{n_{\min}}{2^n} \exp(-q. \min_{k_1\neq k_2} \alpha(k_1,k_2)) \tag{21}$$

where $n_{\min}$ is the number of indexes $k_1 \neq k_2$ reaching the minimum value of $\alpha(k_1,k_2)$.

This simple asymptotic expression can be used to upper-estimate the MI for high values of q. Notice that for any $k_1 \neq k_2$, $\alpha(k_1,k_2) = \alpha(k_2,k_1)$, hence $n_{\min}$ is an even number.

Proof. Let $\mathbf{t} = (t_1, \ldots, t_q)$ be a balanced vector. By Lemma 8, we have

$$-\mathbb{E}_K \log \mathbb{E}_{K_2} \exp\left(-\mathrm{D}(\mathbb{P}_{\mathbf{X}|K\mathbf{t}} || \mathbb{P}_{\mathbf{X}|K_2\mathbf{t}})\right) \underset{q\to\infty}{\sim} -\mathbb{E}_K \log \mathbb{E}_{K_2} \exp\left(-\frac{q.\alpha(K,K_2)}{2\sigma^2}\right)$$

where

$$\begin{aligned}-\mathbb{E}_K \log \mathbb{E}_{K_2} \exp\left(-\frac{q.\alpha(K,K_2)}{2\sigma^2}\right) &= -\mathbb{E}_K \log\left[\frac{1}{2^n}\sum_{k_2} \exp\left(-\frac{q.\alpha(K,k_2)}{2\sigma^2}\right)\right]; \\ &= n - \mathbb{E}_K \log\left[1 + \sum_{k_2\neq K} \exp\left(-\frac{q.\alpha(K,k_2)}{2\sigma^2}\right)\right].\end{aligned}$$

As the value inside the logarithm vanishes as $q \to \infty$, consider its first-order Taylor expansion:

$$\begin{aligned}-\mathbb{E}_K \log \mathbb{E}_{K_2} \exp\left(-\frac{q.\alpha(K,K_2)}{2\sigma^2}\right) &\underset{q\to\infty}{\sim} n - \mathbb{E}_K\left[\sum_{k_2\neq K} \exp\left(-\frac{q.\alpha(K,k_2)}{2\sigma^2}\right)\right]; \\ &= n - \frac{1}{2^n}\sum_{k_1\neq k_2}\left[\exp\left(-\frac{q.\alpha(K,k_2)}{2\sigma^2}\right)\right].\end{aligned}$$

Let $k_1 \neq k_2$ be a couple such that $\alpha(k_1, k_2)$ is the minimum of all the possible α. For any other couple $k_3 \neq k_4$, there are two possibilities:

1. either $\alpha(k_3, k_4) = \alpha(k_1, k_2)$ and the corresponding exponentials will converge at the same rate;
2. or $\alpha(k_3, k_4) > \alpha(k_1, k_2)$ and $\exp\left(-\frac{q}{2\sigma^2}\alpha(k_3, k_4)\right)$ is negligible w.r.t. $\exp\left(-\frac{q}{2\sigma^2}\alpha(k_1, k_2)\right)$.

Hence we can simply count the number of occurrences of the minimum value of α. We have proven that:

$$-\mathbb{E}_K \log \mathbb{E}_{K_2} \exp\left(-\mathrm{D}(\mathbb{P}_{\mathbf{X}|K\mathbf{t}} || \mathbb{P}_{\mathbf{X}|K_2\mathbf{t}})\right) \underset{q\to\infty}{\sim} n - \frac{n_{\min}}{2^n} \exp\left(-\frac{q.\min_{k_1\neq k_2}\alpha(k_1,k_2)}{2\sigma^2}\right).$$

As this expansion is true for any vector $\mathbf{t}$ that is balanced, and is independent of it, this proves the lemma. □

Remark 4. The simplification of Lemma 9 is useful to obtain a simple equivalent form for high values of q. However, it is also possible to compute a tight approximation of the numerical value of $\mathbb{E}_{\mathbf{T}}\mathbb{E}_K \log_2 \mathbb{E}_{K_2} \exp\left(-\mathrm{D}(\mathbb{P}_{\mathbf{X}|K} || \mathbb{P}_{\mathbf{X}|K_2})\right)$.

Remark 5. Interestingly, we notice that parameter $\alpha(k_1, k_2)$ is proportional to the *confusion coefficient* $\kappa(k_1, k_2)$ defined first in [FLD12] for binary leakages, and extended in [GHR15, Equation (45)] for any leakage:

$$\kappa(k_1, k_2) = 4\alpha(k_1, k_2).$$

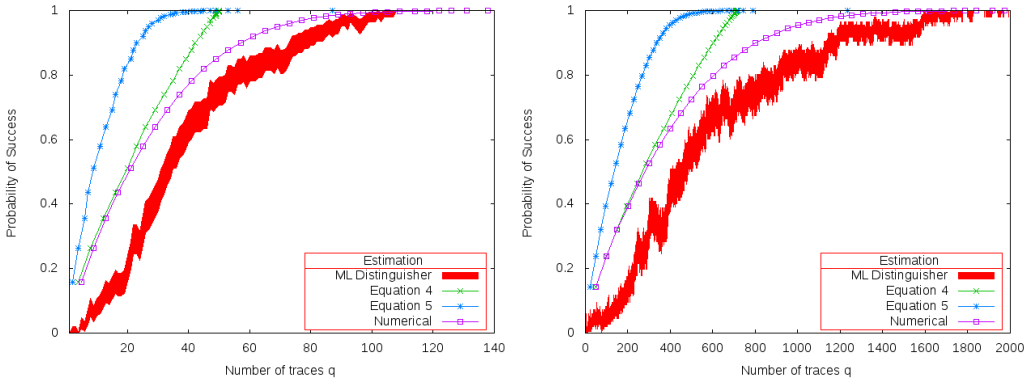
4.3 Example for Monobit Leakage

In this subsection, we consider a *monobit* leakage model:

$$f(t_i \oplus k) = \mathrm{LSB}(\mathrm{S}_{\mathrm{box}}(t_i \oplus k)) \qquad (i = 1, 2, \ldots, q)$$

where $\mathrm{S}_{\mathrm{box}}$ is the AES substitution box and LSB is the least significant bit of a number. Figure 4 represents the success rate of a monobit leakage with additive Gaussian noise (standard deviation $\sigma = 4$). The distinguisher used is the maximum likelihood distinguisher which is optimal [HRG14]. The other curves are the bounds obtained with:

- a numerical estimation of $I(\mathbf{X}; \mathbf{Y} \mid \mathbf{T})$ (using the law of large numbers, as described in Section 3.5);
- MI's upper bound (4);
- MI's upper bound (5).

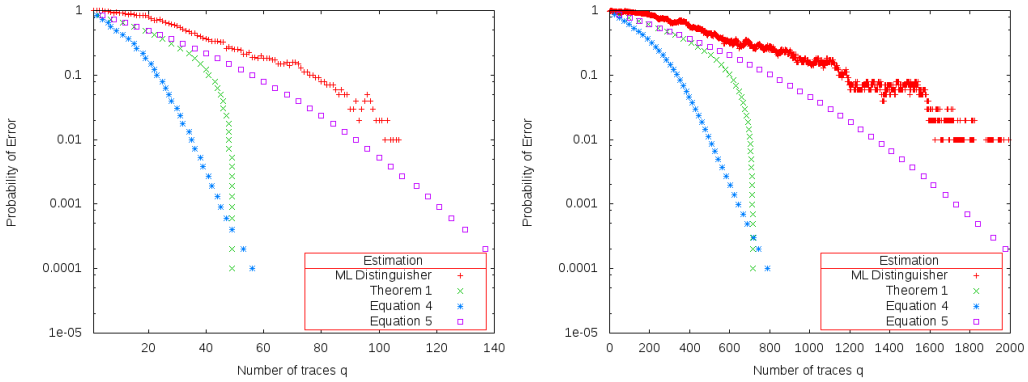


(a) $\sigma = 1$ (b) $\sigma = 4$

Figure 4: Success rates with monobit leakage.

The three bounds curves lie above the success rate curve as expected, the one obtained with a numerical estimation of $I(\mathbf{X}; \mathbf{Y} \mid \mathbf{T})$ being the tightest (since it gives the closest approximation of the MI). The two other curves obtained with Equations (4) and (5) are not as tight but very easy to calculate. Theses results show that the better approximation of the MI we have, the closer we are from the optimal success rate.

In Figure 5, we have plotted the error rate in a semilog scale, so that one can observe that the curves obtained with Equations (4) and (5) actually cross each other. This shows



(a) $\sigma = 1$ (b) $\sigma = 4$

Figure 5: Error rate for a monobit leakage in a logarithmic scale

that, closer to $\mathrm{P}_s = 1$ it is more interesting to choose the approximation of Equation (5), rather than Equation (4).

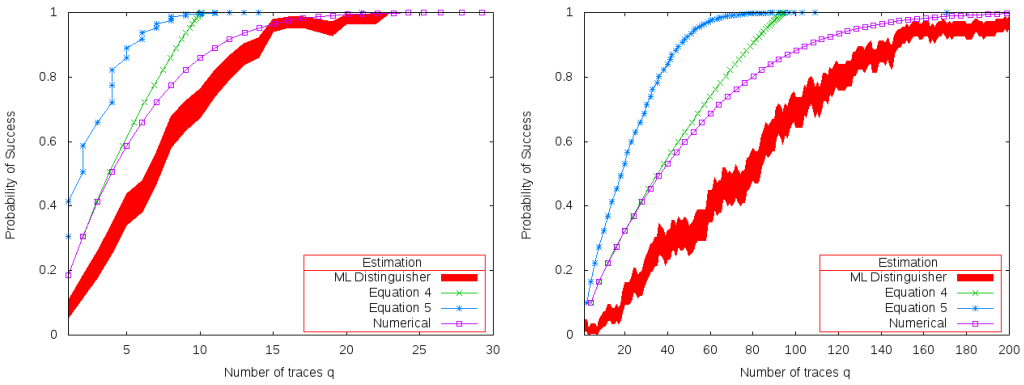
Remark 6. For this leakage model, with a balanced vector $\mathbf{t}$, one needs at least 8 traces to obtain 256 different vectors $\mathbf{y}$, since the function $k \mapsto \mathbf{y}(k)$ is one-to-one.

4.4 Example for Hamming Weight Leakage

In practice, the AES algorithms compute SubBytes with 8 bits. The leakage function are therefore different if we take this into account. Our conclusion is the same. We now consider the leakage model based on the Hamming Weight:

$$y_i = f(t_i \oplus k) = \mathrm{H_w}(\mathrm{S_{box}}(t_i \oplus k)) \qquad (i = 1, 2, \ldots, q)$$

where $\mathrm{S_{box}}$ is the AES substitution box and $\mathrm{H_w}$ is the Hamming weight function. Figure 6 shows the success rate compared with the three other types of estimation with an additive Gaussian noise with two values of standard deviation σ. For this model, we recall that $\mathsf{SNR} = 2/\sigma^2$. Once again, we notice that our bounds are above the optimal distinguisher



(a) $\sigma = 1$ (b) $\sigma = 4$

Figure 6: Success rate for a Hamming weight leakage

and that the closest estimation of the MI gives the tightest bound.

4.5 A Parametric Estimation of $I(\mathbf{X}; \mathbf{Y} \mid \mathbf{T})$

An estimation of $I(\mathbf{X}; \mathbf{Y} \mid \mathbf{T})$ with a simple analytic expression can be obtained by a parametric estimation of the mutual information. This study is based on an empirical model that fits correctly with $I(\mathbf{X}; \mathbf{Y} \mid \mathbf{T})$. The information function $I(q) = I(\mathbf{X}; \mathbf{Y} \mid \mathbf{T})$ has been matched against some classical shapes (e.g., $1 - e^{-q \cdot \alpha}$, as hinted in [GHR15]) with poor accuracy. We found that $I(q) = I(\mathbf{X}; \mathbf{Y} \mid \mathbf{T})$ is best approximated by the error function such as:

$$I(q) \approx n \cdot \mathrm{erf}(q \cdot \alpha), \tag{22}$$

where α is a constant, and erf the error function defined as:

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \mathrm{d}t.$$

In order to verify this hypothesis numerically, for a Hamming weight leakage with additive Gaussian noise, we have plotted in Figure 7 the estimated parameter α for different values of σ and different number of traces. The mutual information is estimated using the law of large numbers and therefore, the parameter α is obtained by:

$$\alpha = \frac{\mathrm{erf}^{-1}(I(\mathbf{X}; \mathbf{Y} \mid \mathbf{T})/n)}{q}$$

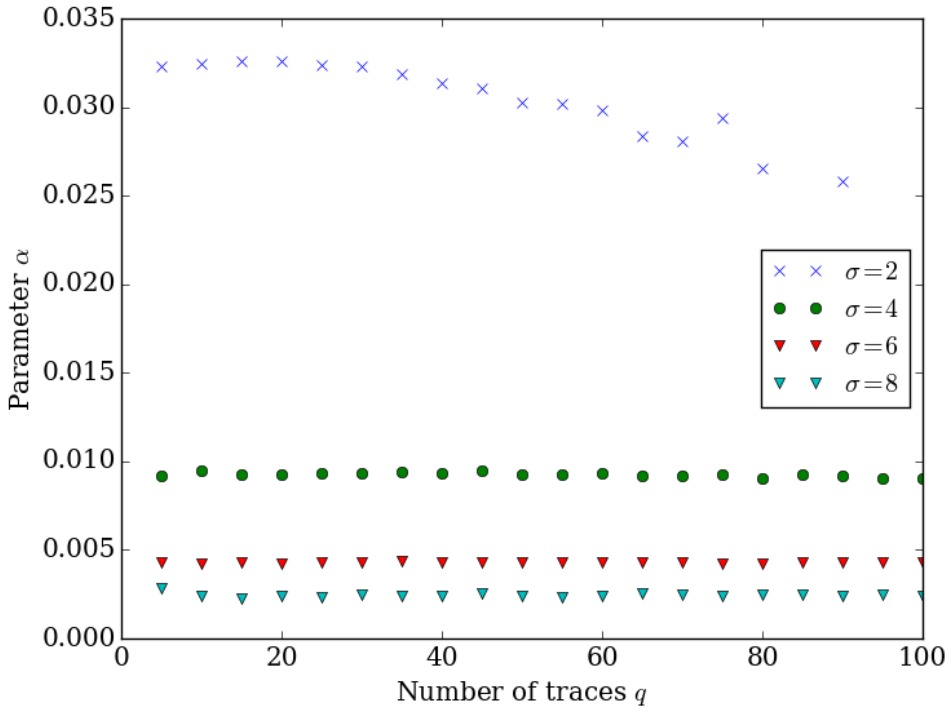


Figure 7: Estimation of parameter α

Notice that for each value of σ, α is constant, which suggest that our empirical model fits the MI well.

We can go even further and find the analytic value of α. Indeed, the first order derivative of our model is $n\alpha\frac{2}{\sqrt{\pi}}e^{-q^2}$, therefore, the slope at the origin is $n\alpha\frac{2}{\sqrt{\pi}}$. We know that $I(0) = 0$ and $I(1) = I(X;Y \mid T) \approx \frac{1}{2}\log_2(1+\mathsf{SNR})$. This means that if we approximate $\frac{\partial I(q)}{\partial q}(0)$ by $I(1) - I(0)$, we have:

$$\frac{1}{2}\log_2(1+\mathsf{SNR}) = n\alpha\frac{2}{\sqrt{\pi}}, \tag{23}$$

that is:

$$\alpha = \frac{\sqrt{\pi}}{4n}\log_2(1+\mathsf{SNR}). \tag{24}$$

Therefore, given the value of the SNR, one can predict the value of MI for additive Gaussian noise. We can see that the approximation (22) holds very well for $\sigma > 2$. This happens for low values of SNR as we encounter in practice when evaluating cryptographic devices. The number of traces needed to reach a given success rate P_s is therefore lower-bounded by:

$$q \geq \frac{4n}{\sqrt{\pi}\log_2(1+\mathsf{SNR})}\,\mathrm{erf}^{-1}\Big(\frac{n - H_2(\mathrm{P}_s) - (1-\mathrm{P}_s)\log_2(2^n-1)}{n}\Big). \tag{25}$$

The interest of such bound is that it requires only the knowledge of an additive Gaussian noise and the calculation of the SNR to be exploited and to therefore predict a tight bound on the number of traces to reach a given success rate.

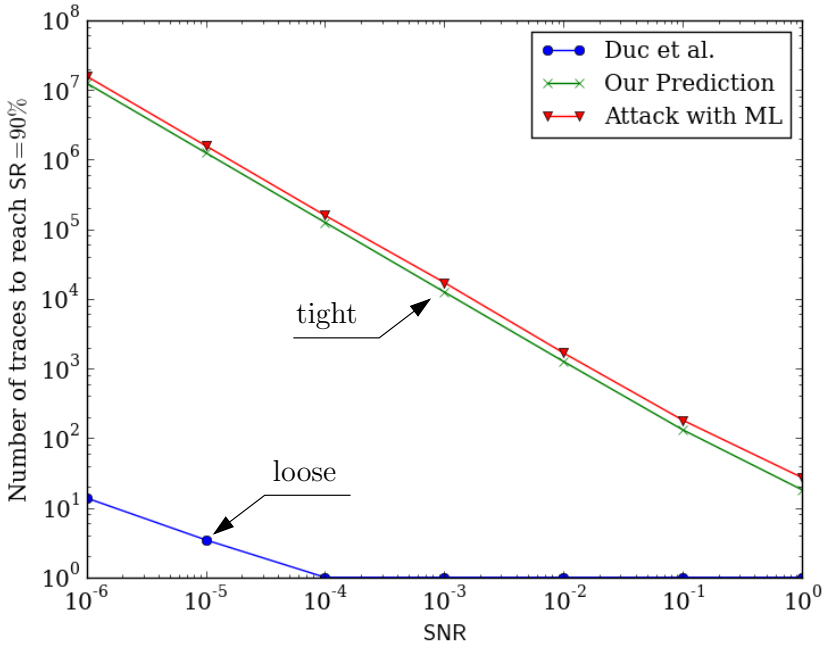


Figure 8: Comparison of our prediction with Duc's bound

4.6 Comparison with Duc's Bound

In order to show that our bounds are tight, we have plotted the number of traces needed to reach a success rate of 90% for a monobit leakage with additional Gaussian noise (same leakage as Section 4.3). In this figure, we compare our bound with the ML distinguisher and the success rate proposed by Duc et al. in [DFS15]. We recall that the ML distinguisher is the best distinguishing rule when the model is known, therefore, the best possible case for the attacker. Mathematically we have:

$$\mathcal{D}_{ML}(\mathbf{x}, \mathbf{t}) = \arg\max_k \mathbb{P}(\mathbf{x} \mid \mathbf{t} \oplus k). \tag{26}$$

To compute our bound, we only suppose that the noise is AWGN and we apply the parametric estimation of the SNR, proposed in the previous subsection (cf. Equation (25)). With AWGN the optimal distinguisher becomes:

$$\mathcal{D}_{ML}(\mathbf{x}, \mathbf{t}) = \arg\min_k \sum_{i=1}^{q} (x_i - y_i(k))^2. \tag{27}$$

In Figure 8, we notice that our bound is always very close to the real success rate, calculated for the best case for the attacker. This means that our predictions give a good idea of the security of any device, and we recall that this prediction has been made with the only knowledge of a Gaussian noise. Therefore, with very low assumptions and very few measurements (needed to calculate the SNR), we are able to predict the number of traces to reach a given success rate with a good approximation. Moreover, our bound is above Duc's bound. This means that our prediction is better.

5 Practical Applications

In practice, the computation of a lower bound on the number of traces, such as that given in equation (25), relies on the value of the SNR. Therefore, it is crucial to estimate the SNR accurately, so as to have a trustworthy bound of the device protection level. In this section, we first propose an algorithm that extracts the SNR of a leakage. Second, in order to compare our results with real world data sets, we apply our method to that obtained within the framework of the "DPA Contest" challenge.

5.1 The SNR estimation

In order to apply Theorem 1 or Equation (25) with the parametric estimation of the Mutual Information, one shall estimate the SNR of the leakage. When the leakage is monovariate, meaning that the attacker has at her disposal one share of the leakage, it is possible to estimate the SNR on-the-fly. The SNR of the leakage can be written as follows:

$$\begin{aligned}\mathsf{SNR} &= \frac{\mathrm{Var}(Y)}{\mathrm{Var}(N)} \\ &= \frac{\mathrm{Var}(Y)}{\mathrm{Var}(X-Y)} \\ &= \frac{\mathrm{Var}(Y)}{\mathrm{Var}(X)-\mathrm{Var}(Y)}.\end{aligned}$$

We also notice that since $X = Y + N$, where the noise N is independent from the signal Y (which depends only on the plain/cipher-text T), we have $Y = \mathbb{E}[X \mid T]$. This means that the SNR can be estimated with:

$$\mathsf{SNR} = \frac{\mathrm{Var}(\mathbb{E}[X \mid T])}{\mathrm{Var}(X) - \mathrm{Var}(\mathbb{E}[X \mid T])}. \tag{28}$$

In this equation, the expression $\mathrm{Var}(\mathbb{E}[X \mid T])$ is the leakage inter-class variance. The equation (28) is valid for algorithms such as AES, since the leakage model of AES does not depend on anything else than the 8 bits of (each individual byte of) the plaintext T.

When the leakage is multivariate, it is possible to compute dimensionality reduction (c.f. [BGH+15, Corollary 4]). In such case, a profiling phase is needed to estimate the noise covariance matrix. Besides, other methods to estimate the SNR can be used such as Linear Discriminant Analysis (LDA) [SA08].

5.2 A Real World Case: the DPA Contest

In order to compare our theoretical results with practical evaluations, we used the data set of the DPA Contest v1 [TEL09]. In the first version of this contest, the goal is to recover the 56-bit key of the DES encrypting algorithm. The device is a Side-channel Attack Standard Evaluation Board (SASEBO) developed by the Japan AIST / RCIS.

According to the data given in the DPA contest, the attacker has at her disposal a high number of traces, each made up of 20003 samples. An example is given in Fig. 9. We will consider here the first round of the algorithm (some attacks consider the last round but the results are very similar).

For example, we have plotted in Fig. 10 the SNR of this leakage considering the first substitution box. In this figure, we notice that the maximum value of the SNR is 0.144 but we notice that other points of interest may be used.

We have computed a simple CPA on the first round of DES with this data set to recover 6 bits of key. Figure 11 shows the partial success rate for all the substitution boxes. This success rate has been obtained with 100 experiments. We have plotted the CPA for

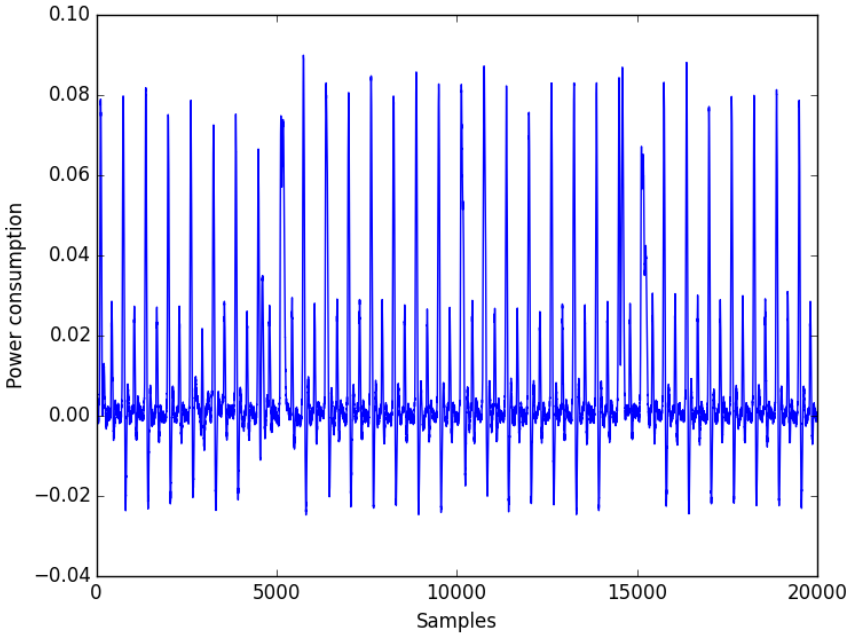


Figure 9: One trace of DES leakage (from DPA contest v1 [TEL09])

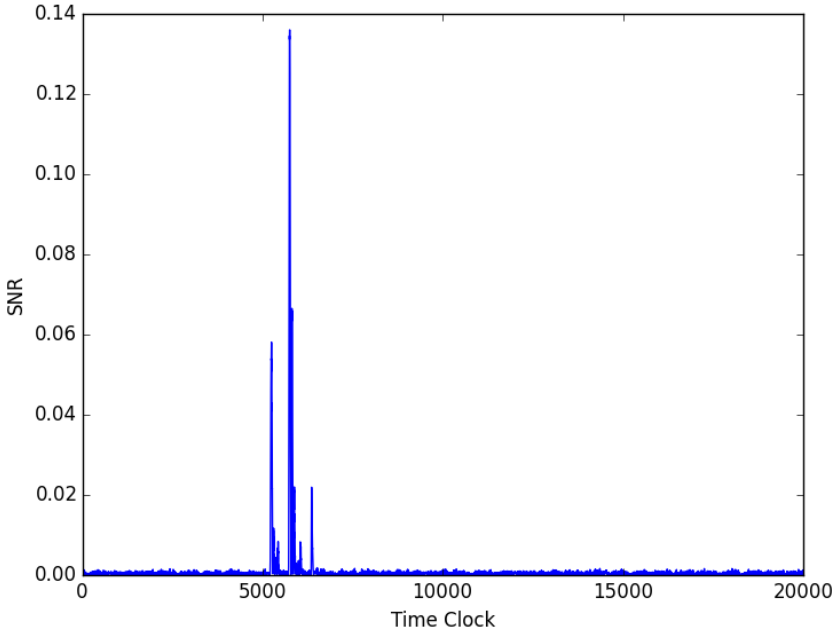


Figure 10: SNR of the first Sbox for the first round of DES.

Table 2: SNR for each Sbox for the DPA contest

Sbox #	SNR	Prediction for 99%	CPA 99%
1	0.144	112	230
2	0.077	203	350
3	0.075	208	350
4	0.071	220	450
5	0.064	243	300
6	0.151	107	190
7	0.079	198	330
8	0.136	118	270

the best time sample (the one that maximizes the SNR) in the green curve and the CPA over all the time samples (the blue curve). The red curves corresponds to the bound of Equation 25. The SNR of each substitution box is reported in Table 2. This table also recall how many traces are needed to get a 99% success rate for each sbox with a simple CPA.

According to the figures of the table, without any pre-processing the attacker will need at least 243 traces to recover the secret key with one sample. This corresponds to the results obtained without pre-processing (cf. http://www.dpacontest.org/hall_of_fame.php) .

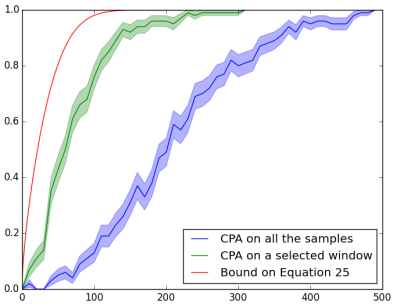
Besides the interest that traces are available and can be used as a benchmark, the DPA contest v1 has also given rise to a serious competition amongst submitters to efficient attacks. We notice that attacks which target (like we did) leakage at S-box of the first round[4] all recover the keys with more traces than our bound. Besides, let us consider the Build-up Sub-keys Correlation Power Analysis (BS-CPA [KSK09]) where the attacker takes into account one broken subkey to recover others. Since the S-Box that is the easiest to break is 6th, this attack (according to our bound) shall require more than 107 traces to succeed. This is what is observed in the DPA contest v1, where Komano, Shimizu and Kawamura require 134 traces to extract the key. In summary, we see that our bounds (represented in Fig. 11) show that the contest has delivered interesting attacks, close to the theoretical bounds, and thus that little margin for improvement was possible.

In addition, it is possible to increase the SNR by selecting several samples (dimensionality reduction). For example, in our case, selecting two samples at each trace (respectively the samples that correspond to the two highest peaks of Figure 10) leads to an SNR of 0.228 for the first substitution box. Such multivariate SNR gives at least 73 traces to reach 99% of success rate.

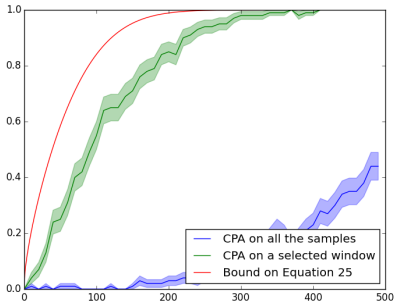
Remark 7. The winning attack [Cla09] combines leakage from the first and the last round of DES, with three samples at each round. This leads to a key recovery in 45 traces in average.

Second Version of the DPA Contest The second version of the DPA Contest (also known as DPAv2) took place between 2010 and 2013 [TEL10]. The targeted device was a FPGA with an unprotected version of AES running on it (cf. http://www.dpacontest.org/v2/documentation.php). Once again, we apply our prediction based on the parametric estimation of the Mutual Information. The estimation of the SNR is made thanks to the template databases and the attack is done on the last round of the algorithm with the ML distinguisher on the sample with the highest SNR. For the 16th byte of the secret key, the success rate and our prediction is shown in Figure 12.

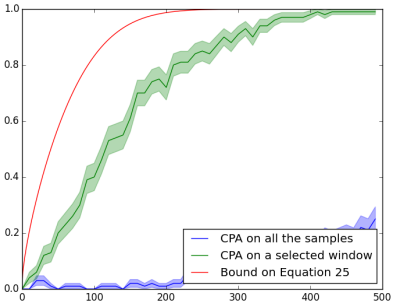
[4] Attack by Hideo SHIMIZU (272 traces), Antonio SOBREIRA and Dejan LAZICH (267 traces).



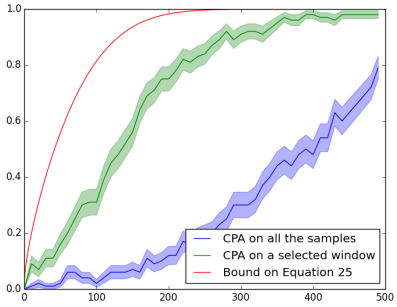
(a) Sbox # 1



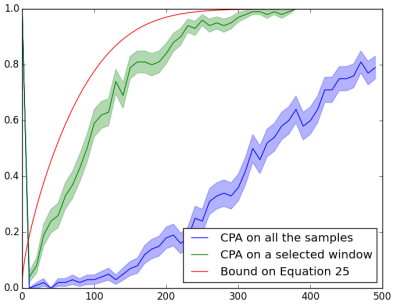
(b) Sbox # 2



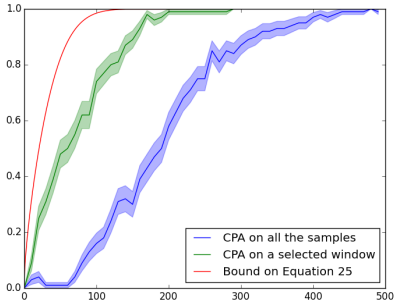
(c) Sbox # 3



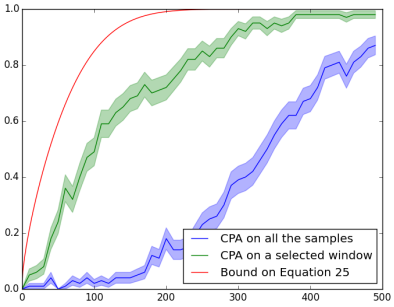
(d) Sbox # 4



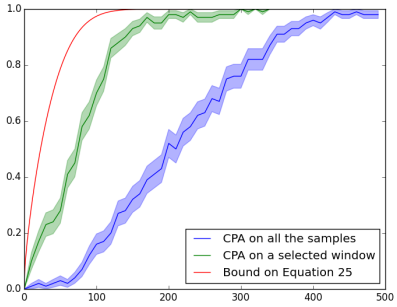
(e) Sbox # 5



(f) Sbox # 6



(g) Sbox # 7



(h) Sbox # 8

Figure 11: Success rates for CPA on the complete trace and on selected points of interest

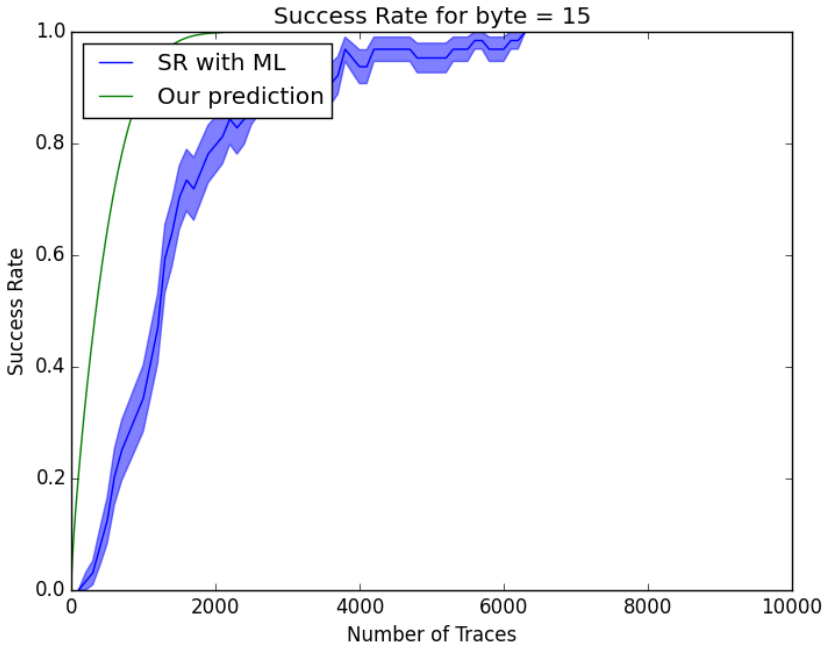


Figure 12: Success Rate vs. our prediction for the 16th byte of the secret key

6 Link with Guessing Entropy

Another way to quantify the quality of an attack is the *Guessing Entropy* [Mas94], defined as $H(K \mid \mathbf{X}, \mathbf{T})$. This metric quantifies the complexity of the exclusive search to recover K knowing the side-channel measurements. Besides, let N_K be the average number of tries to retrieve the secret key K with the knowledge of $\mathbf{X}$ and $\mathbf{T}$. Mathematically, we have:

$$N_K = \mathbb{E}_{\mathbf{XT}}\left[\sum_k \delta_{\mathbf{XT}}(k)\mathbb{P}(k \mid \mathbf{X}, \mathbf{T})\right],$$

where $\delta_{\mathbf{XT}}(\cdot)$ is the permutation that re-orders the probabilities $\mathbb{P}(k \mid \mathbf{X}, \mathbf{T})$ into the decreasing order. There exists a relationship between N_K and $H(K \mid \mathbf{X}, \mathbf{T})$ called the inequality of Massey [Mas94, Section 2]:

$$N_K \geq 2^{H(K|\mathbf{X},\mathbf{T})-2} + 1.$$

We propose here an improved inequality relating N_k with $H(K \mid \mathbf{X}, \mathbf{T})$.

Lemma 10 (Improved Inequality of Massey)**.** *The average number of tries to recover the correct key is upper-bounded by:*

$$N_K > \frac{2^{H(K|\mathbf{X},\mathbf{T})}}{e}. \tag{29}$$

Our inequality improves Massey's inequality as soon as the entropy is greater than $\log_2(\frac{e}{1-e/4})$.

Proof. Let $b_k = \frac{(1-1/N_K)^k}{N_K - 1}$ for all $k \in \mathbb{N}^*$. As $\sum_k b_k = 1$, b_k is a distribution (geometric).

Moreover, by the Gibbs inequality [CT06],

$$\begin{aligned}
H(K \mid \mathbf{X}, \mathbf{T}) &= -\sum_{\mathbf{t},\mathbf{x}} \mathbb{P}(\mathbf{tx}) \sum_{k} \mathbb{P}(k \mid \mathbf{t}, \mathbf{x}) \log_2 \mathbb{P}(k \mid \mathbf{t}, \mathbf{x}) \\
&\leq -\sum_{\mathbf{t},\mathbf{x}} \mathbb{P}(\mathbf{t}, \mathbf{x}) \sum_{k} \mathbb{P}(k \mid \mathbf{t}, \mathbf{x}) \log_2 b_{\delta_{\mathbf{XT}}(k)} \\
&= -\sum_{\mathbf{t},\mathbf{x}} \mathbb{P}(\mathbf{t}, \mathbf{x}) \sum_{k} \mathbb{P}(k \mid \mathbf{t}, \mathbf{x}) \delta_{\mathbf{X},\mathbf{T}}(k) \log_2(1 - 1/N_K) + \log_2(N_K - 1) \\
&= -\log_2(1 - 1/N_K) N_K + \log_2(N_K - 1) \\
&= N_K H_2(1/N_K)
\end{aligned}$$

In fact, the inequality is strict since equality would hold if and only if $\mathbb{P}(k \mid \mathbf{X}, \mathbf{T}) = b_{\delta_{\mathbf{X},\mathbf{T}}(k)}$, which is not the case as the support of $\mathbb{P}$ is finite and the support of b_k is not. Therefore, we have proven that:

$$H(K \mid \mathbf{X}, \mathbf{T}) < N_K H_2(1/N_K).$$

Last, we notice that the function $f(x) = x \log_2(x)$ is convex ($f'(x) = \log_2(ex)$ is increasing). Therefore, fore any x in the range $]0, 1[$, we have:

$$\frac{f(x) - f(x-1)}{x - (x-1)} \leq f'(x) = \log_2(ex).$$

When we apply this for $x = N_K$, we get:

$$\begin{aligned}
N_K H_2(1/N_K) &= N_K \log_2(N_K) - (N_K - 1) \log_2(N_K - 1) \\
&\leq \log_2(eN_K).
\end{aligned}$$

Overall, this means that $H(K \mid \mathbf{X}, \mathbf{T}) < \log_2(eN_K)$ which proves the lemma. □

The lemma can be exploited by replacing $H(K \mid \mathbf{X}, \mathbf{T})$ by $\log_2(eN_K)$ in Subsection 3.2. Therefore, instead of using Fano's inequality, we directly have

$$I((K, \mathbf{T}); (\mathbf{X}, \mathbf{T})) \geq H(K) + H(\mathbf{T}) - \log_2(eN_K),$$

leading to:

$$N_K \geq \frac{2^{-I(\mathbf{X};\mathbf{Y}|\mathbf{T})+H(K)}}{e}. \tag{30}$$

Once more, we can use Theorems (4) and (5) to estimate the mutual information.

For example, we suppose that we have a Gaussian channel, with SNR $= 1/8$ and $q = 40$ traces. We apply Equation (4) to obtain that $I(\mathbf{X}; \mathbf{T} \mid \mathbf{T}) \leq q\frac{1}{2}\log(1 + \text{SNR})$. For a $n = 8$ bits leakage, the average number of tries is lower-bounded by:

$$\begin{aligned}
N_K &\geq \frac{-2^{20*\log_2(1+1/8)+8}}{e} \\
&\approx \frac{2^{4.6}}{e} \\
&\approx 8.9
\end{aligned}$$

This means that, for such a channel, it would take at least 8 tries to recover one byte of the secret key with 40 traces. However, a secret key is made of 16 or even 32 bytes. Supposing that the attacker has only 40 traces for each key-byte, after the attack, one would need at least $8.9^{16} \approx 1.6 \times 10^{15}$ tries in average to recover the entire key as there is no way to check only byte per byte. Note that this method does not deal with key enumeration but brute exhaustive search taking into account the leakage.

7 Conclusion

In this paper, we have linked two metrics used in the field of side-channel analysis: the probability of success of an attack (also known as the success rate) and the mutual information between the leaked traces and the secret key. With such links, designers will be given more precise tools to estimate the security of their cryptographic chips. Our results are of interest to better understand the different factors that impact the success rate of an attack. This is the first time that a study gives *universal* tight bounds to the success rate, in the sense that these bounds are independent of what the attacker may exploit with the measurements.

This is therefore a great improvement for designers. Indeed, in practice they are not able to know how their devices will be attacked in the future, but here, we allow them that to ensure the minimal security of their device in *any* adversarial context.

In addition, the link that we have made with the notion of guessing entropy gives an idea of how many attempts have to be made to recover the key after an attack.

A Proof of Lemma 6

Let $\mathbf{t} \in \mathcal{T}$ and τ be the considered permutation. We have

$$\begin{aligned} H(\mathbf{X} \mid \mathbf{T} = \mathbf{t}) &= -\sum_{\mathbf{x}} \mathbb{P}(\mathbf{x} \mid \mathbf{t}) \log_2 \mathbb{P}(\mathbf{x} \mid \mathbf{t})) \\ &= -\sum_{\mathbf{x}} \left[\sum_k \mathbb{P}(k) \mathbb{P}(\mathbf{x} \mid \mathbf{t}, k) \right] \log_2 \left(\sum_k \mathbb{P}(k) \mathbb{P}(\mathbf{x} \mid \mathbf{t}, k) \right) \\ &= -\sum_{\mathbf{x}} \left[\sum_k \mathbb{P}(k) \prod_{i=1}^{q} \mathbb{P}(x_i \mid t_i, k) \right] \log_2 \left(\sum_k \mathbb{P}(k) \prod_{i=1}^{q} \mathbb{P}(x_i \mid t_i, k) \right) \end{aligned}$$

Re-arranging both products so that they are ordered in accordance with the permutation, we obtain

$$\begin{aligned} H(\mathbf{X} \mid \mathbf{T} = \mathbf{t}) &= -\sum_{\mathbf{x}} \left[\sum_k \mathbb{P}(k) \prod_{i=1}^{q} \mathbb{P}(x_{\tau(i)} \mid t_{\tau(i)}, k) \right] \log_2 \left(\sum_k \mathbb{P}(k) \prod_{i=1}^{q} \mathbb{P}(x_{\tau(i)} \mid t_{\tau(i)}, k) \right) \\ &= -\sum_{\mathbf{x}} \left[\sum_k \mathbb{P}(k) \prod_{i=1}^{q} \mathbb{P}(x_i \mid t_{\tau(i)}, k) \right] \log_2 \left(\sum_k \mathbb{P}(k) \prod_{i=1}^{q} \mathbb{P}(x_i \mid t_{\tau(i)}, k) \right) \\ &= H(\mathbf{X} \mid \mathbf{T} = \tau(\mathbf{t})) \end{aligned}$$

□

B Proof of Equation (5)

We study the sign of the difference

$$\begin{aligned} \Delta &= -\mathbb{E}_Y \log_2 \mathbb{E}_X [\exp(f(X,Y))] + \log_2 \mathbb{E}_X [\exp(\mathbb{E}_Y f(X,Y))]; \\ &= -\log_2 \exp \mathbb{E}_Y \log_2 \mathbb{E}_{X'} [\exp(f(X',Y))] + \log_2 \mathbb{E}_X [\exp(\mathbb{E}_Y \log_2 \exp f(X,Y))]; \\ &= \log_2 \mathbb{E}_X \frac{\exp(\mathbb{E}_Y \log_2 \exp f(X,Y))}{\exp \mathbb{E}_Y \log_2 \mathbb{E}_{X'} [\exp(f(X',Y))]}; \\ &= \log_2 \mathbb{E}_X \exp \mathbb{E}_Y \left[\log_2 \exp f(X,Y) - \log_2 \mathbb{E}_{X'} [\exp(f(X',Y))] \right]; \\ &= \log_2 \mathbb{E}_X \exp \mathbb{E}_Y \left[\log_2 \frac{\exp f(X,Y)}{\mathbb{E}_{X'} [\exp(f(X',Y))]} \right]. \end{aligned}$$

Since the log function is concave:

$$\begin{aligned}
\Delta &\leq \log_2 \mathbb{E}_X \exp \log_2 \mathbb{E}_Y \left[\frac{\exp f(X,Y)}{\mathbb{E}_{X'}[\exp(f(X',Y))]} \right]; \\
&= \log_2 \mathbb{E}_X \mathbb{E}_Y \left[\frac{\exp f(X,Y)}{\mathbb{E}_{X'}[\exp(f(X',Y))]} \right]; \\
&= \log_2 \mathbb{E}_Y \left[\frac{\mathbb{E}_X \exp f(X,Y)}{\mathbb{E}_{X'} \exp(f(X',Y))} \right]; \\
&= \log_2 \mathbb{E}_Y [1]; \\
&= 0.
\end{aligned}$$

□

C Proof of Corollary 1

In Lemma 5, we have proven that:

$$\mathbb{E}_Y \log_2 \mathbb{E}_X[\exp(f(X,Y))] \geq \log_2 \mathbb{E}_X[\exp(\mathbb{E}_Y f(X,Y))]$$

or

$$\mathbb{E}_Y \log_2 \mathbb{E}_X[\exp(f(X,Y))] \geq \log_2 \mathbb{E}_X[\exp(\mathbb{E}_Y \log \exp f(X,Y))].$$

Setting $g(x,y) = \exp(f(x,y))$, we have:

$$\mathbb{E}_Y \log_2 \mathbb{E}_X[g(X,Y)] \geq \log_2 \mathbb{E}_X[\exp(\mathbb{E}_Y \log g(X,Y))].$$

Hence,

$$\begin{aligned}
\exp \mathbb{E}_Y \log_2 \mathbb{E}_X[g(X,Y)] &\geq \exp \log_2 \mathbb{E}_X[\exp(\mathbb{E}_Y \log g(X,Y))] \\
&\geq \mathbb{E}_X[\exp(\mathbb{E}_Y \log g(X,Y))]
\end{aligned}$$

□

D Alternative Proof of (5) and Further Comments

Consider, for any random vector $\mathbf{Y}'$,

$$\begin{aligned}
\Delta &= I(\mathbf{X};\mathbf{Y}) + \mathbb{E}_{\mathbf{Y}} \log \mathbb{E}_{\mathbf{Y}'} \exp \left(\mathbb{E}_{\mathbf{X}|\mathbf{Y}} \log \frac{\mathbb{P}(\mathbf{X} \mid \mathbf{Y}')}{\mathbb{P}(\mathbf{X} \mid \mathbf{Y})} \right) \\
&= \mathbb{E}_{\mathbf{Y}} \mathbb{E}_{\mathbf{X}|\mathbf{Y}} \log \frac{\mathbb{P}(\mathbf{X} \mid \mathbf{Y})}{\mathbb{P}(\mathbf{X})} + \mathbb{E}_{\mathbf{Y}} \log \mathbb{E}_{\mathbf{Y}'} \exp \left(\mathbb{E}_{\mathbf{X}|\mathbf{Y}} \log \frac{\mathbb{P}(\mathbf{X} \mid \mathbf{Y}')}{\mathbb{P}(\mathbf{X} \mid \mathbf{Y})} \right) \\
&= \mathbb{E}_{\mathbf{Y}} \log \exp \mathbb{E}_{\mathbf{X}|\mathbf{Y}} \log \frac{\mathbb{P}(\mathbf{X} \mid \mathbf{Y})}{\mathbb{P}(\mathbf{X})} + \mathbb{E}_{\mathbf{Y}} \log \mathbb{E}_{\mathbf{Y}'} \exp \left(\mathbb{E}_{\mathbf{X}|\mathbf{Y}} \log \frac{\mathbb{P}(\mathbf{X} \mid \mathbf{Y}')}{\mathbb{P}(\mathbf{X} \mid \mathbf{Y})} \right) \\
&= \mathbb{E}_{\mathbf{Y}} \log \mathbb{E}_{\mathbf{Y}'} \exp \mathbb{E}_{\mathbf{X}|\mathbf{Y}} \log \frac{\mathbb{P}(\mathbf{X} \mid \mathbf{Y})}{\mathbb{P}(\mathbf{X})} + \mathbb{E}_{\mathbf{Y}} \log \mathbb{E}_{\mathbf{Y}'} \exp \left(\mathbb{E}_{\mathbf{X}|\mathbf{Y}} \log \frac{\mathbb{P}(\mathbf{X} \mid \mathbf{Y}')}{\mathbb{P}(\mathbf{X} \mid \mathbf{Y})} \right) \\
&= \mathbb{E}_{\mathbf{Y}} \log \mathbb{E}_{\mathbf{Y}'} \exp \mathbb{E}_{\mathbf{X}|\mathbf{Y}} \log \frac{\mathbb{P}(\mathbf{X} \mid \mathbf{Y}) \mathbb{P}(\mathbf{X} \mid \mathbf{Y}')}{\mathbb{P}(\mathbf{X}) \mathbb{P}(\mathbf{X} \mid \mathbf{Y})} \\
&= \mathbb{E}_{\mathbf{Y}} \log \mathbb{E}_{\mathbf{Y}'} \exp \mathbb{E}_{\mathbf{X}|\mathbf{Y}} \log \frac{\mathbb{P}(\mathbf{X} \mid \mathbf{Y}')}{\mathbb{P}(\mathbf{X})}
\end{aligned}$$

By the concavity of the log function,

$$\begin{aligned}\Delta &\leq \mathbb{E}_{\mathbf{Y}} \log \mathbb{E}_{\mathbf{Y}'} \exp \log \mathbb{E}_{\mathbf{X}|\mathbf{Y}} \frac{\mathbb{P}(\mathbf{X} \mid \mathbf{Y}')}{\mathbb{P}(\mathbf{X})} \\ &= \mathbb{E}_{\mathbf{Y}} \log \mathbb{E}_{\mathbf{X}|\mathbf{Y}} \frac{\mathbb{E}_{\mathbf{Y}'}\mathbb{P}(\mathbf{X} \mid \mathbf{Y}')}{\mathbb{P}(\mathbf{X})} \\ &= \mathbb{E}_{\mathbf{Y}} \log \mathbb{E}_{\mathbf{X}|\mathbf{Y}} \frac{\mathbb{P}(\mathbf{X}')}{\mathbb{P}(\mathbf{X})}\end{aligned}$$

where the $\mathbf{X}'$ distribution is given by $\mathbb{P}(\mathbf{x}') = \mathbb{E}_{\mathbf{Y}'}\mathbb{P}(\mathbf{x} \mid \mathbf{Y}')$. It is important to note that this derivation can be applied for any random vector $\mathbf{Y}'$. The derivations made in Section 3 were made for $\mathbf{Y}'$ following the same distribution as $\mathbf{Y}$. In this case $\mathbb{P}(\mathbf{X}') = \mathbb{P}(\mathbf{X})$ and

$$\Delta \leq \mathbb{E}_{\mathbf{Y}} \log \mathbb{E}_{\mathbf{X}|\mathbf{Y}} \frac{\mathbb{P}(\mathbf{X})}{\mathbb{P}(\mathbf{X})} = 0$$

which proves inequality (5). □

Another choice is to take an i.i.d. vector $\mathbf{Y}'$ having the same marginals as $\mathbf{Y}$. Then

$$\Delta = \mathbb{E}_{\mathbf{Y}} \log \mathbb{E}_{\mathbf{Y}'} \exp \mathbb{E}_{\mathbf{X}|\mathbf{Y}} \log \frac{\mathbb{P}(\mathbf{X} \mid \mathbf{Y}')}{\mathbb{P}(\mathbf{X})}$$

and by Corollary 1,

$$\begin{aligned}\Delta &\leq \mathbb{E}_{\mathbf{Y}} \log \exp \mathbb{E}_{\mathbf{X}|\mathbf{Y}} \log \mathbb{E}_{\mathbf{Y}'} \frac{\mathbb{P}(\mathbf{X} \mid \mathbf{Y}')}{\mathbb{P}(\mathbf{X})} \\ &= \mathbb{E}_{\mathbf{Y}} \mathbb{E}_{\mathbf{X}|\mathbf{Y}} \log \frac{\mathbb{E}_{\mathbf{Y}'}\mathbb{P}(\mathbf{X} \mid \mathbf{Y}')}{\mathbb{P}(\mathbf{X})} \\ &= \mathbb{E}_{\mathbf{X},\mathbf{Y}} \log \frac{\prod_i \mathbb{P}(X_i)}{\mathbb{P}(\mathbf{X})} \times \frac{\mathbb{P}(\mathbf{X} \mid \mathbf{Y})}{\mathbb{P}(\mathbf{X} \mid \mathbf{Y})} \\ &= I(\mathbf{X}; \mathbf{Y}) - qI(X; Y)\end{aligned}$$

which is to be compared to Lemma 3. This proves that if applying our second bound with such an i.i.d. distribution $\mathbf{Y}'$ would lead to a bound that would be worse than the first upper bound (4).

References

[Ari73] Suguru Arimoto. On the converse to the coding theorem for discrete memoryless channels (corresp.). *IEEE Transactions on Information Theory*, 19(3):357–359, May 1973.

[BCO04] Éric Brier, Christophe Clavier, and Francis Olivier. Correlation power analysis with a leakage model. In Marc Joye and Jean-Jacques Quisquater, editors, *Cryptographic Hardware and Embedded Systems - CHES 2004: 6th International Workshop Cambridge, MA, USA, August 11-13, 2004. Proceedings*, volume 3156 of *Lecture Notes in Computer Science*, pages 16–29. Springer, 2004.

[BGH+15] Nicolas Bruneau, Sylvain Guilley, Annelie Heuser, Damien Marion, and Olivier Rioul. Less is More - Dimensionality Reduction from a Theoretical Perspective. In Tim Güneysu and Helena Handschuh, editors, *Cryptographic Hardware and Embedded Systems - CHES 2015 - 17th International Workshop, Saint-Malo, France, September 13-16, 2015, Proceedings*, volume 9293 of *Lecture Notes in Computer Science*, pages 22–41. Springer, 2015.

[BGP+11] Lejla Batina, Benedikt Gierlichs, Emmanuel Prouff, Matthieu Rivain, François-Xavier Standaert, and Nicolas Veyrat-Charvillon. Mutual Information Analysis: a Comprehensive Study. *J. Cryptology*, 24(2):269–291, 2011.

[BR12] Normand J. Beaudry and Renato Renner. An intuitive proof of the data processing inequality. *Quantum Info. Comput.*, 12(5-6):432–441, May 2012.

[BR14] Lejla Batina and Matthew Robshaw, editors. *Cryptographic Hardware and Embedded Systems - CHES 2014 - 16th International Workshop, Busan, South Korea, September 23-26, 2014. Proceedings*, volume 8731 of *Lecture Notes in Computer Science*. Springer, 2014.

[Cla09] Christophe Clavier. DPA Contest 2008–2009, Less than 50 traces allow to recover the key, September 6-9 2009. CHES Special Session 1: DPA Contest. Lausanne, Switzerland, (slides).

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, July 18 2006. ISBN-10: ISBN-10: 0471241954, ISBN-13: 978-0471241959, 2nd edition.

[DFS15] Alexandre Duc, Sebastian Faust, and François-Xavier Standaert. Making masking security proofs concrete - or how to evaluate the security of any leaking device. In Elisabeth Oswald and Marc Fischlin, editors, *Advances in Cryptology - EUROCRYPT 2015 - 34th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Sofia, Bulgaria, April 26-30, 2015, Proceedings, Part I*, volume 9056 of *Lecture Notes in Computer Science*, pages 401–429. Springer, 2015.

[DSV14] François Durvaux, François-Xavier Standaert, and Nicolas Veyrat-Charvillon. How to Certify the Leakage of a Chip? In Phong Q. Nguyen and Elisabeth Oswald, editors, *Advances in Cryptology - EUROCRYPT 2014 - 33rd Annual International Conference on the Theory and Applications of Cryptographic Techniques, Copenhagen, Denmark, May 11-15, 2014. Proceedings*, volume 8441 of *Lecture Notes in Computer Science*, pages 459–476. Springer, 2014.

[FLD12] Yunsi Fei, Qiasi Luo, and A. Adam Ding. A Statistical Model for DPA with Novel Algorithmic Confusion Analysis. In Emmanuel Prouff and Patrick Schaumont, editors, *CHES*, volume 7428 of *Lecture Notes in Computer Science*, pages 233–250. Springer, 2012.

[GBTP08] Benedikt Gierlichs, Lejla Batina, Pim Tuyls, and Bart Preneel. Mutual information analysis. In *CHES, 10th International Workshop*, volume 5154 of *Lecture Notes in Computer Science*, pages 426–442. Springer, August 10-13 2008. Washington, D.C., USA.

[GHR15] Sylvain Guilley, Annelie Heuser, and Olivier Rioul. A Key to Success - Success Exponents for Side-Channel Distinguishers. In Alex Biryukov and Vipul Goyal, editors, *Progress in Cryptology - INDOCRYPT 2015 - 16th International Conference on Cryptology in India, Bangalore, India, December 6-9, 2015, Proceedings*, volume 9462 of *Lecture Notes in Computer Science*, pages 270–290. Springer, 2015.

[GS18] Vincent Grosso and François-Xavier Standaert. Masking Proofs Are Tight and How to Exploit it in Security Evaluations. In Jesper Buus Nielsen and Vincent Rijmen, editors, *Advances in Cryptology - EUROCRYPT 2018 - 37th Annual International Conference on the Theory and Applications of Cryptographic*

Techniques, Tel Aviv, Israel, April 29 - May 3, 2018 Proceedings, Part II, volume 10821 of *Lecture Notes in Computer Science*, pages 385–412. Springer, 2018.

[HRG14] Annelie Heuser, Olivier Rioul, and Sylvain Guilley. Good Is Not Good Enough - Deriving Optimal Distinguishers from Communication Theory. In Batina and Robshaw [BR14], pages 55–74.

[KDOP15] Su Min Kim Kim, Tan Tai Do, Tobias J. Oechtering, and Gunnar Peters. On the Entropy Computation of Large Complex Gaussian Mixture Distributions. *IEEE Transactions on Signal Processing*, 63(17):4710–4723, Sept 2015.

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

[Koc96] Paul C. Kocher. Timing Attacks on Implementations of Diffie-Hellman, RSA, DSS, and Other Systems. In Neal Koblitz, editor, *Advances in Cryptology - CRYPTO '96, 16th Annual International Cryptology Conference, Santa Barbara, California, USA, August 18-22, 1996, Proceedings*, volume 1109 of *Lecture Notes in Computer Science*, pages 104–113. Springer, 1996.

[KSK09] Yuichi Komano, Hideo Shimizu, and Shinichi Kawamura. Built-in determined sub-key correlation power analysis. Cryptology ePrint Archive, Report 2009/161, 2009. http://eprint.iacr.org/2009/161.

[LPR+14] Victor Lomné, Emmanuel Prouff, Matthieu Rivain, Thomas Roche, and Adrian Thillard. How to Estimate the Success Rate of Higher-Order Side-Channel Attacks. In Batina and Robshaw [BR14], pages 35–54.

[Man04] Stefan Mangard. Hardware Countermeasures against DPA – A Statistical Analysis of Their Effectiveness. In *CT-RSA*, volume 2964 of *Lecture Notes in Computer Science*, pages 222–235. Springer, 2004. San Francisco, CA, USA.

[Mas94] James L. Massey. Guessing and entropy. In *Proceedings of 1994 IEEE International Symposium on Information Theory*, pages 204–, Jun 1994.

[NIS01] NIST/ITL/CSD. Advanced Encryption Standard (AES). FIPS PUB 197, Nov 2001. http://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.197.pdf (also ISO/IEC 18033-3:2010).

[PR13] Emmanuel Prouff and Matthieu Rivain. Masking against Side-Channel Attacks: A Formal Security Proof. In Thomas Johansson and Phong Q. Nguyen, editors, *Advances in Cryptology - EUROCRYPT 2013, 32nd Annual International Conference on the Theory and Applications of Cryptographic Techniques, Athens, Greece, May 26-30, 2013. Proceedings*, volume 7881 of *Lecture Notes in Computer Science*, pages 142–159. Springer, 2013.

[Riv08] Matthieu Rivain. On the Exact Success Rate of Side Channel Analysis in the Gaussian Model. In *Selected Areas in Cryptography*, volume 5381 of *LNCS*, pages 165–183. Springer, August 14-15 2008. Sackville, New Brunswick, Canada.

[SA08] François-Xavier Standaert and Cédric Archambeau. Using Subspace-Based Template Attacks to Compare and Combine Power and Electromagnetic Information Leakages. In *CHES*, volume 5154 of *Lecture Notes in Computer Science*, pages 411–425. Springer, August 10–13 2008. Washington, D.C., USA.

[SMY09] François-Xavier Standaert, Tal Malkin, and Moti Yung. A Unified Framework for the Analysis of Side-Channel Key Recovery Attacks. In *EUROCRYPT*, volume 5479 of *LNCS*, pages 443–461. Springer, April 26-30 2009. Cologne, Germany.

[SPAQ06] François-Xavier Standaert, Eric Peeters, Cédric Archambeau, and Jean-Jacques Quisquater. Towards security limits in side-channel attacks. In *CHES*, volume 4249 of *Lecture Notes in Computer Science*, pages 30–45. Springer, October 10-13 2006. Yokohama, Japan.

[TEL09] TELECOM ParisTech SEN research group. DPA Contest, 2008–2009. http://www.DPAcontest.org/.

[TEL10] TELECOM ParisTech SEN research group. DPA Contest (2nd edition), 2009–2010. http://www.DPAcontest.org/v2/.

[vW01] Manfred von Willich. A technique with an information-theoretic basis for protecting secret data from differential power attacks. In Bahram Honary, editor, *Cryptography and Coding, 8th IMA International Conference, Cirencester, UK, December 17-19, 2001, Proceedings*, volume 2260 of *Lecture Notes in Computer Science*, pages 44–62. Springer, 2001.