Success through confidence: Evaluating the effectiveness of a side-channel attack.

Adrian Thillard, Emmanuel Prouff, and Thomas Roche

ANSSI, 51, Bd de la Tour-Maubourg, 75700 Paris 07 SP, France
firstname.name@ssi.gouv.fr

Abstract. Side-channel attacks usually apply a divide-and-conquer strategy, separately recovering different parts of the secret. Their efficiency in practice relies on the adversary ability to precisely assess the success or unsuccess of each of these recoveries. This makes the study of the attack success rate a central problem in side channel analysis. In this paper we tackle this issue in two different settings for the most popular attack, namely the Correlation Power Analysis (CPA). In the first setting, we assume that the targeted subkey is known and we compare the state of the art formulae expressing the success rate as a function of the leakage noise and the algebraic properties of the cryptographic primitive. We also make the link between these formulae and the recent work of Fei *et al.* at CHES 2012. In the second setting, the subkey is no longer assumed to be known and we introduce the notion of *confidence level* in an attack result, allowing for the study of different heuristics. Through experiments, we show that the rank evolution of a subkey hypothesis can be exploited to compute a better confidence than considering only the final result.

1 Introduction

Embedded devices performing cryptographic algorithms may leak information about the processed intermediate values. *Side channel attacks* (SCA) aim to exploit this leakage (usually measures of the power consumption or the electromagnetic emanations) to deduce a secret manipulated by the device.

SCA against block cipher implementations usually consider the secret as a tuple of so-called *subkeys* and apply a *divide-and-conquer* strategy to recover them separately. During the conquering phase, a *partial attack*, limited in time and space, is run against each subkey. Heuristics are then applied to decide on the success or unsuccess of each of these attacks. Subkeys corresponding to attack failures are deduced by exhaustive search. In practice, this last step is often executed either for efficiency reasons or because it is assumed that there is no chance to get the missing subkeys directly by side channel analysis. This description makes apparent that the attack effectiveness greatly depends on the heuristic applied by the adversary. Indeed, incorrect heuristics leave the subsequent exhaustive search little chance to succeed.

Formally, a partial attack is performed on a finite set of measurements $\mathbf{L}$ and aims at the recovery of a correct subkey k_0 among a small set $\mathcal{K}$ of hypotheses (usually, $|\mathcal{K}| = 2^8$ or 2^{16}). For such a purpose, a *score* is computed for every subkey hypothesis $k \in \mathcal{K}$, leading to an ordered *scores vector*. The position r_k of an hypothesis k in this vector is called its *rank*. The attack is said to be *successful* if r_{k_0} equals 1. Extending this notion, an attack is said *o-th order successful* if r_{k_0} is lower than or equal to o.

Under the assumption that the secret k_0 is known, the success of a partial attack can be unambiguously stated. This even allows for the estimation of its *success rate*, by simply dividing the number of attack successes (for which $r_{k_0} \leq o$) by the total number of attacks. If this *known secret assumption* is relaxed, the adversary chooses a candidate which is the most likely according to some *selection rules*. In this case, the success can only be decided *a posteriori* and a *confidence level* must hence be associated *a priori* to the choice before the decision is made. Clearly the soundness of the latter process depends on both the selection and the confidence, which must hence be carefully defined. In particular, to be effective in a practical setting, the confidence associated to a decision must be accurately evaluated even for a small number of observations.

This need is illustrated in Figure 1. An usual selection rule is to simply choose the best ranked key. Using 280 observations, this rule would lead to the choice of the right subkey, whereas a wrong subkey would have been chosen using 420 observations. An optimal heuristic would then deem the first attack a success, and the second one a failure.

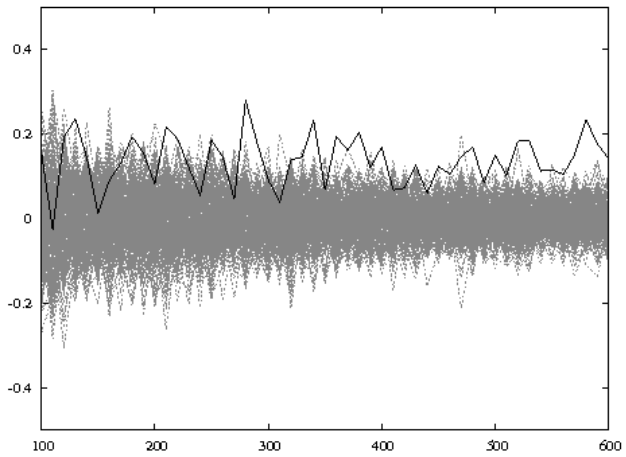


Fig. 1. Correlation coefficients obtained from a CPA on AES. The correct hypothesis is plotted in black.

To evaluate the confidence, we follow a similar approach as in [2] and [9], and we consider the rank of a key and the success rate of an attack as random variables depending on the number of observations. We therefore study the *sampling distribution* of these variables, that is, their distribution when derived from a random sample of finite size.

As an illustration of the sampling distribution of the rank, we run an experiment where several CPA targeting the output of the AES sbox are performed, assuming a Hamming weight leakage model with a Gaussian noise of standard deviation 3. A random subkey k_0 is drawn, and N leakage observations are generated. Then, the rank $r_{k,N}$ of each hypothesis k is computed. This experiment is repeated several times with new leakage observations, and the mean and variance of the associated random variables $R_{k,N}$ are computed. We then perform the same experiment on a leakage of standard deviation 10. The results can be seen in Figure 2.

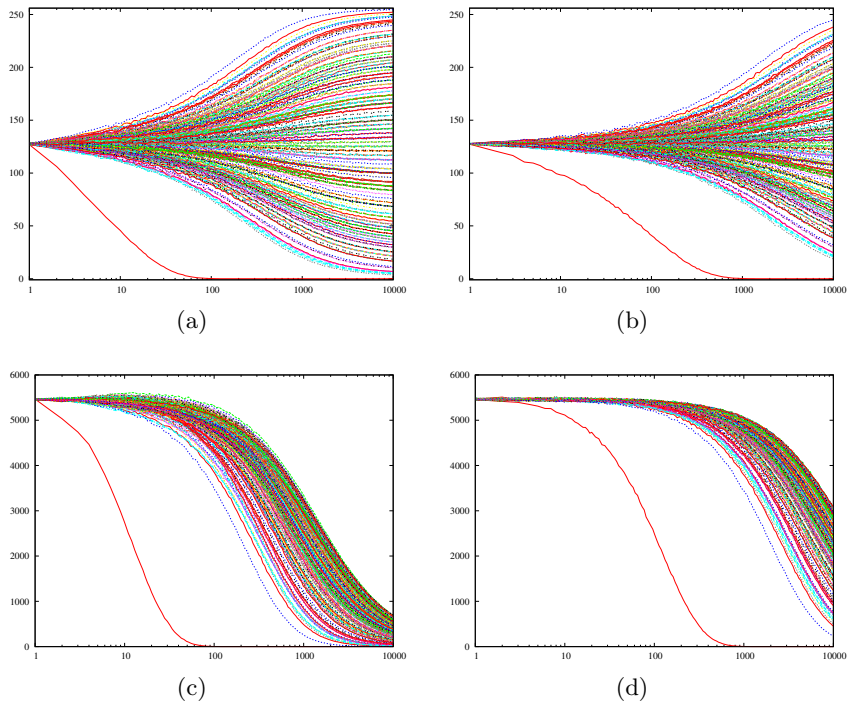


Fig. 2. Results of CPA experiments on the AES sbox. The averages of the ranks are plotted, in function of the number of measurements used for each attack (logscaled), in (a) and (b) for Gaussian noises of standard deviation respectively equal to 3 and 10. Their respective variances are plotted in (c) and (d).

Interestingly, the repetition of this process using a different correct key k'_0 results in the exact same curves, but none of them is associated with the same hypothesis. In fact, the distribution of $R_{k,N}$ does not depend on the value of the hypothesis k, but on its (bit-wise) difference to the correct key k_0. As already mentioned in [9], this can be formally argued by observing that the difference $k \oplus k_0$ can be rewritten as $(k \oplus k_0 \oplus k'_0) \oplus k'_0$. Experiments also show that the rate of convergence is substantially higher for the correct hypothesis, and that the variance of the correct key rank decreases faster than the variance of any wrong key rank. Moreover, the increase of the noise standard deviation only impacts the number of measurements required to observe these patterns.

Figure 2 also hints that the evolution of the sampling distribution of every R_k is eventually related to the value of the correct key and hence brings information about it. In other terms, the full vector of ranks gives information on the correct key (and not only the hypothesis ranked first). Based on this observation, it seems natural to use this information to increase the attack efficiency and/or the confidence in the attack results. To be able to precisely assess both kinds of increase, the distributions of all the variables R_k therefore need to be understood. Bearing this in mind, we now formalize some information that an adversary can obtain while performing a side-channel attack on a set $\mathbf{L}$ of N independent observations. Scores are computed using a *progressive approach*, *i.e.* taking an increasing number of traces into account. Namely, the scores are computed after $N_1 < N$ observations, then again after $N_2 > N_1$ observations, and so on until the N observations in $\mathbf{L}$ have been considered. This approach enables the computation of the matrix:

$$\mathcal{M}_s = \begin{pmatrix} s(1, N_1) & s(1, N_2) & \cdots & s(1, N) \\ \vdots & \vdots & \ddots & \vdots \\ s(|\mathcal{K}|, N_1) & s(|\mathcal{K}|, N_2) & \cdots & s(|\mathcal{K}|, N) \end{pmatrix},$$

where $s(k, N_i)$ denotes the score of the hypothesis k computed using N_i observations.

According to the Neyman-Pearson lemma [8], an optimal selection rule would then require the knowledge of the statistical distribution of this matrix when the correct subkey is known. In a real attack setup however, the latter subkey is unknown and one then has to proceed with a likelihood-ratio approach in order to retrieve it. Even optimal from an effectiveness point of view, this approach is not realistic as it reposes on two major issues: the knowledge of the distribution of the matrix (which requires a theoretical study over highly dimensional data) and the computation and storage of every score (which may require a lot of time and memory). Moreover, one could wonder if all the information contained in the matrix is relevant, or if there is some redundancy. On the opposite side, the actual attacks only use small parts of the available information. For example, the classical selection of the best ranked key simply amounts to choose the maximum of the last column of scores in $\mathcal{M}_s$. Between those two extrem approaches, one could wonder if other tractable parts of the matrix can be used to give better selection rules or better confidence estimators.

Related work The problem of evaluating the success of an attack has already been tackled in several papers [2,6,9,10]. In [6] and [10], the CPA success rate is evaluated by using Fisher's transformation (see for instance [3]): simple formulae are exhibited to estimate the success rate in terms of both the noise standard deviation and the correlation corresponding to the correct key. These works were a first important step towards answering our problem. However, they are conducted under the assumption that wrong hypotheses are uncorrelated to the leakage. As illustrated in Figure 2 (and as already noticed in several papers), this assumption, sometimes called *wrong key randomization hypothesis* [5], does not fit with the reality: each hypothesis score indeed actually depends on the bit-wise difference between the hypothesis and the correct key. The error induced by the assumption is not damaging when one only needs to have an idea about the general attack trends. It is however not acceptable when the purpose is to have a precise understanding of the attack success behavior and of the effect of the sbox properties on it. This observation has been the starting point of the analyses conducted in [2] and [9], where the wrong key randomization hypothesis is relaxed. In Rivain's paper, a new and more accurate success rate evaluation formula is proposed for the CPA. In [2], Fei *et al.* introduce the notion of *confusion coefficient*, and use it to precisely express the success rate of the monobit DPA. This work can be viewed as a specification of Rivain's, as monobit DPA is a particular case of a CPA [1]. This point is formally stated in Section 2.3.

Several criteria indicating the effectiveness of side-channels have also been studied to compare side-channel attacks (e.g. [11]). Among those, the particular behavior of the right subkey ranking have been exploited in [7] to propose an improvement of the attack efficiency when the correct key is unknown. This approach illustrates the importance of such criteria in practical attacks, but it is purely empirical.

Contributions In this paper, we focus on the estimation of the success of an attack in both contexts of known and unknown correct key. In Section 2, state of the art evaluations of the CPA success rate are compared under the Hamming weight leakage model. In Section 3, the impact of the evolution of ranks on the confidence level is studied, and the success rate is used to give a theoretical ground to these results. Finally, conclusions are drawn and new questions are opened in Section 4.

2 CPA success rate

2.1 Notations

Vectors (resp. matrices) with coordinates x_i (resp. x_{ij}) are denoted by $(x_i)_i$ (resp. $(x_{ij})_{i,j}$). Indices bounds are omitted if not needed. For any random variable X, we denote by $\mathbf{E}[X]$ the expectation of X. We denote by $\mathcal{X}$ the set of possible values that can be taken by X. We also denote by $\mathbf{Cov}[X, Y]$ the covariance of X with the random variable Y. When X follows a normal distribution of mean μ

and variance σ^2, we denote it by $X \sim \mathcal{N}(\mu, \sigma^2)$. The set of subkey hypotheses is denoted by $\mathcal{K}$, and $k_0 \in \mathcal{K}$ denotes the correct key, *i.e.* the subkey actually used by the algorithm. We assume that $\mathcal{K}$ is a group for the bit-wise addition and for any $\delta \in \mathcal{K}$, we denote by k_δ the element such that $k_\delta = k_0 \oplus \delta$. Furthermore, we denote by X a (discrete) random variable whose realizations are known to the attacker, by Z_δ the random variable associated to the output of a function f such that $Z_\delta = f(X \oplus k_\delta)$, and by L the random variable associated to the leakage on Z_0. For any i, we denote by x_i and l_i the i-th realization of X and L, and by $z_{\delta,i}$ the i-th realization of Z_δ. For a fixed number N of observations, we denote by ρ_δ the Pearson correlation coefficient between $(l_1, l_2, \cdots, l_N)$ and $(z_{\delta,1}, z_{\delta,2}, \cdots, z_{\delta,N})$. Eventually, we denote the rank of k_δ by R_δ. By definition, it is equal to the number of hypotheses $k_{\delta'}$ such that $\rho_{\delta'} > \rho_\delta$. We will sometimes use the notation $\rho_\delta(N)$ and $R_\delta(N)$ to reveal the functional dependency between ρ_δ (respectively R_δ) and N.

2.2 Theoretical success rate

In this section we aim to compare the theoretical evaluations of the CPA success rate given by [6], [10] and [9]. We recall that, according to the introduced notations, the success rate SR of an attack satisfies:

$$SR = \mathbf{P}(R_0(N) = 1), \tag{1}$$

or equivalently

$$SR = \mathbf{P}(\rho_0(N) - \rho_1(N) > 0, \cdots, \rho_0(N) - \rho_{|\mathcal{K}|-1}(N) > 0). \tag{2}$$

Mangard's study in [6] is conducted in the particular case where $|\mathcal{K}| = 2$ (*i.e.* when there are only two subkey candidates to test). It is moreover based on the three following assumptions:

Assumption 1 [Input uniformity] *The input random variable X is uniformly distributed.*

Assumption 2 [Gaussian distribution of the leakage] *The i-th leakage satisfies $l_i = f(x_i \oplus k_0) + \beta_i$, where β_i is the realization of an independent random variable $B \sim \mathcal{N}(0, \sigma^2)$, and f is a known function.*

Remark 1. Usually, f is of the form $\varphi \circ S$, where φ is surjective and S is a balanced function.

Assumption 3 [Nullity of the wrong hypotheses' correlation coefficients] *The correlation coefficient corresponding to a wrong hypothesis is asymptotically null.*

Using Fisher's Z-transformation, the following approximation of (1) is then obtained:

$$SR \simeq \left(\int_0^\infty \frac{1}{\frac{1}{\sqrt{N-3}}\sqrt{2\pi}} \exp - \frac{(x - \frac{1}{\sqrt{1+\sigma^2}})^2}{\frac{2}{N-3}} dx \right). \tag{3}$$

The latter approximation has been further extended to any subkey set of size $|\mathcal{K}|$ by Standaert *et al.* in [10]:

$$SR \simeq \left(\int_0^{\infty} \frac{1}{\frac{1}{\sqrt{N-3}}\sqrt{2\pi}} \exp - \frac{(x - \frac{1}{\sqrt{1+\sigma^2}})^2}{\frac{2}{N-3}} dx \right)^{|\mathcal{K}|-1}. \quad (4)$$

In subsequent works, Rivain [9] and Fei *et al.* [2] have argued that Assumption 3 is usually not satisfied, which induces an error (possibly high) in (3) and (4) approximations. This observation led Rivain to conduct a new theoretical study of the success rate where the latter assumption is relaxed, and Assumption 1 is replaced by the following one:

Assumption 1 bis [Equality of the inputs occurrences] *Every possible value* $x \in \mathcal{X}$ occurs the same number of times in the sample used for the attack.

Remark 2. This assumption implicitly considers that the study is done by fixing the values taken by X (which is hence no longer a random variable).

Remark 3. When the plaintexts used in the attack are generated uniformly at random and if their number is reasonably high, then the occurrences of every possible value x are very likely to be close to each other.

Under Assumption 1 bis, Rivain has shown that the distribution of the scores vector $(\rho_0(N), \rho_1(N), \cdots, \rho_{|\mathcal{K}|-1}(N))$ produces the same ranking as a new vector $\mathbf{d}(N)$ called the *distinguishing vector* and defined such that $\mathbf{d}(N) = (\Gamma_0(N), \Gamma_1(N), \cdots, \Gamma_{|\mathcal{K}|-1}(N))$, where $\Gamma_\delta(N)$ is the random variable associated to the sum $\frac{1}{N}\sum_{i=1}^{N} z_{\delta,i} l_i$. It is also observed that evaluating the rank $R_\delta(N)$ of a key hypothesis k_δ (at a difference δ of the correct key k_0) amounts to study the number of positive coordinates in the $(|\mathcal{K}|-1)$-dimensional *comparison vector* $\mathbf{c}_\delta(N) = (\Gamma_\delta(N) - \Gamma_0(N), \cdots \Gamma_\delta(N) - \Gamma_{|\mathcal{K}|-1}(N))$ (*i.e.* the vector obtained by subtracting $\mathbf{d}(N)$ to $(\Gamma_\delta(N), \cdots, \Gamma_\delta(N))$, followed by the deletion of the δ-th coordinate). Thanks to this rewriting of the CPA success rate estimation in terms of $\mathbf{d}(N)$ and $\mathbf{c}_\delta(N)$, and considering an independent noise, Rivain proves the two following theorems[1]:

Theorem 1. *[9] In a CPA exploiting N observations leakages, the distinguishing vector $\mathbf{d}(N)$ follows a multivariate normal distribution $\mathcal{N}(\mu_d, \Sigma_d(N))$, such that:*

$$\mu_d = (\kappa_0, \kappa_1, \cdots, \kappa_{|\mathcal{K}|-1}),$$

where $\kappa_\delta = \frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} z_{x,0} z_{x,\delta}$ *and*

$$\Sigma_d(N) = \frac{\sigma^2}{N}(\kappa_{i\oplus j})_{0\le i,j\le|\mathcal{K}|-1}$$

[1] respectively corresponding to Corollary 1 and Section 6 in [9].

Theorem 2. *[9] In a CPA exploiting N observation leakages, the comparison vector $\mathbf{c}_\delta(N)$ follows a multivariate normal distribution $\mathcal{N}(\mu_\delta, \Sigma_\delta(N))$, such that:*

$$\mu_\delta = (\kappa_\delta - \kappa_i)_{i \neq \delta}$$

and

$$\Sigma_\delta(N) = \frac{\sigma^2}{N}(\kappa_0 - \kappa_i - \kappa_j + \kappa_{i \oplus j})_{i,j \neq \delta}.$$

These theorems allow to accurately deduce the distribution of the vectors $\mathbf{d}(N)$ and $\mathbf{c}_\delta(N)$, from the noise variance σ^2 and a modeling of φ. They therefore permit the computation of the probability $\mathbf{P}(R_\delta(N) = 1)$ for any δ (*i.e.* the probability that the hypothesis at difference δ of the correct key is ranked first). According to (1), it may consequently be applied to compute the CPA success rate, which leads to the following success rate evaluation[2]:

$$SR = \Phi_{|\mathcal{K}|-1}(\sqrt{N}\Sigma_0(N)^{-\frac{1}{2}}\mu_0), \tag{5}$$

where $\Phi_{|\mathcal{K}|-1}$ denotes the cdf of the $(|\mathcal{K}|-1)$-dimensional standard normal distribution. In Section 2.3, this new approximation is compared to (4) and it is indeed shown to be more precise.

The coefficient κ_i in Theorems 1 and 2 can be seen as an extension of the definition of the *confusion coefficient* introduced by Fei *et al.* in [2] to estimate the efficiency of a monobit DPA. By analogy with [2], we hence propose the following definition:

Definition 1 (CPA confusion coefficient). *Let k_0 be the correct hypothesis and k_δ be an element of $\mathcal{K}$, for $x \in \mathcal{X}$, let $z_{x,0}$ and $z_{x,\delta}$ be defined such that $z_{x,0} = f(x \oplus k_0)$ and $z_{x,\delta} = f(x \oplus k_\delta)$ for some function f. The* CPA confusion coefficient *κ_δ is then defined by[3]:*

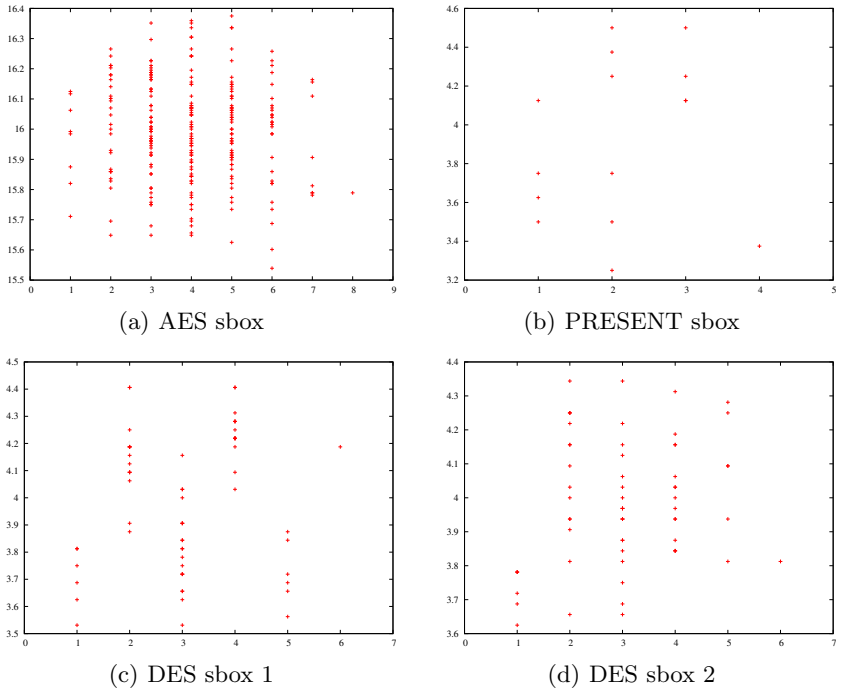
$$\kappa_\delta = \frac{1}{|\mathcal{X}|}\sum_{x \in \mathcal{X}} z_{x,0} z_{x,\delta}.$$

In Figure 3, we illustrate the CPA confusion coefficient in the case where f is the composition of the Hamming weight with some classical sbox. Moreover, Definition 1 implies that, similarly to the expression of the success rate of the DPA proposed in [2], the formula for the CPA success rate can be related to confusion coefficients capturing the impact of the algebraic properties of the cryptographic primitive on the attack efficiency.

In the following section, we compare the formulae of [10] and [9] against experimental simulations of CPA on AES.

[2] This estimation supposes that the covariance matrix $\Sigma_0(N)$ is not singular. When $\Sigma_0(N)$ is singular, other numerical evaluations can be performed (e.g. [4]). In both cases, empirical evaluations of SR can be performed by simulating random vectors $\mathbf{d}(N)$ or $\mathbf{c}_0(N)$ following respectively $\mathcal{N}(\mu_d, \Sigma_d(N))$ or $\mathcal{N}(\mu_0, \Sigma_0(N))$.

[3] Under Assumption 1, when a large enough number of realizations of X are observed, κ_δ is likely to be close to $\mathbf{E}[Z_0 Z_\delta]$.



(a) AES sbox (b) PRESENT sbox

(c) DES sbox 1 (d) DES sbox 2

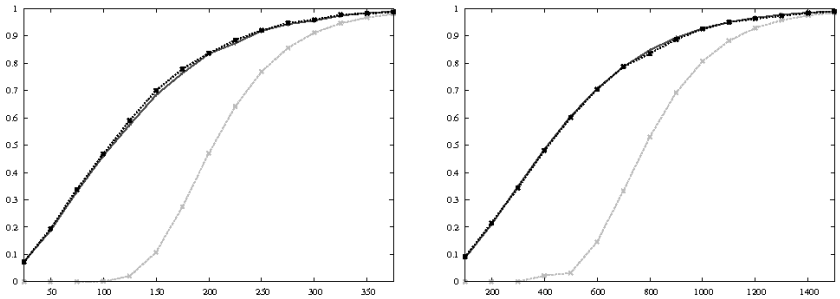
Fig. 3. Values of κ_δ under the assumption that φ is the Hamming weight function, for different sboxes S, in function of the Hamming weight of δ.

2.3 Comparison on AES

In the following, we suppose that the function S is the AES sbox, and that the function φ is the Hamming weight function. First, we estimate the success rate of a CPA empirically, by performing several thousands of attacks. Then, we evaluate Formula (4). Finally, we compute all confusion coefficients, deducing μ_0 and $\Sigma_0(N)$, and we estimate the success rate by evaluating Formula (5). The results are plotted in Figure 4. Formula (5) matches the empirical results quite well. This is mainly due to the relaxing of Assumption 3.

3 Confidence in a result

When performing an attack without the knowledge of the correct subkey k_0, the adversary needs to determine *how* to select the most likely hypothesis, and *when* (*i.e.* after which number of observations). Usually, the *how* problem is answered by using a *selection rule*, such as "choosing the best ranked subkey". To answer the *when* problem, this rule is conditioned by the observation of some pattern, like the stabilization of the rank of the best hypothesis. Figure 5 aims at experimentally validating the latter approach. In the first case, we perform



(a) Noise standard deviation $\sigma = 5$ (b) Noise standard deviation $\sigma = 10$

Fig. 4. Evaluations of the CPA success rate in function of the number of measurements, according to either empirical results (plain black), Formula (4) (dashed light grey) and Formula (5) (dashed dark grey).

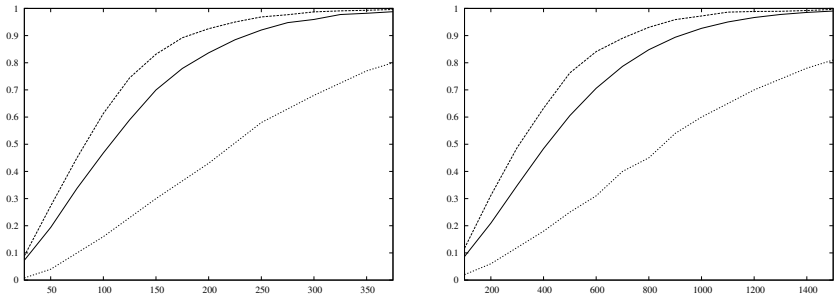
several CPA using an increasing number N of observations and we compute the attack success rate as a function of N. In the second case, we perform the same CPA but we output a candidate subkey only if it has been ranked first both with N and $\frac{N}{2}$ observations. For the latter experiment, we plot the attack success rate considering either the total number of experiments in dotted light grey and considering only the experiments where a key candidate was output (*i.e.* appeared ranked first with N and $\frac{N}{2}$ observations) in dashed light grey.

As it can be seen on Figure 5, the attack based on the stabilization criterion has a better chance (up to 15%) to output a correct result if it outputs anything. However, its overall success rate is significantly lower than the classical CPA success rate. The candidate selection rule hence increases the *confidence* in the selected subkey but decreases the success rate. In fact, we argue here that the two notions are important when studying an attack effectiveness. When attacking several subkeys separately, the assessment of a wrong candidate as a subpart of the whole secret key will lead to an indubitable failure, whereas a subkey that is not found (because the corresponding partial attack does not give a satisfying confidence level) will be bruteforced.

In the following, we give a theoretical justification to this empirical and natural attack effectiveness improvement. To this end, we introduce the notion of *confidence*, which aims at helping the adversary to assess the success or failure of an attack with a known error margin.

3.1 Confidence in an hypothesis

Applying the notations introduced in Section 1, we assume that a partial attack is performed on a set of N independent observations and aims at the recovery of a correct subkey k_0 among a set of hypotheses. For our analysis, the score of each candidate is computed at different steps of the attack (*i.e.* for an increasing number of traces). Namely, the scores are computed after $N_1 < N$ observations,



(a) Noise standard deviation $\sigma = 5$ (b) Noise standard deviation $\sigma = 10$

Fig. 5. Evaluations of the correctness of the output of attacks in function of the number of observations N in different contexts: 1) the best ranked subkey is always returned (plain dark grey, 2)) the best ranked subkey is returned only when it was also ranked first with $\frac{N}{2}$ observations and the success is computed against the number of times both attacks returned the same result (dashed light grey) 3) the best ranked subkey is returned only when it was also ranked first with $\frac{N}{2}$ observations and the success is computed against the number of times the attack has been launched (dotted light grey).

then again after $N_2 > N_1$ observations, and so on until the N observations are considered. In the sequel, the attack on N_i observations is called the i-th attack. All those attacks result in a matrix $\mathcal{M}_s$ containing the scores $s(k, N_i)$ for every hypothesis k and every number N_i of observations. With this construction, the last column vector $(s(k, N))_k$ corresponds to the final attack scores, whereas $(s(k, N_i))_k$ corresponds to intermediate scores (for the i-th attack). In other terms, the right-column of $\mathcal{M}_s$ is the attack result, and the rest of the matrix corresponds to the attack history. With this formalism in hand, the key candidate selection may be viewed as the application of some selection rule $\mathcal{R}$ to $\mathcal{M}_s$, returning a subkey candidate $K^{\mathcal{R}}$. The question raised in the preamble of this section may then be rephrased as: "For some rule $\mathcal{R}$, what is the confidence one can have in $K^{\mathcal{R}}$?". To answer this question, we introduce hereafter the notion of *confidence* in $K^{\mathcal{R}}$.

Definition 2 (Confidence). *For an attack aiming at the recovery of a key k_0 and applying a selection rule $\mathcal{R}$ to output a candidate subkey $K^{\mathcal{R}}$, the* confidence *is defined by:*

$$c(K^{\mathcal{R}}) = \frac{\mathbf{P}(K^{\mathcal{R}} = k_0)}{\sum_{k \in \mathcal{K}} \mathbf{P}(K^{\mathcal{R}} = k)}.$$

Remark 4. The *confidence level* associated to a rule $\mathcal{R}$ merges with the notion of success rate only when the selection rule always outputs a subkey candidate, eg. the rule $\mathcal{R}_0$ defined in the following.

Let us illustrate the application of the confidence level with the comparison of the two following rules, corresponding to the criterion described in the preamble of this section:

- Rule $\mathcal{R}_0$: output the candidate ranked first at the end of the $N-th$ attack.
- Rule $\mathcal{R}_t$: output the candidate ranked first at the end of the $N-th$ attack, only if it was also ranked first for all attacks performed using N_t to N observations.

By definition of $\mathcal{R}_0$, and using the notations of Section 2, the confidence associated to $\mathcal{R}_0$ satisfies:

$$c(K^{\mathcal{R}_0}) = \frac{\mathbf{P}(R_0(N)=1)}{\sum_\delta \mathbf{P}(R_\delta(N)=1)} = \mathbf{P}(R_0(N)=1),$$

which can be computed thanks to Theorem 2.

With a similar reasoning, we have:

$$c(K^{\mathcal{R}_t}) = \frac{\mathbf{P}(R_0(N_t)=1, R_0(N_{t+1})=1, \cdots, R_0(N)=1)}{\sum_\delta \mathbf{P}(R_\delta(N_t)=1, , \cdots, R_\delta(N)=1)},$$

whose evaluation requires more development than that of $c(K^{\mathcal{R}_0})$. For such a purpose, the distribution of the ranks vector $(R_\delta(N_t), R_\delta(N_{t+1}), \cdots, R_\delta(N))$ needs to be studied[4]. We thus follow a similar approach as in Section 2, and we build the *progressive comparison vector* $\mathbf{c}_{\delta,t}(N) = (\mathbf{c}_\delta(N_t)||\mathbf{c}_\delta(N_{t+1})||\cdots||\mathbf{c}_\delta(N))$ where $||$ denotes the vector concatenation operator. We then apply the following proposition, whose proof is given in Annex A:

Proposition 1. *For a CPA exploiting a number N of observations, the progressive comparison vector $\mathbf{c}_{\delta,t}(N)$ follows a multivariate normal distribution $\mathcal{N}(\mu_{\delta,t}, \Sigma_{\delta,t}(N))$, where $\mu_{delta,t}$ is a $|\mathcal{K}|(N-N_t)$ vector and $\Sigma_{\delta,t}$ is a $|\mathcal{K}| \times (N-N_t) \times |\mathcal{K}| \times (N-N_t)$ matrix, satisfying:*

$$\mu_{\delta,t} = (\kappa_\delta - \kappa_0, \cdots, \kappa_\delta - \kappa_{|\mathcal{K}|-1}, \kappa_\delta - \kappa_0, \cdots, \kappa_\delta - \kappa_{|\mathcal{K}|-1}),$$

and

$$\Sigma_{\delta,t}(N) = \left(\frac{N}{max(i,j)}\Sigma_\delta\right)_{N_t \leq i,j \leq N}$$

Proposition 1 allows for the evaluation of the distribution of $\mathbf{c}_{\delta,t}(N)$, and thus for the evaluation of $\mathbf{P}(R_\delta(N_t)=1, R_\delta(N_{t+1})=1, \cdots, R_\delta(N)=1)$ for all hypotheses k_δ. We are then able to compute the confidence $c(K^{\mathcal{R}_t})$.

As an illustration, we study the case where a single intermediate ranking is taken into account, *i.e.* we study the probability $\mathbf{P}(R_\delta(\frac{N}{2})=1, R_\delta(N)=1)$, and we plot in Figure 6 the obtained confidences.

As we can see, the confidence estimation matches the empirical results of Figure 5. At any number of observations, the rule $\mathcal{R}_t$ actually increases the confidence in the output of an attack compared to the rule $\mathcal{R}_0$.

[4] It is worth noting at this point that the variable $R_\delta(N_i)$ does not verify the Markov property, and that the whole vector has to be studied.

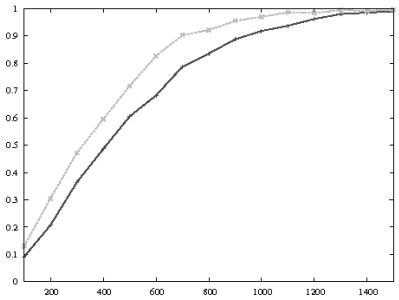


Fig. 6. Evaluation of confidences in function of the number of measurements for $\mathcal{R}_0$ (plain dark grey), and for $\mathcal{R}_{\frac{N}{2}}$ (dashed light grey), with $\sigma = 10$.

3.2 Discussion and empirical study of convergence rules

The accurate evaluation of the confidence level allows a side-channel attacker to assess the success or failure of a partial attack with a known margin of error. For example, and as illustrated in previous section, applying the selection rule $\mathcal{R}_0$ for a CPA on 800 noisy observations (with noise standard deviation equal to 10) leads to an attack failure in 18% of the cases. As a consequence, to reach a 90% confidence level, the attacker has either to perform the attack on more observations (1000 in our example), or to use an other selection rule. Indeed, different selection rules lead to different confidence levels, as they are based on different information. Though a rule based on the whole matrix $\mathcal{M}_s$ would theoretically give the best results, the estimation of the confidence level in such a case would prove to be difficult. An interesting open problem is to find an acceptable tradeoff between the computation of the involved probabilities and the accuracy of the obtained confidence.

In this section, we study a new rule exploiting the *convergence* of the best hypothesis' rank, echoing the observation made in Section 1. To this end, we consider a rule $\mathcal{R}_t^\gamma$ (with $1 \leq \gamma \leq |\mathcal{K}|$) and define it as a slight variation of $\mathcal{R}_t$. The rule $\mathcal{R}_t^\gamma$ returns the best ranked key candidate after the N-th attack only if it was ranked *lower than* γ for the attack on N_t observations. As in previous section, we simulate the simple case where only the ranking obtained with an arbitrary number x of observations is taken into account. We hence experimentally estimate the confidence given by $\mathcal{R}_x^\gamma$ for all γ in Figure 7.

For example, when the final best ranked key is ranked lower than 50 using 200 messages, the confidence is around 94% (compared to 92% when using $\mathcal{R}_0$).

Eventually, the analysis conducted in this section shows that though a stabilization of the rank brings a strong confidence, its convergence can also bring some information to an adversary. This confirms the intuition discussed in Section 1. We propose in Annex B the study of another selection rule commonly considered in the literature.

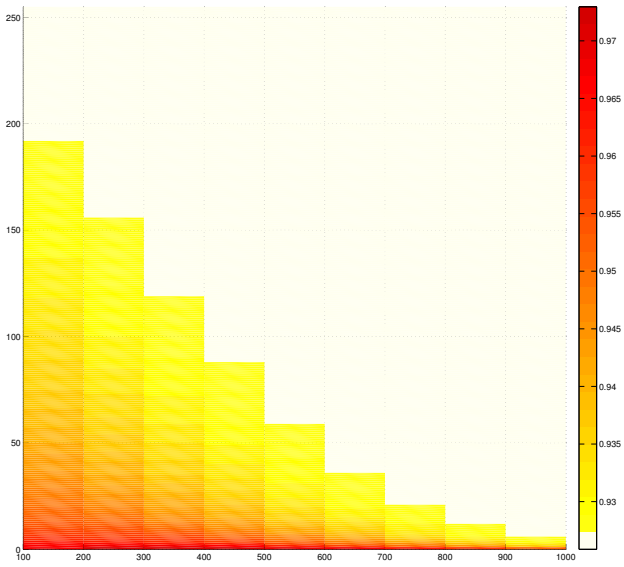


Fig. 7. Confidence in the key ranked first after a CPA on 1000 observations with $\sigma = 10$, knowing that it was ranked below a given rank γ (in y-axis) on a smaller number of measurements N_t (in x-axis).

4 Conclusion

Results presented in this paper are twofold. We first compared several state of the art theoretical evaluations for the success rate of the CPA, and we linked them with the notion of confusion coefficient, capturing the effect of the cryptographic primitive on the difference between the correct hypothesis and the wrong ones. Secondly, we give a rationale for the use of some empirical criteria (such as the convergence of the best hypothesis' rank towards 1) as indicators of the attack success. We hence involve the notion of *confidence* to allow for the accurate estimation of this success.

As an avenue for further research, this work opens the new problem of the exhibition of novel selection rules allowing to efficiently and accurately evaluate the confidence in a side-channel attack while conserving an acceptable success rate.

Acknowledgments We would like to thank Matthieu Rivain and the anonymous reviewers for their fruitful comments.

References

1. J. Doget, E. Prouff, M. Rivain, and F.-X. Standaert. Univariate Side Channel Attacks and Leakage Modeling. *Journal of Cryptographic Engineering*, 1(2):123–144, 2011.
2. Y. Fei, Q. Luo, and A. A. Ding. A Statistical Model for DPA with Novel Algorithmic Confusion Analysis. In E. Prouff and P. Schaumont, editors, *CHES*, volume 7428 of *Lecture Notes in Computer Science*, pages 233–250. Springer, 2012.
3. R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society*, 1922.
4. A. Genz and K. shing Kwong. Numerical evaluation of singular multivariate normal distributions. *Journal of Statistical Computation and Simulation*, 68:1–21, 1999.
5. C. Harpes. Cryptanalysis of iterated block ciphers. In *ETH Series in Information Processing*, volume 7. Hartung-Gorre Verlag, 1996.
6. S. Mangard. Hardware Countermeasures against DPA – A Statistical Analysis of Their Effectiveness. In T. Okamoto, editor, *Topics in Cryptology – CT-RSA 2004*, volume 2964 of *Lecture Notes in Computer Science*, pages 222–235. Springer, 2004.
7. M. Nassar, Y. Souissi, S. Guilley, and J.-L. Danger. "Rank Correction": A New Side-Channel Approach for Secret Key Recovery. In M. Joye, D. Mukhopadhyay, and M. Tunstall, editors, *InfoSecHiComNet*, volume 7011 of *Lecture Notes in Computer Science*, pages 128–143. Springer, 2011.
8. J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933.
9. M. Rivain. On the Exact Success Rate of Side Channel Analysis in the Gaussian Model. In R. Avanzi, L. Keliher, and F. Sica, editors, *Selected Areas in Cryptography*, Lecture Notes in Computer Science, pages 165–183. Springer, 2008.
10. F.-X. Standaert, E. Peeters, G. Rouvroy, and J.-J. Quisquater. An overview of power analysis attacks against field programmable gate arrays. *IEEE*, 94(2):383–394, 2006.
11. C. Whitnall and E. Oswald. A Comprehensive Evaluation of Mutual Information Analysis Using a Fair Evaluation Framework. In P. Rogaway, editor, *CRYPTO*, volume 6841 of *Lecture Notes in Computer Science*, pages 316–334. Springer, 2011.

A Proof of proposition 1

By its construction, the progressive comparison vector $\mathbf{c}_{\delta,t}(N)$ follows a multivariate normal law $\mathcal{N}(\mu_{\delta,t}, \Sigma_{\delta,t}(N))$. Its mean vector $\mu_{\delta,t}$ is trivially deduced from the expression of μ_δ given in Section 2. To compute the expression of $\Sigma_{\delta,t}(N)$, we hence only need to prove the following lemma:

Lemma 1. *For any hypotheses $(i, j, j') \in [0, |\mathcal{K}| - 1]^3$ and for any sets of observations of sizes N_t and N (such that $N_t < N$), Assumptions 2 and 4 imply:*

$$\mathbf{Cov}[\Gamma_i(N) - \Gamma_j(N), \Gamma_i(N_t) - \Gamma_{j'}(N_t)] = \frac{N_t}{N} \mathbf{Cov}[\Gamma_i(N_t) - \Gamma_j(N_t), \Gamma_i(N_t) - \Gamma_{j'}(N_t)].$$

Proof. By the definitions of $\Gamma_i(N)$ *and* $\Gamma_j(N)$*, the following equality holds:* $\Gamma_i(N) - \Gamma_j(N) = \frac{1}{N}(\sum_{t=1}^{N_t} l_t(z_{i,t} - z_{j,t}) + \sum_{t=N_t+1}^{N} l_t(z_{i,t} - z_{j,t}))$*. This can be rewritten as* $\Gamma_i(N) - \Gamma_j(N) = \frac{1}{N}(N_t(\Gamma_i(N_t) - \Gamma_j(N_t)) + \sum_{t=N_t+1}^{N} l_t(z_{i,t} - z_{j,t}))$*. The independence of all observations and the bilinearity of the covariance then suffice to prove the lemma.* □

The coefficients of $\Sigma_{\delta,t}(N)$ can hence be easily computed, using this Lemma.

B Confidence gain with the difference of scores

We study a transverse approach to the one described in Section 3, by observing the last vector of scores (instead of the rank obtained from intermediate attacks). Namely, we focus on a rule outputting the best ranked candidate when the difference between its score and the score of every other hypothesis is greater than a certain value. This criterion is considered for example in [11]. We simulate this rule, for several bounds, and we plot the results in Figure 8. It is of particular

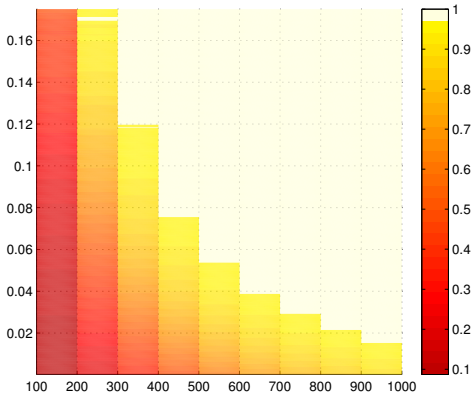


Fig. 8. Confidence in the best ranked key after a CPA with $\sigma = 10$, on a given number of observations (in x-axis), knowing that its score is higher by a certain value (in y-axis) than every other hypothesis score.

interest to note that this rule can bring a huge confidence. Indeed, if the difference using 500 observations is higher than 0.06, then the obtain confidence is around 96% (while 1000 observations would not suffice to attain this level using $\mathcal{R}_0$).