Bivariate attacks and confusion coefficients

Sylvain Guilley and Liran Lerman

March 23, 2017

Abstract

We solve the problem of finding the success rate of an optimal side-channel attack targeting at once the first and the last round of a block cipher. We relate the results to the properties of the direct and inverse substitution boxes (when they are bijective), in terms of confusion coefficients.

1 Introduction

We are interested in the link between the success rate of an attack and the properties of the targetted substitution box (Sbox). For monovariate attacks, the problem is solved. We recall how in the next section 1.1. However, for bivariate attacks, which exploit both the first and the last round of the cipher, the problem is open. We introduce it in section 1.2, and solve it in section 2. Conclusions are in section 3.

1.1 Attack success rate for one Sbox

The link between the success rate and the targetted only holds for the *optimal attack*. Indeed, when the attack is non-optimal, it is possible that the relationship between Sbox properties and the success rate is singular due to the attack singularity. The optimal attack (or maximum likelihood attack) when the noise is Gaussian consists in the Euclidean distance between measurements and model.

The success rate can be modeled using a *first-order exponent* [2, Chap. 11]: for all attack, including the optimal one, there is a constant SE such that[1]:

$$1 - \mathsf{SR} \approx \exp(-q \cdot \mathsf{SE}),$$

where q is the number of traces for the expected success rate to be equal to SR.

[1] Quoting definition 7 of [5], we say that a function $f(x)$ has *first order exponent* $\xi(x)$ if $\big(\ln f(x)\big)/\xi(x) \to 1$ as $x \to +\infty$, in which case we write $f(x) \approx \exp \xi(x)$.

The first-order exponent SE for monovariate attacks takes the following form [5, Proposition 5]:

$$\mathsf{SE}^{D=1} = \min_{k \neq k^\star} \frac{\kappa_{k^\star,k}^2/2}{\kappa''_{k^\star,k} - \kappa_{k^\star,k}^2 + \kappa_{k^\star,k}/\mathrm{SNR}}, \tag{1}$$

where $\kappa_{k^\star,k}$ and $\kappa''_{k^\star,k}$ are two versions of confusion coefficients (which generalize that of [4]), and SNR is the signal-to-noise ratio.

When the SNR is low, then Eqn. (1) simplifies as:

$$\mathsf{SE}^{D=1} \approx \frac{1}{2} \min_{k \neq k^\star} \kappa_{k^\star,k} \cdot \mathrm{SNR}. \tag{2}$$

The value of the success exponent can be interpreted as follows:

Corollary 1. *When the noise is large, the number of traces q to succeed the attack with success rate SR (say $\mathsf{SR} = 90$ %) is inversely proportional to SE, namely:*

$$q = \frac{-\ln(1-\mathsf{SR})}{\mathsf{SE}} = \frac{2.30}{\mathsf{SE}}.$$

This formula allows the following rule of thumb:

- when the SNR is divided by two, then the number of traces to succeed the attack is doubled;
- when in presence of dth-order shuffling, then the number of traces to succeed the attack multiplied by d;
- when first-order masking is applied[2], then the number of traces to succeed the attack is increased from $\frac{2.30}{\mathsf{SE}} = \frac{2.30}{\frac{1}{2}\min_{k \neq k^\star} \kappa_{k^\star,k} \cdot \mathrm{SNR}}$ to $\frac{2.30}{\frac{1}{2}\min_{k \neq k^\star} \kappa_{k^\star,k} \cdot \mathrm{SNR}^2}$.

1.2 Problem of attack success rate for two Sboxes

In side-channel analysis on block ciphers, the attacker shall target a part of the algorithm with two contradictory constraints:

1. it shall be sensitive, i.e., depend on both some controllable data (e.g., plaintext and/or ciphertext) and on a part of the key, and
2. it shall be enumerable, i.e., depend on only a management key portion.

This means that the suitable attack points are shallow in the algorithm. That is, targetted variables shall be close to the plaintext or the ciphertext, where the diffusion is not complete (otherwise, the attacker needs to guess too large portions of the key). Notice that the diffusion of the few last rounds is considered

[2] Notice that the number of traces to succeed the attack decreases only provided the SNR is strictly smaller than one (which is the case in practice).

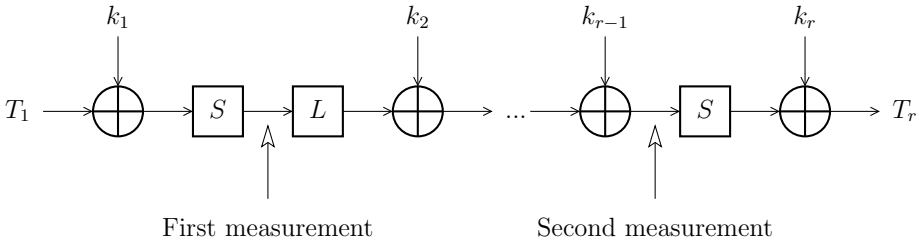

Figure 1: Bi-variate attack setup

with respect to the ciphertext, and not with respect to the plaintext. On the opposite, the attack will gain efficiency if the targetted value is mixed with the key through some function which brings confusion, such as an Sbox. Therefore, in general, the preferred attack points are:

1. after the Sbox in the first round, and
2. before the Sbox in the last round.

This is illustrated on Fig. 1, where the number of rounds is r and there is no diffusion L in the last round (as for the AES). When the attacker knows both the plaintext T_1 and the ciphertext T_r, obviously, he would gain benefit to conduct a bivariate attack. The question is now to quantify the gain of the this bivariate attack, compared to monovariate attacks.

1.3 Notations for the bivariate setting

From now on, we focus on bivariate attacks, targeting the leakage at the first and last rounds, especially in ciphers for which the substitution box is inversible. Examples of such ciphers are AES, PRESENT, Add-Rotate-Xor ciphers (ARX, such as SPECK), etc. Those algorithms have different number of rounds r. In the sequel, as we are interested only on the first and the last rounds, we neglect inner rounds, hence fix the value of r to 2. Thus first round is indicated by index 1, and last round by index 2.

We consider a leakage model:

$$\mathbf{X} = \mathbf{A}\mathbf{Y}^\star + \mathbf{N},$$

where:

- $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \in \mathbb{R}^2$ is the leakage at first and last rounds,
- $\mathbf{A} = \mathrm{diag}(\alpha_1, \alpha_2) = \begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{pmatrix} \in \mathbb{R}^{2\times 2}$ is the amplitude of each leakage,

- $\mathbf{Y}^\star = \begin{pmatrix} Y_1^\star \\ Y_2^\star \end{pmatrix} \in \mathbb{R}^2$ is the centered and normalized model at first and last rounds, for the correct key hypothesis (hence the star),
- $\mathbf{N} = \begin{pmatrix} N_1 \\ N_2 \end{pmatrix}$ is the noise, assumed bivariate normal. Without loss of generality, we also assume that $\mathbf{N}$ is centered (otherwise, a simple pre-processing on $\mathbf{X}$ consisting in centering the traces suffices for the hypothesis to be valid).

Let us consider that T_i and k_i are n-bit words. We denote by w_H the Hamming weight, i.e., the number of bits equal to one in a bit string. For instance, the model can be:

$$Y_i = \frac{2}{\sqrt{n}} \left(w_H(S_i(T_i \oplus k_i)) - \frac{n}{2} \right), \tag{3}$$

where $i = 1$ related to the plaintext and $i = 2$ to the ciphertext, (T_i, k_i) are the text and keys for the considered round (first or last), and $(S_1, S_2) = (S, S^{-1})$ are the *direct* and *inverse* substitution boxes. It can be checked that, provided T_i is uniformly distributed, Y_i is a centred and normal random variable, i.e.,

$$\mathbb{E}(Y_i) = 0 \qquad \text{and} \qquad \mathsf{Var}(Y_i) = \mathbb{E}(Y_i^2) - (\mathbb{E}(Y_i))^2 = 1.$$

Regarding notations, we use Y_i ($i \in \{1, 2\}$) to designate $Y_i(T_i, k_i)$ (recall Eqn. (3)). However, when the context is clear, and in order to keep light notations, we drop the dependence in the text T_i (plaintext when $i = 1$, ciphertext when $i = 2$), and in the key k_i. We make the difference between $Y_i(T_i, k_i)$ and $Y_i(T_i, k_i^\star)$, where k_i is any key and $k_i^\star$ is the correct key, by using respectively Y_i and $Y_i^\star$.

Remark 1 (Regarding the discussion between LL and SG at ULB on 15-16 Sept)**.** *We do not need Y_1 and Y_2 to use the same key. Thus, the conclusions in the sequel will be very general, in particular, not tied to an Even-Mansour scheme (where all keys in the schedule are identical).*

Regarding the noise $\mathbf{N}$, it is centered bivariate normal, hence has a covariance matrix: $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix}$, with $\sigma_1 > 0$, $\sigma_2 > 0$, $-1 < \rho < +1$. This matrix is symmetrical, and inversible (given our assumptions, namely we have neither $\sigma_1 = 0$, nor $\sigma_2 = 0$, nor $\rho = \pm 1$). The inverse is equal to $\Sigma^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} \sigma_1^{-2} & -\sigma_1^{-1}\sigma_2^{-1}\rho \\ -\sigma_1^{-1}\sigma_2^{-1}\rho & \sigma_2^{-2} \end{pmatrix}$.

For the rest of the analysis, we also make a couple of (realistic) assumptions:

Assumption 1 (Independence of plaintext, ciphertext, and noise)**.** *We abstract the block cipher under attack as a PRF (pseudo-random function), hence the plaintext and the ciphertext are independent from each other. In particular, Y_1 and Y_2 are independent. They are also independent with the noise, as customarily assumed in side-channel analysis [7, Sec. 4.1].*

Remark 2. *As will appear in the rest of the developments, the independence between Y_1 and Y_2 imply that the correlation between N_1 and N_2 has an impact on the results only through the value of ρ.*

2 Solution of the success rate for two Sboxes

2.1 Optimal distinguisher for bivariate attacks

Let us assume that the attacker has measured Q traces (each random variable is independent for query q and $q' \neq q$). The measured traces are denoted by lower-case, e.g., $\mathbf{x}_q$ for a side-channel measurement obtained from $\mathbf{X}_q$.

It is shown in [6] that the optimal attack strategy, namely the one which maximizes the success rate SR is the Maximum Likelihood (ML):

$$\hat{\mathbf{k}} = \text{argmax}_{\mathbf{k}}\, p(\mathbf{x}_{1\leq q\leq Q}|\mathbf{y}_{1\leq q\leq Q}(\mathbf{k})), \tag{4}$$

where $\mathbf{k} = \begin{pmatrix} k_1 \\ k_2 \end{pmatrix}$ is the pair of keys to guess.

As already shown in [1], the optimal key guess of Eqn. (4) can also be rewritten in terms of a quadratic form:

Lemma 1 (Optimal distinguisher when the noise is normal [1, Theorem 2])**.** *We have:*

$$\hat{\mathbf{k}} = \text{argmin}_{\mathbf{k}} \sum_{q=1}^{Q} (\mathbf{x}_q - \mathbf{A}\mathbf{y}_q)^{\mathsf{T}} \Sigma^{-1} (\mathbf{x}_q - \mathbf{A}\mathbf{y}_q)\,. \tag{5}$$

Proof. In Eqn. (4), the argument to maximize is:

$$\begin{aligned}
&p(\mathbf{x}_{1\leq q\leq Q}|\mathbf{y}_{1\leq q\leq Q}(\mathbf{k})) \\
&= \prod_{q=1}^{Q} p(\mathbf{X} = \mathbf{x}_q|\mathbf{Y} = \mathbf{y}_q(\mathbf{k})) \\
&= \prod_{q=1}^{Q} p_{\mathbf{N}}(\mathbf{x}_q - \mathbf{A}\mathbf{y}_q(\mathbf{k})) \\
&= \prod_{q=1}^{Q} (2\pi)^{-1} \det \Sigma^{-\frac{1}{2}} \exp -\frac{1}{2}(\mathbf{x}_q - \mathbf{A}\mathbf{y}_q(\mathbf{k}))^{\mathsf{T}} \Sigma^{-1} (\mathbf{x}_q - \mathbf{A}\mathbf{y}_q(\mathbf{k})) \\
&= (2\pi)^{-Q} \det \Sigma^{-\frac{Q}{2}} \exp -\frac{1}{2} \sum_{q=1}^{Q} (\mathbf{x}_q - \mathbf{A}\mathbf{y}_q(\mathbf{k}))^{\mathsf{T}} \Sigma^{-1} (\mathbf{x}_q - \mathbf{A}\mathbf{y}_q(\mathbf{k})).
\end{aligned}$$

This value is maximum when $-\sum_{q=1}^{Q} (\mathbf{x}_q - \mathbf{A}\mathbf{y}_q(\mathbf{k}))^{\mathsf{T}} \Sigma^{-1} (\mathbf{x}_q - \mathbf{A}\mathbf{y}_q(\mathbf{k}))$ is maximum because the exponential function is increasing. □

2.2 Success exponent for bivariate attacks

Since the optimal distinguisher is additive (See Eqn. (5)), the success exponent takes the following value [5, Eqn. (44) of Corollary 1]:

$$\mathsf{SE}^{D=2} = \min_{\mathbf{k}\neq\mathbf{k}^\star} \frac{\frac{1}{2}}{\frac{(\mathbb{E}(\Delta\mathcal{D}))^2}{\mathsf{Var}(\Delta\mathcal{D})} - 1}, \tag{6}$$

where:

- $\mathcal{D}$ is the distinguisher, namely $(\mathbf{X} - \mathbf{AY}(\mathbf{k}))^{\mathsf{T}}\Sigma^{-1}(\mathbf{X} - \mathbf{AY}(\mathbf{k}))$ (as identified from Eqn. (5)[3]),
- $\Delta\mathcal{D} = \mathcal{D}(\mathbf{k}^\star) - \mathcal{D}(\mathbf{k})$.

Let us denote

$$\Delta\mathbf{Y} = \mathbf{Y}(\mathbf{k}^\star) - \mathbf{Y}(\mathbf{k}).$$

As in Eqn. (45) of Definition 8 and Eqn. (63) of Definition 10, we denote the confusion coefficients (for $i \in \{1, 2\}$):

$$\kappa_{i,k_i^\star,k_i} = \mathbb{E}\left(\frac{\Delta Y_i}{2}\right)^2 = \mathbb{E}\left(\frac{Y_i(k_i^\star) - Y_i(k_i)}{2}\right)^2,$$
$$\kappa''_{i,k_i^\star,k_i} = \mathbb{E}\left(\frac{\Delta Y_i}{2}\right)^4 = \mathbb{E}\left(\frac{Y_i(k_i^\star) - Y_i(k_i)}{2}\right)^4.$$

Notice that $\kappa_{i,k_i^\star,k_i}$ and $(\kappa''_{i,k_i^\star,k_i} - \kappa^2_{i,k_i^\star,k_i})$ are respectively the *expectation* and the *variance* of $\frac{\Delta Y_i}{2}$.

As Y_1 and Y_2 are independent, we have:

$$\mathbb{E}\left(\Delta Y_i \Delta Y_j\right) = \begin{cases} 4\kappa_{i,k_i^\star,k_i} & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Theorem 1. *The success exponent for the optimal bivariate attack is:*

$$\mathsf{SE}^{D=2} = \min_{\mathbf{k}\neq\mathbf{k}^\star} \frac{\frac{1}{2}\left(\sum_{i\in\{1,2\}} \frac{\alpha_i^2}{\sigma_i^2}\kappa_{i,k_i^\star,k_i}\right)^2}{\sum_{i\in\{1,2\}} \frac{\alpha_i^4}{\sigma_i^4}\left(\kappa''_{i,k_i^\star,k_i} - \kappa^2_{i,k_i^\star,k_i}\right) + (1+\rho^2)\sum_{i\in\{1,2\}} \frac{\alpha_i^2}{\sigma_i^2}\kappa_{i,k_i^\star,k_i} + 4\rho^2 \prod_{i\in\{1,2\}} \frac{\alpha_i^2}{\sigma_i^2}\kappa_{i,k_i^\star,k_i}}. \tag{8}$$

Proof. We have:

$$\Delta\mathcal{D} = -(\mathbf{A}\Delta\mathbf{Y})^{\mathsf{T}}\Sigma^{-1}(\mathbf{A}\Delta\mathbf{Y}) - 2\mathbf{N}^{\mathsf{T}}\Sigma^{-1}\mathbf{A}\Delta\mathbf{Y}, \tag{9}$$
$$(\Delta\mathcal{D})^2 = (\mathbf{A}\Delta\mathbf{Y})^{\mathsf{T}}\Sigma^{-1}(\mathbf{A}\Delta\mathbf{Y})(\mathbf{A}\Delta\mathbf{Y})^{\mathsf{T}}\Sigma^{-1}(\mathbf{A}\Delta\mathbf{Y}) \tag{10}$$
$$+ 4\mathbf{N}^{\mathsf{T}}\Sigma^{-1}(\mathbf{A}\Delta\mathbf{Y})(\mathbf{A}\Delta\mathbf{Y})^{\mathsf{T}}\Sigma^{-1}\mathbf{N} \tag{11}$$
$$+ 4\mathbf{N}^{\mathsf{T}}\Sigma^{-1}(\mathbf{A}\Delta\mathbf{Y})(\mathbf{A}\Delta\mathbf{Y})^{\mathsf{T}}\Sigma^{-1}(\mathbf{A}\Delta\mathbf{Y}). \tag{12}$$

[3]Indeed, by the law of large numbers, $\frac{1}{Q}(\mathbf{x}_q - \mathbf{A}\mathbf{y}_q(\mathbf{k}))^{\mathsf{T}}\Sigma^{-1}(\mathbf{x}_q - \mathbf{A}\mathbf{y}_q(\mathbf{k})) \longrightarrow \mathbb{E}\left((\mathbf{X} - \mathbf{AY}(\mathbf{k}))^{\mathsf{T}}\Sigma^{-1}(\mathbf{X} - \mathbf{AY}(\mathbf{k}))\right)$ when $Q \longrightarrow +\infty$.

Regarding $\Delta\mathcal{D}$ (Eqn. (9)), one can compute:

$$\begin{aligned}\Delta\mathcal{D} &= -(\mathbf{A}\Delta\mathbf{Y})^{\mathsf{T}}\Sigma^{-1}(\mathbf{A}\Delta\mathbf{Y}) \\ &= \frac{-1}{1-\rho^2}\begin{pmatrix}\alpha_1\Delta Y_1 & \alpha_2\Delta Y_2\end{pmatrix}\begin{pmatrix}\sigma_1^{-2} & -\sigma_1^{-1}\sigma_2^{-1}\rho \\ -\sigma_1^{-1}\sigma_2^{-1}\rho & \sigma_2^{-2}\end{pmatrix}\begin{pmatrix}\alpha_1\Delta Y_1 \\ \alpha_2\Delta Y_2\end{pmatrix} \\ &= \frac{-1}{1-\rho^2}\left(\frac{\alpha_1^2}{\sigma_1^2}\Delta Y_1^2 - 2\frac{\alpha_1\alpha_2}{\sigma_1\sigma_2}\rho\Delta Y_1\Delta Y_2 + \frac{\alpha_2^2}{\sigma_2^2}\Delta Y_2^2\right) .\end{aligned}$$

Hence, by applying Eqn. (7), we have:

$$\mathbb{E}(\Delta\mathcal{D}) = \frac{-4}{1-\rho^2}\left(\frac{\alpha_1^2}{\sigma_1^2}\kappa_{1,k_1^\star,k_1} + \frac{\alpha_2^2}{\sigma_2^2}\kappa_{2,k_2^\star,k_2}\right).$$

Regarding $(\Delta\mathcal{D})^2$, we focus on terms (10) and (11), since the cross-product term (12) has an expectation equal to 0 (since the expression is a multiple of $\mathbf{N}$ which is centered).

The term (10) rewrites as:

$$\begin{aligned}&(\mathbf{A}\Delta\mathbf{Y})^{\mathsf{T}}\Sigma^{-1}(\mathbf{A}\Delta\mathbf{Y})(\mathbf{A}\Delta\mathbf{Y})^{\mathsf{T}}\Sigma^{-1}(\mathbf{A}\Delta\mathbf{Y}) \\ &= \frac{1}{(1-\rho^2)^2}\left(\frac{\alpha_1^2}{\sigma_1^2}\Delta Y_1^2 + \frac{\alpha_2^2}{\sigma_2^2}\Delta Y_2^2 - 2\rho\frac{\alpha_1\alpha_2}{\sigma_1\sigma_2}\Delta Y_1\Delta Y_2\right)^2 .\end{aligned}$$

By taking the expectation, one gets:

$$\begin{aligned}&\mathbb{E}((\mathbf{A}\Delta\mathbf{Y})^{\mathsf{T}}\Sigma^{-1}(\mathbf{A}\Delta\mathbf{Y})(\mathbf{A}\Delta\mathbf{Y})^{\mathsf{T}}\Sigma^{-1}(\mathbf{A}\Delta\mathbf{Y})) \\ &= \frac{16}{(1-\rho^2)^2}\left(\frac{\alpha_1^4}{\sigma_1^4}\kappa''_{1,k_1^\star,k_1} + \frac{\alpha_2^4}{\sigma_2^4}\kappa''_{2,k_2^\star,k_2} + 2\frac{\alpha_1^2\alpha_2^2}{\sigma_1^2\sigma_2^2}\kappa_{1,k_1^\star,k_1}\kappa_{2,k_2^\star,k_2}(1+2\rho^2)\right).\end{aligned}$$

Besides, the term (11) rewrites as:

$$\begin{aligned}&4\mathbf{N}^{\mathsf{T}}\Sigma^{-1}(\mathbf{A}\Delta\mathbf{Y})(\mathbf{A}\Delta\mathbf{Y})^{\mathsf{T}}\Sigma^{-1}(\mathbf{A}\Delta\mathbf{Y}) \\ &= \frac{4}{(1-\rho^2)^2}\left(\frac{N_1^2}{\sigma_1^2}\left(\frac{\alpha_1^2}{\sigma_1^2}\Delta Y_1^2 + \rho^2\frac{\alpha_2^2}{\sigma_2^2}\Delta Y_2^2 - 2\rho\frac{\alpha_1\alpha_2}{\sigma_1\sigma_2}\Delta Y_1\Delta Y_2\right) + \right. \\ &\qquad \frac{N_1N_2}{\sigma_1\sigma_2}\left(-\rho\left(\frac{\alpha_1^2}{\sigma_1^2}\Delta Y_1^2 + \frac{\alpha_2^2}{\sigma_2^2}\Delta Y_2^2\right) + (1+\rho^2)\frac{\alpha_1\alpha_2}{\sigma_1\sigma_2}\Delta Y_1\Delta Y_2\right) + \\ &\qquad \left.\frac{N_2^2}{\sigma_2^2}\left(\rho^2\frac{\alpha_1^2}{\sigma_1^2}\Delta Y_1^2 + \frac{\alpha_2^2}{\sigma_2^2}\Delta Y_2^2 - 2\rho\frac{\alpha_1\alpha_2}{\sigma_1\sigma_2}\Delta Y_1\Delta Y_2\right)\right).\end{aligned}$$

Therefore:

$$\begin{aligned}&\mathbb{E}(4\mathbf{N}^{\mathsf{T}}\Sigma^{-1}(\mathbf{A}\Delta\mathbf{Y})(\mathbf{A}\Delta\mathbf{Y})^{\mathsf{T}}\Sigma^{-1}(\mathbf{A}\Delta\mathbf{Y})) \\ &= \frac{16}{(1-\rho^2)^2}\left(\frac{\alpha_1^2}{\sigma_1^2}\kappa_{1,k_1^\star,k_1} + \frac{\alpha_2^2}{\sigma_2^2}\kappa_{2,k_2^\star,k_2}\right).\end{aligned}$$

Eventually,

$$\frac{\frac{1}{2}(\mathbb{E}(\Delta\mathcal{D}))^2}{\mathbb{E}(\Delta\mathcal{D}^2) - (\mathbb{E}(\Delta\mathcal{D}))^2} = \frac{\frac{1}{2}\left(\sum_{i\in\{1,2\}} \frac{\alpha_i^2}{\sigma_i^2}\kappa_{i,k_i^\star,k_i}\right)^2}{\sum_{i\in\{1,2\}} \frac{\alpha_i^4}{\sigma_i^4}\left(\kappa''_{i,k_i^\star,k_i} - \kappa^2_{i,k_i^\star,k_i}\right) + (1+\rho^2)\sum_{i\in\{1,2\}} \frac{\alpha_i^2}{\sigma_i^2}\kappa_{i,k_i^\star,k_i} + 4\rho^2 \prod_{i\in\{1,2\}} \frac{\alpha_i^2}{\sigma_i^2}\kappa_{i,k_i^\star,k_i}},$$

and $\mathsf{SE}^{D=2}$ is the minimum of this expression over all $\mathbf{k} \neq \mathbf{k}^\star$. □

Corollary 2. *When the noise is large, that is when $SNR_i = \alpha_i^2/\sigma_i^2 \ll 1$ (for $i \in \{1,2\}$), the success exponent for the optimal bivariate attack simplifies to:*

$$\begin{aligned}\mathsf{SE}^{D=2} &\approx \frac{1}{2(1+\rho^2)} \min_{k_1\neq k_1^\star} \frac{\alpha_1^2}{\sigma_1^2}\kappa_{1,k_1^\star,k_1} + \frac{1}{2(1+\rho^2)} \min_{k_2\neq k_2^\star} \frac{\alpha_2^2}{\sigma_2^2}\kappa_{2,k_2^\star,k_2} \\ &= \frac{1}{1+\rho^2}\left(\mathsf{SE}_1^{D=1} + \mathsf{SE}_2^{D=1}\right). \end{aligned} \tag{13}$$

In this expression, the symbol "≈" means an equivalence $\mathcal{O}(SNR_i^2)$ in Bachmann-Landau notation.

Proof. The Taylor expansion of Eqn. (8) when $\mathrm{SNR}_i \longrightarrow 0$ yields:

$$\mathsf{SE}^{D=2} \approx \min_{\mathbf{k}\neq\mathbf{k}^\star} \frac{1}{2(1+\rho^2)}\left(\frac{\alpha_1^2}{\sigma_1^2}\kappa_{1,k_1^\star,k_1} + \frac{\alpha_2^2}{\sigma_2^2}\kappa_{2,k_2^\star,k_2}\right).$$

As there is no cross-coupling term between the leakage at sample 1 and 2, the minimization can be carried out independently over k_1 and k_2. Hence the result, which is indeed equal to $1/(1+\rho^2)$ multiplied by the sum of success exponents for the univariate optimal distinguisher (recall Eqn. (2)) at samples 1 and 2. □

The approximate expression of $\mathsf{SE}^{D=2}$ (Eqn. (13)) depends in the substitution boxes only in the sum of the confusion coefficients at each end of the cipher, weighted by the signal-to-noise ratio at these ends.

Remark 3 (Regaring the study of substitution boxes properties)**.** *The criteria to make substitution boxes resistant against bivariate side-channel attacks which exploit both the first and the last rounds is thus the following: the algorithm is all the more secure as the (weighted) sum of the (minimum value over all keys except the correct one of) confusion coefficients at each end is small.*

That is, the designer of substitution box S (as in [3] for monovariate *attacks) in a* bivariate *case, searches S, a bijection $\mathbb{F}_2^n \to \mathbb{F}_2^n$, along with its inverse S^{-1} such that*

$$\min_{\mathbf{k}\neq\mathbf{k}^\star}\left(\frac{\alpha_1^2}{\sigma_1^2}\kappa_{1,k_1^\star,k_1} + \frac{\alpha_2^2}{\sigma_2^2}\kappa_{2,k_2^\star,k_2}\right)$$

is minimized. Notice that this objective is independent from the correlation coefficient ρ of the noise between the two substitution box calls.

We also see that the expression (13) is maximum when $\rho = 0$, i.e., when the noise at samples 1 and 2 is independent. In this case, we have that the success exponent of the bivariate attack is the sum of the success exponents for the two univariate attacks at each end of the cipher.

The worst case is when $\rho = \pm 1$, in which case the success exponent of the bivariate attack is the average of the success exponents for the two univariate attacks at each end of the cipher.

2.3 Number of traces to recover the key(s)

Using Proposition 1, we can relate these results to the number of traces to extract the keys $k_1^\star$ and $k_2^\star$. In this section 2.3, we assume that the noise is large, e.g., we rely on the simplified formula of Corollary 2.

We first need a simple lemma:

Lemma 2. *Let $a, b > 0$. Then, we have:*

$$\frac{1}{\frac{1}{a} + \frac{1}{b}} \leq \frac{a+b}{4},$$

with equality if and only if $a = b$.

Proof. The inverse function is convex on $]0, +\infty[$, hence $\forall a, b > 0, \forall t \in [0,1]$, it holds that $\frac{1}{t \cdot a + (1-t) \cdot b} \leq t \cdot \frac{1}{a} + (1-t) \cdot \frac{1}{b}$, with equality if and only if $a = b$. Apply the formula with $t = 1/2$. □

Now, let us introduce those notations:

- $q_1^{D=1}$: number of traces to recover the key in a monovariate attack on the first round,
- $q_2^{D=1}$: idem, but on the last round, and
- $q^{D=2}$: number of traces to recover the pair of keys in a bivariate attacks on first and last rounds.

These three quantities depend on the success rate of the attack. For the sake of comparison, we assume they apply to the same success rate SR.

We can now state the important result regarding the attack data complexity:

Theorem 2. *When the keys to recover are the same, and when the noise is large,*

$$q^{D=2} \leq \frac{1+\rho^2}{4} \left(q_1^{D=1} + q_2^{D=1} \right).$$

Moreover, this inequality is tight, as it is an equality when $q_1^{D=1} = q_2^{D=1}$.

Proof. Let $\mathsf{SR} \in]0,1[$ a given success rate. We have:

$$\begin{aligned}
q^{D=2} &= \frac{-\ln(1-\mathsf{SR})}{\mathsf{SE}^{D=2}} && \text{(by corollary 1)} \\
&= \frac{-\ln(1-\mathsf{SR})}{\frac{1}{(1+\rho^2)}(\mathsf{SE}_1^{D=1}+\mathsf{SE}_2^{D=1})} && \text{(by corollary 2)} \\
&= \frac{1+\rho^2}{\frac{1}{q_1^{D=1}}+\frac{1}{q_2^{D=1}}} && \text{(by definition of } q_i^{D=1}\text{, for } i \in \{1,2\}) \\
&\leq \frac{1+\rho^2}{4}\left(q_1^{D=1}+q_2^{D=1}\right) && \text{(by lemma 2),}
\end{aligned}$$

with equality if and only if $q_1^{D=1} = q_2^{D=1}$. □

Remarkably, theorem 2 holds irrespective of the targetted success rate. Moreover, it also holds whatever the intrinsic properties of the substitution box.

We can now interpret theorem 2.

Same keys. When the keys $k_1^\star$ and $k_2^\star$ are dependent, then the success rate is the same in the monovariate and bivariate cases (there is only one key to recover). In the extreme case where the keys to guess are the same (as in Even-Mansour case, where there is no key scheduling), then $q_1^{D=1} + q_2^{D=1}$ is twice the number of traces to guess the key (assuming equal SNR at first and last round). Then, theorem 2 states that a bivariate attack is *always faster* than two monovariate attacks. Numerically, the gain of the bivariate attack over the twain monovariate attacks is:

- none (1 times faster) when ($|\rho| = 1$), and
- 2 times faster when ($|\rho| = 0$).

There is no gain when the two attack points (first and last rounds) convey the same information (the same key is to be guessed) and have the same noise (or opposite), which is sensible: there is thus no *diversity*.

Independent keys. In this case, the demonstration of Theorem 2 would not hold, since the success rate to recover two keys $\mathsf{SR}^{D=2}$ is not comparable to the success rate to recover one single key $\mathsf{SR}_i^{D=1}$, for $i \in \{1,2\}$. In fact, one would have $\mathsf{SR}^{D=2} = \mathsf{SR}_1^{D=1} \times \mathsf{SR}_2^{D=1}$, which makes formal derivations complex.

3 Conclusions

We analyse bivariate side-channel attacks targeting at once plaintext and ciphertext, thereby improving the attack success rate. We establish the probability of success for such attack, and relate it to the direct (first round) and inverse (last round) substitution boxes properties.

References

[1] Nicolas Bruneau, Sylvain Guilley, Annelie Heuser, Damien Marion, and Olivier Rioul. Optimal Side-Channel Attacks for Multivariate Leakages and Multiple Models, August 20 2016. Santa Barbara, CA, USA. Online reference: http://www.proofs-workshop.org/2016/program.html. To appear in the Journal of Cryptographic Engineering.

[2] 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.

[3] Baris Ege, Kostas Papagiannopoulos, Lejla Batina, and Stjepan Picek. Improving DPA resistance of S-boxes: How far can we go? In *2015 IEEE International Symposium on Circuits and Systems, ISCAS 2015, Lisbon, Portugal, May 24-27, 2015*, pages 2013–2016. IEEE, 2015.

[4] 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 *LNCS*, pages 233–250. Springer, 2012.

[5] 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.

[6] Annelie Heuser, Olivier Rioul, and Sylvain Guilley. Good Is Not Good Enough - Deriving Optimal Distinguishers from Communication Theory. In 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*, pages 55–74. Springer, 2014.

[7] Stefan Mangard, Elisabeth Oswald, and Thomas Popp. *Power Analysis Attacks: Revealing the Secrets of Smart Cards*. Springer, December 2006. ISBN 0-387-30857-1, http://www.dpabook.org/.