Good is Not Good Enough
Deriving Optimal Distinguishers from Communication Theory

Annelie Heuser[1⋆], Olivier Rioul[1], and Sylvain Guilley[1,2]

[1] Télécom ParisTech, Institut Mines-Télécom, CNRS LTCI
Department Comelec
{firstname.lastname@telecom-paristech.fr}
[2] Secure-IC S.A.S.

Abstract. We find mathematically optimal side-channel distinguishers by looking at the side-channel as a communication channel. Our methodology can be adapted to any given scenario (device, signal-to-noise ratio, noise distribution, leakage model, etc.). When the model is known and the noise is Gaussian, the optimal distinguisher outperforms CPA and covariance. However, we show that CPA is optimal when the model is only known on a proportional scale. For non-Gaussian noise, we obtain different optimal distinguishers, one for each noise distribution. When the model is imperfectly known, we consider the scenario of a weighted sum of the sensitive variable bits where the weights are unknown and drawn from a normal law. In this case, our optimal distinguisher performs better than the classical linear regression analysis.

Keywords: Side-channel analysis, distinguisher, communication channel, maximum likelihood, correlation power analysis, uniform noise, Laplacian noise.

1 Introduction

Any embedded system that contains secrets, such as a cryptographic key $k^\star$, is prone to side-channel attacks, which proceed in two steps. First, a leakage (power consumption, electromagnetic radiations, time, etc.) is measured, which is a noisy signal dependent on internally manipulated data, some of which are sensitive, meaning that they depend on the secret key $k^\star$ and on some plain-text or cipher-text (denoted by T). A distinguisher is then used to quantify the similarity between the measured leakage and an assumed leakage model. The result is an estimation $\hat{k}$ of the secret key $k^\star$.

In the literature, side-channel distinguishers are customarily presented as statistical operators that confront the leakage and the sensitive variable, both seen as random variables, in order to extract the secret key. Different choices of

⋆ Annelie Heuser is a Google European fellow in the field of privacy and is partially founded by this fellowship.

distinguishers as statistical tools yield different performances, depending on the scenario (device, signal-to-noise ratio, noise distributions, leakage models, etc.)

There are certainly various ways to appreciate the quality of distinguishers. In this article, we focus on distinguishers that maximize the probability of revealing the correct key. In the field of side-channel analysis , somewhat paradoxically, most of the academic works have eluded the precise mathematical derivation of the best distinguisher given a precise attack scenario. Specifically, the community has introduced popular statistical tools (maximum likelihood (ML), difference of means (DoM), covariance, Pearson correlation coefficient (correlation power analysis (CPA)), Kolmogorov-Smirnov distance, etc.) and addressed two questions: *Q1: what distinguishes known distinguishers in terms of distinctive features?*, and *Q2: given a side-channel context what is the best distinguisher among all known ones?*

As for *Q1*, there have been some publications that attempt to highlight specificities of distinguishers. For instance, Doget et al. [4] show that some distinguishers seemingly have different expressions, but are in practice the same one fed with different variants of leakage models. Mangard et al. [11] argue that some distinguishers achieve success performance all the more similar as the noise variance increases; they conclude that only "statistical artifacts" can explain the difference of success probability between a class of selected distinguishers (notably maximum likelihood and correlation). Souissi et al. [20] note that the closer the noise is to a normal distribution (measured by a gaussianity metric), the better the correlation compared to other distinguishers. Besides, it was noticed by Prouff and Rivain [16] that the way a distinguisher is estimated seriously impacts its success rate. This is especially true for information-theoretic side-channel distinguishers, because probability density functions are to be estimated, which is a notoriously difficult problem. In contrast, Whitnall and Oswald [24] defined metrics (such as RDM, the relative distinguishing margin) to rank distinguishers according to exact values, independently of the way they are estimated (notably mutual information). However, the RDM has recently been found questionable in some situations [17]. All in one, it appears difficult to identify salient features that make one distinguisher in particular more appropriate than another.

Regarding question *Q2*, a usual practice is to estimate the success rate using enough simulations or experiments until an unambiguous ranking of the distinguishers can be carried out. In [21], Standaert et al. also consider the quality of the profiling stage when comparing distinguishers. But the fundamental shortcoming of this approach is that the pool of investigated distinguishers is always limited and does not necessarily contain the best possible distinguisher in every scenario.

Contributions. In this paper, we answer the ultimate version of *Q2*, which is also related to *Q1*, namely: *Q3: given a side-channel scenario what is the best distinguisher among all possible ones?* The "*best*" distinguisher is to be understood in terms of *success probability maximization*. Our analyses show that such an objective coincides with the one pursued in digital communication theory [5, 23],

where it is rather formulated as the minimization of the *error probability* (i.e., one minus the success probability). Interestingly, in this approach, it is not necessary to investigate how a distinguisher can be estimated as a stochastic tool, since our analysis already gives the optimal way of estimating the secret key from the measured data.

We show that, when the leakage model is perfectly known by the attacker (on a direct scale [25]), the optimal distinguisher depends only on the noise distribution, not necessarily Gaussian. Consideration of different noise models (Gaussian, uniform, Laplacian) shows that there is no "universal" distinguisher, only one best distinguisher per noise distribution type. Surprisingly, in the additive Gaussian noise case, we find that neither the DoM, nor the CPA are optimal: we exhibit the optimal distinguisher that slightly outperforms them all. The optimal distinguishers for uniform and Laplacian noise are different from Pearson correlation or covariance, and simulations show that they can be much more efficient. When the leakage model is only known on a proportional scale [25] (i.e., $ax + b$ where a and b are unknown) and when the noise is Gaussian, we show that the optimal expression leads exactly to Pearson correlation coefficient. This in particular explains optimality of CPA in this context.

When the model drifts away from Hamming weight (or Hamming distance) and is thus (at least partially) unknown to the attacker, we use a stochastic linear leakage model with unknown coefficients drawn from a normal distribution and derive an optimal distinguisher that outperforms the linear regression attack [4]. Our result has the merit of showing that a rigorous derivation of the optimal attack is possible and that it yields a new expression, which is interpretable in terms of *stochastic* vs. *epistemic* noise[3].

Outline. The remainder of the paper is organized as follows. We express the problem of side-channel analysis (SCA) as a communication problem in Sect. 2. The mathematical derivation of the optimal distinguishers in various scenarios is carried out in Sect. 3 when the leakage model is known. Section 4 derives the optimal distinguisher when the leakage model is partially known to the attacker. Then, Sect. 5 validates the results using simulations. Conclusions and perspectives are in Sect. 6.

2 Side-channel Analysis as a Communication Problem

2.1 Notations

Calligraphic letters (e.g., $\mathcal{X}$) denote sets, capital letters (e.g., X) denote random variables taking values in these sets, and the corresponding lowercase letters (e.g., x) denote their realizations. We write $\mathbb{P}$ for probability distributions, p for densities, and let p_X denote the density of X. Symbols in bold are vectors: $\mathbf{X}$ or $\mathbf{x}$;

[3] In our paper, we use the term *stochastic* for the independent noise N added to the leakage model, and we resort to the term *epistemic* to characterize the distribution of the leakage model when it is not deterministically known.

implicitly, the length of all vectors is m, which is the number of queries (i.e., $\mathbf{X} = (X_i)_{1 \le i \le m}$). We denote the average of $\mathbf{x}$ by $\overline{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m} x_i$, and the scalar product between $\mathbf{x}$ and $\mathbf{y}$ by $\langle \mathbf{x} | \mathbf{y} \rangle = \sum_{i=1}^{m} x_i y_i$. The norms $1, 2, \ldots, q, \ldots, \infty$ are denoted as $\|\mathbf{x}\|_1 = \sum_{i=1}^{m} |x_i|$(*Manhattan norm*), $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{m} x_i^2}$ (*Euclidean norm*), ..., $\|\mathbf{x}\|_q = (\sum_{i=1}^{m} |x_i|^q)^{\frac{1}{q}}$ (*q-norm*) with $q \in \mathbb{R}$, ..., and $\|\mathbf{x}\|_\infty = \max_{i \in [\![1,m]\!]} |x_i|$ (*uniform norm*), respectively. Let k denote any possible key hypothesis from the keyspace $\mathcal{K}$, let $k^\star$ denote the secret secret cryptographic key, and let T be the input or cipher text in the cryptographic algorithm.

2.2 Modeling Through a Communication Channel

In this section, we rewrite the SCA problem as a communication channel problem (Fig. 1). Our setup resembles the one presented by Standaert et al. [22], but focuses specifically on key recovery.

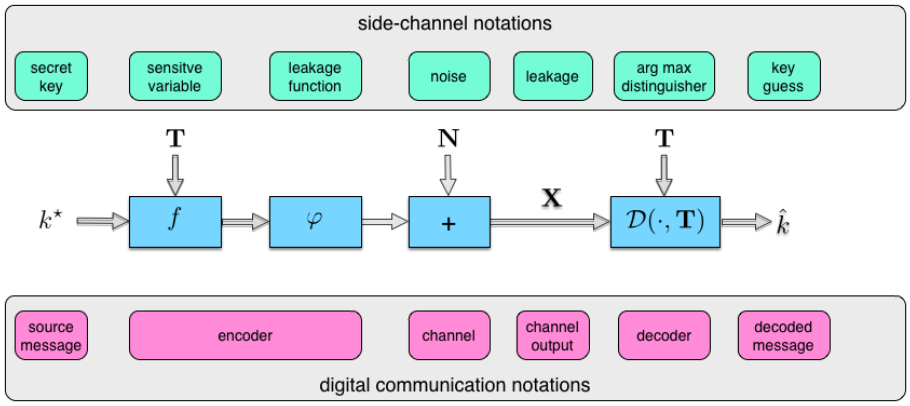


Fig. 1: Side-channel analysis as a communication channel

The *input message* is the secret key $k = k^\star$ (assumed uniformly distributed over $\mathbb{F}_2^n$ in a Bayesian approach). The key is most often recovered piece by piece (independently) using a divide-and-conquer strategy, so n is typically equal to 8 (as in AES, a byte-oriented block cipher). The *encoder* can be any function $\varphi(f(k, \mathbf{T}))$. In SCA, the *sensitive variable* $f(\mathbf{T}, k)$ is normally assumed to be known, since it is part of the algorithm's specification. Depending on the scenario, the *leakage function* $\varphi : \mathbb{F}_2^n \to \mathbb{R}$ can be known (see Sect. 3) or partly unknown (see Sect. 4). Accordingly, $\varphi(f(k, \mathbf{T}))$ can be known or partly unknown. The *communication channel* is the side-channel, typically with *additive* noise $\mathbf{N}$. The decoder to be optimized maximizes the value of the distinguisher by taking its maximal argument over the keyspace[4]. The output of the decoder is then the

[4] Given a function $g(k)$, we use the notation $\arg\max_k \; g(k)$ to denote the value of k that maximizes $g(k)$.

decoded message $\hat{k} = \mathcal{D}(\mathbf{X}, \mathbf{T})$, where $\mathcal{D}$ is the *optimal distinguishing rule*. Notice that we consider the distinguisher as a deterministic mapping, which allows us to rigorously derive optimal expressions. There is an additional *side information*[5] $\mathbf{T}$, which corresponds to the message or the ciphertext, which is assumed to be known both at the encoder and the decoder.

Capturing m measurements means that the channel is used m times. Specifically, the output of the encoder is an independent and identically distributed (i.i.d.) sequence ("codeword") $\varphi(f(k^\star, T_1)), \varphi(f(k^\star, T_2)), \ldots, \varphi(f(k^\star, T_m))$ depending on the i.i.d. sequence of side information $\mathbf{T} = (T_1, T_2, \ldots, T_m)$. The channel is assumed memoryless so that $\mathbf{X} = (X_1, X_2, \ldots, X_m)$ ("received noisy codeword") again forms an i.i.d. sequence; this implies in particular that the additive noise (if present) is white, and successive noise samples $\mathbf{N} = (N_1, N_2, \ldots, N_m)$ are i.i.d.

The problem is to determine the optimum distinguishing (or decoding) rule $\mathcal{D}$ so as to minimize the probability of error

$$\mathbb{P}_e = \mathbb{P}\{\hat{k} \neq k^\star\}, \tag{1}$$

or equivalently to maximize the success probability $\mathbb{P}_s = 1 - \mathbb{P}_e$, which is also referred to as the theoretical or exact success rate [18]).

Theorem 1 (Optimal distinguishing rule). *The optimal distinguishing rule is given by the* maximum a posteriori probability (MAP) *rule*

$$\mathcal{D}(\mathbf{x}, \mathbf{t}) = \arg\max_k \Big(\mathbb{P}\{k\} \cdot p(\mathbf{x}|\mathbf{t}, k) \Big). \tag{2}$$

If the keys are assumed equiprobable, i.e., $\mathbb{P}\{k\} = 2^{-n}$*, Eq.* (2) *reduces to the* maximum likelihood (ML) rule

$$\mathcal{D}(\mathbf{x}, \mathbf{t}) = \arg\max_k p(\mathbf{x}|\mathbf{t}, k). \tag{3}$$

Proof. This is similar to a classical result in communication theory [23, Chap. 2] or [5, Chap. 8], except that one should take the side information into account. The optimal distinguishing rule maximizes

$$\mathbb{P}_s = 1 - \mathbb{P}_e = \mathbb{P}\{\hat{k} = k^\star\} = \mathbb{P}\{k^\star = \mathcal{D}(\mathbf{X}, \mathbf{T})\} \tag{4}$$

$$= \sum_{\mathbf{t}} \mathbb{P}\{\mathbf{t}\} \int p(\mathbf{x}|\mathbf{t}) \cdot \mathbb{P}\{k^\star = \mathcal{D}(\mathbf{x}, \mathbf{t})|\mathbf{x}, \mathbf{t}\} \, \mathrm{d}\mathbf{x}. \tag{5}$$

Since $\mathbb{P}\{\mathbf{t}\} \geq 0$ and $p(\mathbf{x}|\mathbf{t}) \geq 0$, it suffices to maximize the *a posteriori* probability $\mathbb{P}\{k|\mathbf{x}, \mathbf{t}\}$ for every value of $(\mathbf{x}, \mathbf{t})$. Thus the optimal distinguishing rule is $\mathcal{D}(\mathbf{x}, \mathbf{t}) = \arg\max_k \ \mathbb{P}\{k|\mathbf{x}, \mathbf{t}\}$. To evaluate the latter distribution, we apply the Bayes' rule $\mathbb{P}\{k|\mathbf{x}, \mathbf{t}\} = \mathbb{P}\{k\} \cdot p(\mathbf{x}, \mathbf{t}|k)/p(\mathbf{x}, \mathbf{t})$. This gives the MAP optimal distinguishing rule $\mathcal{D}(\mathbf{x}, \mathbf{t}) = \arg\max_k \ \mathbb{P}\{k\} \cdot p(\mathbf{x}, \mathbf{t}|k)$. Furthermore, since $\mathbf{T}$ is obviously key-independent, one can simplify $p(\mathbf{x}, \mathbf{t}|k) = \mathbb{P}\{\mathbf{t}|k\}p(\mathbf{x}|\mathbf{t}, k) = \mathbb{P}\{\mathbf{t}\}p(\mathbf{x}|\mathbf{t}, k)$ so that the MAP and ML rules become as stated. □

[5] This term, not to be confused with the side-channel, is used in communication theory to refer to a variable that is shared unaltered between the encoder and the decoder.

Remark 1. Distinguishing rule in Eq. (2) is useful if there is some a priori knowledge about the distribution of the secret key $k^\star$ (e.g., weak or semi-weak keys in DES [12]).

Remark 2. Provided $p(\mathbf{x}, \mathbf{t}|k)$ is known (for instance through a *profiling* stage), optimal distinguishing rules (2) and (3) can be readily used as an attack. They are known as *template attacks* [2], which are indeed optimal.

3 Optimal Attacks when the Leakage Model is Known

3.1 Derivation

We first consider the scenario of an attacker who knows precisely the leakage model of the device under attack on a "direct scale", in such a way that the *leakage prediction* $Y(k)$ coincides exactly with the deterministic part of the leakage. For example, in an AES software implementation, the device might leak in the Hamming weight (HW) model as $X = \mathsf{HW}[\mathtt{Sbox}[T \oplus k^\star]] + N$, where $\mathtt{Sbox}$ is the SubBytes transformation and $Y(k) = \mathsf{HW}[\mathtt{Sbox}[T \oplus k]]$ for all $k \in \mathcal{K}$.

Proposition 2 (Maximum likelihood). *When f and φ are known to the attacker and $\mathbf{Y}(k) = \varphi(f(k, \mathbf{T}))$, the optimal decision becomes*

$$\mathcal{D}(\mathbf{x}, \mathbf{t}) = \arg\max_k \Big(\mathbb{P}\{k\} \cdot p(\mathbf{x}|\mathbf{y}(k)) \Big). \tag{6}$$

For equiprobable keys this reduces to

$$\mathcal{D}(\mathbf{x}, \mathbf{t}) = \arg\max_k \ p(\mathbf{x}|\mathbf{y}(k)). \tag{7}$$

Proof. Since $(k, \mathbf{T}) \to \mathbf{Y}(k) \to \mathbf{X}$ forms a Markov chain, we have the identity $p(\mathbf{x}|\mathbf{t}, k) = p(\mathbf{x}|\mathbf{t}, k, \mathbf{y}(k)) = p(\mathbf{x}|\mathbf{y}(k))$. Apply Theorem 1. □

Corollary 3. *When the leakage arises from $\mathbf{X} = \mathbf{Y}(k^\star) + \mathbf{N}$,*

$$p(\mathbf{x}|\mathbf{y}(k)) = p_{\mathbf{N}}(\mathbf{x} - \mathbf{y}(k)) = \prod_{i=1}^{m} p_{N_i}(x_i - y_i(k)). \tag{8}$$

This expression, which can be substituted in Eq. (6) *or* (7)*, depends only on the noise probability distribution $p_{\mathbf{N}}$.*

Proof. Trivial, since $\mathbf{N}$ is independent of $\mathbf{Y}(k)$. □

Most publications [2, 13, 18] examine the scenario of Gaussian noise, which we consider next. However, this might not always be valid in practice. Due to other activities on the device, or to some sampling/quantization process for $\mathbf{X}$, or even due to countermeasures, the distribution of the noise might differ from Gaussian. This is addressed in SubSect. 3.3.

3.2 Gaussian Noise Assumption

Theorem 4 (Optimal expression for Gaussian noise). *When the noise is zero mean Gaussian, $N \sim \mathcal{N}(0, \sigma^2)$, the optimal distinguishing rule is*

$$\mathcal{D}_{opt}^{M,G}(\mathbf{x}, \mathbf{t}) = \arg\max_k \ \langle \mathbf{x} | \mathbf{y}(k) \rangle - \frac{1}{2} \|\mathbf{y}(k)\|_2^2. \tag{9}$$

Proof. Applying Corollary 3, a straightforward computation yields

$$\begin{aligned} \arg\max_k p(\mathbf{x}|\mathbf{y}(k)) &= \arg\max_k \ \frac{1}{(\sigma\sqrt{2\pi})^m} e^{-\frac{\|\mathbf{x}-\mathbf{y}(k)\|_2^2}{2\sigma^2}} \\ &= \arg\min_k \ \|\mathbf{x} - \mathbf{y}(k)\|_2^2 \qquad (10) \\ &= \arg\min_k \ \|\mathbf{x}\|_2^2 + \|\mathbf{y}(k)\|_2^2 - 2\langle \mathbf{x} | \mathbf{y}(k) \rangle. \qquad (11) \end{aligned}$$

Since $\|\mathbf{x}\|_2^2$ is not key dependent, we obtain Eq. (9). □

Remark 3. Notice that the optimal distinguisher corresponding to the optimal distinguishing rule of Eq. (9) is $\mathbb{E}\left\{X \cdot Y(k) - \frac{1}{2} Y(k)^2\right\}$, which does not normally reduce to a covariance or correlation coefficient.

Remark 4. The scalar product $\langle \mathbf{x} | \mathbf{y}(k) \rangle$ can be negative, but the optimal expression in Eq. (9) does not involve absolute values. This would only be necessary if the sign of the model was unknown.

Remark 5. In the mono-bit case (i.e., $Y_i(k)$ takes two opposite values), the distinguisher simplifies to $\arg\max_k \langle \mathbf{x} | \mathbf{y}(k) \rangle$. However, somewhat surprisingly, this distinguisher is not the same as the usual DoM from the literature [3, 8] and empirical results show that indeed our optimal distinguishing rule is slightly more efficient. This is detailed in Appendix A.

Remark 6. For a very large number of traces $\frac{1}{2}\|\mathbf{y}(k)\|_2^2$ becomes key independent[6]. However, as we will show in Sect. 5 this factor plays an important role, especially when the signal-to-noise ratio (SNR) is high and thus the number of traces needed to reveal the secret key is low.

We insist that the expression in Eq. (9) is a deterministic value that can be computed from a series of m sampled pairs of leakages and corresponding texts. As the second term $(-\frac{1}{2}\|\mathbf{y}(k)\|_2^2)$ becomes key independent when $m \to \infty$, this expression approximates to $\langle \mathbf{x} | \mathbf{y}(k) \rangle$ or even $\langle \mathbf{x} | \mathbf{y}(k) - \overline{\mathbf{y}}(k) \rangle$ (similar assumption as done in Footnote 6), which is an estimator of the covariance. This is why it can be claimed that when the leakage model is known, the noise is Gaussian

[6] Informally, let us make the hypothesis that T is uniformly distributed in $\mathbb{F}_2^n$ and that $Y(k)$ has the following expression $Y(k) = \varphi(f(T \oplus k))$; then, for large m, we have $\frac{1}{m}\sum_{i=1}^{m} \varphi(f(t_i \oplus k)) \approx \frac{1}{2^n}\sum_{t \in \mathbb{F}_2^n} \varphi(f(t \oplus k)) = \frac{1}{2^n}\sum_{t' \in \mathbb{F}_2^n} \varphi(f(t'))$ which clearly does not depend on k. See also the EIS (Equal Images under the Same key) assumption in [19].

and $m \to \infty$ the optimal distinguisher is very close to the covariance (or to the correlation, since the normalization factor of the Pearson correlation coefficient is also key-independent for large m).

Remark 7. As already mentioned in Remark 2, the best distinguisher when the model is known boils down to a template attack (ML). When the model is known and the noise is Gaussian, it specializes to an equivalent distinguisher which is all the closer to correlation as the SNR is low (by previous Remark 6). This is an independent proof of the main result of [11]. More precisely, the CPA is tolerant to any scaling of the leakage function, provided it is *positive*; otherwise, the attacker must resort to the absolute value of the Pearson correlation coefficient. It is known to be less efficient as depicted in our empirical results in Sect. 5 since there exists more rivals, and the soundness can even be impacted (e.g., if there exists a key $k^c \neq k^\star$ that satisfies $f(k^\star, t) = -f(k^c, t)$ for all $t \in \mathbb{F}_2^n$).

Remark 8. The expression in Eq. (9) can be computed only if the leakage model is known, including its scaling factor (denoted *direct scale* in [25]). In contrast, for CPA the relationship between X and $Y(k)$ is only known up to some affine law (denoted *proportional scale* in [25]) such that $X = aY(k^\star) + b + N$, where a and b are unknown. These coefficients have to be estimated such as to maximize the attacker's performance, i.e., minimize $\|\mathbf{x} - a\mathbf{y}(k) - b\|_2$ in Eq. (10) so as to maximize the likelihood. The following theorem shows that this is equivalent to CPA.

Theorem 5 (Correlation power analysis). *When the leakage arises from $X = aY(k^\star) + b + N$ where N is zero-mean Gaussian, $\hat{k} = \arg\min_k \min_{a,b} \|\mathbf{x} - a\mathbf{y}(k) - b\|^2$, is equivalent to maximizing the absolute value of the empirical Pearson's coefficient:*

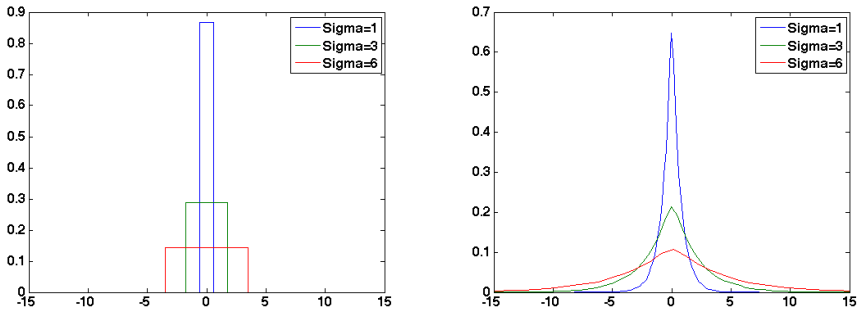
$$\hat{k} = \arg\max_k |\hat{\rho}(k)| = |\widehat{\mathrm{Cov}}(\mathbf{x}, \mathbf{y}(k))| \Big/ \sqrt{\widehat{\mathrm{Var}}(\mathbf{x})\widehat{\mathrm{Var}}(\mathbf{y}(k))} \tag{12}$$

where the empirical (co)variances are defined by $\widehat{\mathrm{Cov}}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})$ and $\widehat{\mathrm{Var}}(\mathbf{x}) = \widehat{\mathrm{Cov}}(\mathbf{x}, \mathbf{x})$.

Proof. The minimization $\min_{a,b} \|\mathbf{x} - a\mathbf{y}(k) - b\|^2$ corresponds to the well-known linear regression analysis (ordinary least squares) [6]. The optimal values of a and b are $a^* = \widehat{\mathrm{Cov}}(\mathbf{x}, \mathbf{y})/\widehat{\mathrm{Var}}(\mathbf{y})$, $b^* = \bar{x} - a^*\bar{y}$, and the minimized mean-squared error takes the well-known expression $\min_{a,b} \|\mathbf{x} - a\mathbf{y} - b\|^2 = \widehat{\mathrm{Var}}(\mathbf{x}) \cdot (1 - \hat{\rho}^2)$ therefore minimizing $\min_{a,b} \|\mathbf{x} - a\mathbf{y} - b\|^2$ amounts to maximizing $|\hat{\rho}|$. □

3.3 Non-Gaussian Noise

The assumption of Gaussian noise may not always hold in practice. We first consider the case of uniform $\mathcal{U}(0, \sigma^2)$ and Laplacian noise distribution $\mathcal{L}(0, \sigma^2)$ as depicted in Fig. 2.



(a) Uniform distribution (b) Laplacian distribution

Fig. 2: Probability distributions for $\sigma \in \{1, 3, 6\}$

Definition 6 (Noise distributions). *Let N be a zero-mean variable with variance σ^2 modeling the noise. Its distribution is:*

- *Uniform, $N \sim \mathcal{U}(0, \sigma^2)$ if* $p_N(n) = \begin{cases} \frac{1}{2\sigma\sqrt{3}} & \text{for } n \in [-\sqrt{3}\sigma, \sqrt{3}\sigma], \\ 0 & \text{otherwise.} \end{cases}$
- *Laplacian, $N \sim \mathcal{L}(0, \sigma^2)$ if* $p_N(n) = \frac{1}{\sqrt{2}\sigma} e^{-\frac{|n|}{\sigma/\sqrt{2}}}$.

For example, uniform noise can arise in side-channel analysis in the case where the only measurement error is the quantization noise. "Oscilloscopes" or most "digital sampling devices" use Analog-to-Digital Converters with only 8 bit resolution. Appendix B shows that Laplacian noise is a good approximation to the noise when combining multiplicatively two (or more) leakage samples.

Theorem 7 (Optimal expression for uniform and Laplacian noises). *When f and φ are known such that $Y(k) = \varphi(f(k, T))$, and the leakage arises from $X = Y(k^\star) + N$ with $N \sim \mathcal{U}(0, \sigma^2)$ or $N \sim \mathcal{L}(0, \sigma^2)$, then the optimal distinguishing rule becomes*

- *Uniform noise distribution:* $\mathcal{D}_{opt}^{M,U}(\mathbf{x}, \mathbf{t}) = \arg\max_k \ -\|\mathbf{x} - \mathbf{y}(k)\|_\infty$,
- *Laplace noise distribution:* $\mathcal{D}_{opt}^{M,L}(\mathbf{x}, \mathbf{t}) = \arg\max_k \ -\|\mathbf{x} - \mathbf{y}(k)\|_1$.

Proof. In case of a uniform noise distribution $\mathcal{U}(0, \sigma^2)$ we have

$$p(\mathbf{x}|\mathbf{y}(k)) = p_N(\mathbf{x} - \mathbf{y}(k)) = \begin{cases} 0 \text{ if } \exists i \mid x_i - y_i(k) \notin [-\sqrt{3}\sigma, \sqrt{3}\sigma], \\ (2\sigma\sqrt{3})^{-m} \text{ otherwise.} \end{cases} \tag{13}$$

Hence, $\arg\max_k \ p_N(\mathbf{x}|\mathbf{y}(k)) = 0$ if and only if $\|\mathbf{x} - \mathbf{y}(k)\|_\infty > \sqrt{3}\sigma$, i.e., $\mathcal{D}_{opt}^{M,U}(\mathbf{x}, \mathbf{t}) = \arg\min_k \ \|\mathbf{x} - \mathbf{y}(k)\|_\infty = \arg\max_k \ -\|\mathbf{x} - \mathbf{y}(k)\|_\infty$.

Assuming a Laplacian noise distribution $\mathcal{L}(0, \sigma^2)$ we have

$$\arg\max_k p(\mathbf{x}|\mathbf{y}(k)) = \arg\max_k \ (\sqrt{2}\sigma)^{-m} \cdot e^{-\frac{\|\mathbf{x} - \mathbf{y}(k)\|_1}{\sigma/\sqrt{2}}}, \tag{14}$$

which reduces to $\arg\max_k \ -\|\mathbf{x}-\mathbf{y}(k)\|_1$. □

We can even be more general. Let $q \in \mathbb{R}$. Consider the *generalized Gaussian* noise distributions [14] of variance σ^2:

$$p(\mathbf{x}|\mathbf{y}(k)) = \left(\frac{q}{2\alpha\Gamma\left(\frac{1}{q}\right)}\right)^m e^{-\left(\frac{\|\mathbf{x}-\mathbf{y}(k)\|_q}{\alpha}\right)^q}, \tag{15}$$

where $\Gamma(\cdot)$ is the *Gamma function* and $\alpha = \sqrt{\frac{\Gamma(1/q)}{\Gamma(3/q)}}\,\sigma$. The optimal distinguishing rule becomes $\mathcal{D}^{M,q}_{opt}(\mathbf{x},\mathbf{t}) = \arg\max_k \ -\|\mathbf{x}-\mathbf{y}(k)\|_q^q = \arg\max_k \ -\|\mathbf{x}-\mathbf{y}(k)\|_q$. The Gaussian, Laplacian and uniform distributions are particular cases obtained for $q = 2, 1, \infty$, respectively.

4 Optimal Attacks when the Leakage Model is Partially Unknown

For standard technologies, the leakage model is either predictable or can be profiled accurately, while being portable from one implementation to another. However, in some contexts, profiling is not possible (the key can neither be chosen nor varied), or changes from one device to the other because of the technological dispersion. Accordingly, the model might not be known exactly to the attacker yielding *epistemic noise*. We now extend our assumptions made in Sect. 3. We assume a linear leakage model as in [4, 19, 25] arising from a weighted sum of the bits of the sensitive variable and additive Gaussian noise N, i.e.,

$$X = \sum_{j=1}^{n} \alpha_j [f(T, k^\star)]_j + N, \tag{16}$$

where $[\cdot]_j : \mathbb{F}_2^n \to \mathbb{F}_2$ is the projection mapping onto the j^{th} bit. But now, the attacker has no knowledge about $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)$ (except that $\boldsymbol{\alpha}$ is distributed according to a given law). This $\boldsymbol{\alpha}$ is unknown but fixed for the whole experiment (series of m measurements). This setting is just one (stochastic) way of considering a leakage model that is not entirely known[7]. See e.g. [1] for a motivation of this scenario, and [4, 24] for assuming and evaluating similar scenarios.

Theorem 8 (Optimal expression for unknown weights). *Let* $\mathbf{Y}_{\boldsymbol{\alpha}}(k) = \sum_{j=1}^{n} \alpha_j [f(\mathbf{T}, k)]_j$ *and* $\mathbf{Y}_j(k) = [f(\mathbf{T}, k)]_j$*, where the weights are independently deviating normally from the Hamming weight model, i.e.,* $\forall j \in [\![1, 8]\!], \alpha_j \sim \mathcal{N}(1, \sigma_\alpha^2)$*. Then the optimal distinguishing rule is*

$$\begin{aligned}\mathcal{D}^{\alpha,G}_{opt}(\mathbf{x},\mathbf{t}) = \arg\max_k \ & (\gamma\langle\mathbf{x}|\mathbf{y}(k)\rangle + \mathbf{1})^t \cdot (\gamma Z(k) + I)^{-1} \cdot (\gamma\langle\mathbf{x}|\mathbf{y}(k)\rangle + \mathbf{1}) \\ & - \sigma_\alpha^2 \ln\det(\gamma Z(k) + I),\end{aligned} \tag{17}$$

[7] For example, diversion of bit loads due to routing, fanout gates, etc. are difficult to model; we used randomly weighted bit sums, randomization being due to technological dispersing (like for PUFS, analog characterization is highly device-dependent due to unpredictable manufacturing defects) and with the idea that the design is balanced (e.g., FPGA, full costume ASIC designs) so that α_j's have equal means.

where $\gamma = \frac{\sigma_\alpha^2}{\sigma^2}$ *is the epistemic to stochastic noise ratio (ESNR),* $\langle \mathbf{x}|\mathbf{y}\rangle$ *is the vector with elements* $(\langle \mathbf{x}|\mathbf{y}(k)\rangle)_j = \langle \mathbf{x}|\mathbf{y}_j(k)\rangle$*,* $Z(k)$ *is the* $n \times n$ *Gram matrix with entries* $Z_{j,j'}(k) = \langle \mathbf{y}_j(k)|\mathbf{y}_{j'}(k)\rangle$*,* $\mathbf{1}$ *is the all-one vector, and* I *is the identity matrix.*

Proof. Again we start from Eq. (7):

$$\mathcal{D}(\mathbf{x},\mathbf{t}) = \arg\max_k \; p(\mathbf{x}|\mathbf{y}_{\boldsymbol{\alpha}}(k)) = \arg\max_k \int_{\mathbb{R}^n} p(\mathbf{x}|\mathbf{y}_{\boldsymbol{\alpha}}(k),\boldsymbol{\alpha})\; p(\boldsymbol{\alpha})\; \mathrm{d}\boldsymbol{\alpha} \tag{18}$$

$$= \arg\max_k \int_{\mathbb{R}^n} \frac{1}{(\sqrt{2\pi}\sigma)^m} e^{-\frac{1}{2\sigma}\|\mathbf{x}-\mathbf{y}_{\boldsymbol{\alpha}}(k)\|_2^2} \frac{1}{(\sqrt{2\pi}\sigma_\alpha)^n} e^{-\frac{1}{2\sigma_\alpha}\|\boldsymbol{\alpha}-\mathbf{1}\|_2^2}\; \mathrm{d}\boldsymbol{\alpha}$$

$$= \arg\max_k \int_{\mathbb{R}^n} \frac{1}{(\sqrt{2\pi}\sigma)^m} \exp\big(-\frac{1}{2\sigma^2}\|\mathbf{x} - \sum_{j=1}^n \alpha_j \mathbf{y}_j(k)\|^2\big) \times$$

$$\frac{1}{(\sqrt{2\pi}\sigma_\alpha)^n} \exp\big(-\frac{1}{2\sigma_\alpha^2}\sum_{j=1}^n (\alpha_j - 1)^2\big) \quad \mathrm{d}\boldsymbol{\alpha}. \tag{19}$$

Now expanding the squares and dropping all multiplicative constants that are independent of k, the distinguishing rule takes the form

$$\arg\max_k \int_{\mathbb{R}^n} \exp\big(-R(\boldsymbol{\alpha})/2\big)\, \mathrm{d}\boldsymbol{\alpha}, \tag{20}$$

$$R(\alpha) = \frac{1}{\sigma^2}(\|\sum_{j=1}^n \alpha_j \mathbf{y}_j\|^2 - 2\sum_{j=1}^n \alpha_j \langle \mathbf{x} \mid \mathbf{y}_j\rangle) + \frac{1}{\sigma_\alpha^2}\sum_{j=1}^n (\alpha_j^2 - 2\alpha_j) \tag{21}$$

$$= \sum_{j,j'=1}^n \alpha_j \alpha_{j'} (\sigma^{-2}\langle \mathbf{y}_j(k) \mid \mathbf{y}_{j'}(k)\rangle + \sigma_\alpha^{-2}\delta_{j,j'}) - 2\sum_{j=1}^n \alpha_j(\sigma^{-2}\langle \mathbf{x}|\mathbf{y}_j(k)\rangle + \sigma_\alpha^{-2})\,.$$

Using an $n \times n$ matrix notation as $\boldsymbol{\alpha}^t Q \boldsymbol{\alpha} = \sum_{j,j'=1}^n \alpha_j \alpha_{j'} Q_{j,j'}$ and $\mathbf{a}^t\boldsymbol{\alpha} = \sum_{j=1}^n a_j\alpha_j$, Eq. (21) takes the form $\boldsymbol{\alpha}^t Q\boldsymbol{\alpha} - 2\mathbf{a}^t\boldsymbol{\alpha}$, where $Q = \sigma^{-2}Z(k) + \sigma_\alpha^{-2} I = \sigma_\alpha^{-2}(\gamma Z(k) + I)$, $\mathbf{a} = \sigma^{-2}\langle \mathbf{x}|\mathbf{y}(k)\rangle + \sigma_\alpha^{-2}\mathbf{1} = \sigma_\alpha^{-2}(\gamma\langle \mathbf{x}|\mathbf{y}(k)\rangle + \mathbf{1})$ and I is the identity matrix, Z is the Gram matrix with entries $Z_{j,j'}(k) = \langle \mathbf{y}_j(k)|\mathbf{y}_{j'}(k)\rangle$, $\mathbf{1}$ is the all-one vector, $\langle \mathbf{x}|\mathbf{y}\rangle$ is the vector with entries $(\langle \mathbf{x}|\mathbf{y}\rangle)_j = \langle \mathbf{x}|\mathbf{y}_j\rangle$. Now, $\boldsymbol{\alpha}^t Q\boldsymbol{\alpha} - 2\mathbf{a}^t\boldsymbol{\alpha} = (\boldsymbol{\alpha} - Q^{-1}\mathbf{a})^t Q(\boldsymbol{\alpha} - Q^{-1}\mathbf{a}) - \mathbf{a}^t Q^{-1}\mathbf{a}$. So,

$$\arg\max_k \int \exp\big(-\tfrac{1}{2}((\boldsymbol{\alpha} - Q^{-1}\mathbf{a})^t Q(\boldsymbol{\alpha} - Q^{-1}\mathbf{a}) - \mathbf{a}^t \cdot Q^{-1} \cdot \mathbf{a})\big)\, \mathrm{d}\boldsymbol{\alpha} \tag{22}$$

$$= \arg\max_k \; (2\pi)^{n/2} |\det Q|^{-1/2} \exp(\tfrac{1}{2}\mathbf{a}^t Q^{-1}\mathbf{a}) \tag{23}$$

$$= \arg\max_k \; \frac{1}{2}\mathbf{a}^t \; Q^{-1}\mathbf{a} - \frac{1}{2}\ln\det Q. \tag{24}$$

Finally, multiplying by $2\sigma_\alpha^2$ we achieve the optimal distinguishing rule. □

Remark 9. For Eq. (17) to work the ESNR γ should be somehow known from some experiments (e.g., Pelgrom coefficients [15] for σ_α and platform noise for σ).

Remark 10. If the ESNR γ is small, i.e., σ_α is small w.r.t. σ, expanding about $\gamma = 0$ and neglecting the term $\sigma_\alpha^2\gamma$ in the expansion of the logarithm gives (at first order in γ):

$$(\mathbf{1} + \gamma\langle \mathbf{x}|\mathbf{y}(k)\rangle)^t (I + \gamma Z(k))^{-1} (\mathbf{1} + \gamma\langle \mathbf{x}|\mathbf{y}(k)\rangle) \tag{25}$$

$$\approx n + 2\gamma\ \mathbf{1}^t\langle \mathbf{x}|\mathbf{y}(k)\rangle - \gamma \mathbf{1}^t Z(k) \cdot \mathbf{1}^t. \tag{26}$$

Since $\mathbf{1}^t\mathbf{y}(k) = \sum_{j=1}^n y_j(k) = \mathsf{HW}[\mathbf{y}]$ and

$$\mathbf{1}^t Z(k) \mathbf{1}^t = \sum_{j,j'=1}^{n} \langle y_j(k)|y_{j'}(k)\rangle = \langle \sum_{j=1}^{n} y_j(k)| \sum_{j'=1}^{n} y_{j'}(k)\rangle = \|\mathsf{HW}[\mathbf{y}]\|_2^2, \tag{27}$$

Eq. (26) boils down to maximizing $\langle \mathbf{x}|\mathsf{HW}[\mathbf{y}]\rangle - \frac{1}{2}\|\mathsf{HW}[\mathbf{y}]\|_2^2$. As expected, we recover the optimal distinguishing rule when the Hamming weight model is assumed to be known and $\alpha_j \approx 1$ (see SubSect. 3.2).

Remark 11. If ESNR γ is large (σ_α is large w.r.t. σ), a similar calculation as done in Remark 10 shows that the optimal distinguishing rule becomes

$$\gamma\langle \mathbf{x}|\mathbf{y}(k)\rangle^t \cdot Z^{-1}(k) \cdot \langle \mathbf{x}|\mathbf{y}(k)\rangle - \sigma_\alpha^2 \ln\det(Z(k)), \tag{28}$$

where $\det(Z(k)) = \|\mathbf{y}_1(k) \wedge \cdots \wedge \mathbf{y}_n(k)\|_2^2$ is the Gram determinant, the squared norm of the exterior product of the $\mathbf{y}_j(k)$'s . This simpler formula can be useful to be directly implemented for small stochastic noise.

Remark 12. Note that, in contrast to the linear regression attack (LRA) [4], $\mathcal{D}^{\alpha,G}$ does not require an estimation of $\boldsymbol{\alpha}$ explicitly; $\mathcal{D}^{\alpha,G}$ is already optimal given the *a priori* probability distribution of $\boldsymbol{\alpha}$. An empirical comparison is shown in Subsec 5.2.

5 Experimental Validation

5.1 Known Model: *Stochastic* Noise

As an application we choose $Y = \mathsf{HW}[\mathtt{Sbox}[T \oplus k]]$ and $X = Y(k^\star) + N$, where $\mathtt{Sbox} : \mathbb{F}_2^8 \to \mathbb{F}_2^8$ is the AES Substitution box and T is uniformly distributed over $\mathbb{F}_2^8$. We simulated noise from several distributions p_N and for $\sigma \in \{1, 3, 6\}$ resulting in an SNR of $\frac{Var(Y)}{Var(N)} = \frac{2}{\sigma^2} \in \{2, 0.222, 0.056\}$. Note that since the SNR is equivalent for all noise distributions, we can compare the performance of the distinguishers across different noise distributions. For reliability, we conducted 500 independent experiments in each setting with uniformly distributed $k^\star$ to compute the empirical success rate (noted $\hat{\mathbb{P}}_s$). Moreover, as suggested in [10], when plotting the empirical success rate, we highlight the standard deviation of the success rate by error bars. In particular, since $\hat{\mathbb{P}}_s$ follows a binomial distribution, we shaded the confidence interval $\left[\hat{\mathbb{P}}_s \pm \sqrt{\frac{\hat{\mathbb{P}}_s(1-\hat{\mathbb{P}}_s)}{n_{\text{exp}}}}\right]$, where $n_{\text{exp}} = 500$ is the

number of experiments. If the error bars do not overlap, we can unambiguously conclude that one distinguisher is better than the other [10].

In the scenario where the model is known, we implemented the following distinguishers, where the labels for the figures are put within parentheses:

$$\mathcal{D}_{opt}^{M,G}(\mathbf{x},\mathbf{t}) = \arg\max_k \langle \mathbf{x}|\mathbf{y}(k)\rangle - \frac{1}{2}\|\mathbf{y}(k)\|_2^2, \quad \text{(Euclidean norm)} \tag{29}$$

$$\mathcal{D}_{opt\text{-}s}^{M,G}(\mathbf{x},\mathbf{t}) = \arg\max_k \langle \mathbf{x}|\mathbf{y}(k)\rangle, \quad \text{(Scalar product)} \tag{30}$$

$$\mathcal{D}_{opt}^{M,L}(\mathbf{x},\mathbf{t}) = \arg\max_k -\|\mathbf{x}-\mathbf{y}(k)\|_1, \quad \text{(Manhattan norm)} \tag{31}$$

$$\mathcal{D}_{opt}^{M,U}(\mathbf{x},\mathbf{t}) = \arg\max_k -\|\mathbf{x}-\mathbf{y}(k)\|_\infty, \quad \text{(Uniform norm)} \tag{32}$$

$$\mathcal{D}_{Cov}(\mathbf{x},\mathbf{t}) = \arg\max_k |\langle \mathbf{x}-\overline{\mathbf{x}}|\mathbf{y}(k)\rangle|, \quad \text{(Covariance)} \tag{33}$$

$$\mathcal{D}_{CPA}(\mathbf{x},\mathbf{t}) = \arg\max_k \left| \frac{\langle \mathbf{x}-\overline{\mathbf{x}}|\mathbf{y}(k)\rangle}{\|\mathbf{x}-\overline{\mathbf{x}}\|_2 \cdot \|\mathbf{y}(k)-\overline{\mathbf{y}(k)}\|_2} \right|. \quad \text{(CPA)} \tag{34}$$

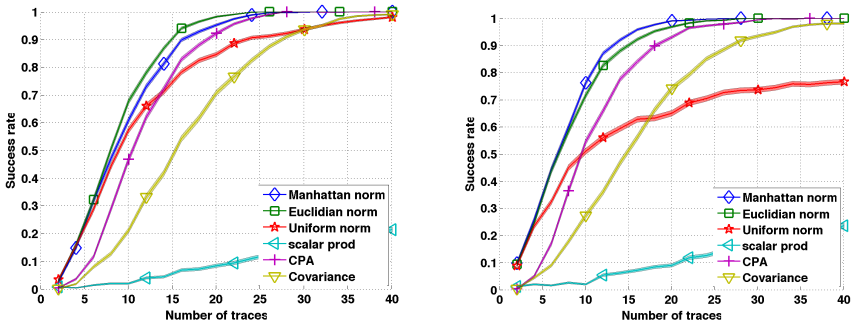
Figures 3a, 3c and 3e show empirical success rate curves for Gaussian noise. One can see that for all levels of SNR $\mathcal{D}_{opt}^{M,G}$ outperforms the other distinguishers, including CPA. As expected from Remark 6, scalar product, covariance, and correlation have poorer but comparable performance than $\mathcal{D}_{opt}^{M,G}$ for high noise.

Figures 3b, 3d and 3f show the empirical success rate curves for Laplacian noise. For low noise, $\mathcal{D}_{opt}^{M,L}$ is the most efficient and $\mathcal{D}_{opt}^{M,G}$ is the nearest rival, whereas $\mathcal{D}_{CPA}$ and $\mathcal{D}_{Cov}$ are less efficient. As the noise increases the difference becomes more significant. As expected, $\mathcal{D}_{CPA}$ and $\mathcal{D}_{Cov}$ become equivalent for high noise, and $\mathcal{D}_{opt}^{M,U}$ fails to distinguish.

In case of uniform noise (see Fig. 4) all optimal distinguishers behave similarly for $\sigma = 1$, whereas CPA, covariance and the scalar product are less efficient. When the noise increases, $\mathcal{D}_{opt}^{M,U}$ is the most efficient distinguisher. One can see that $\mathcal{D}_{opt}^{M,U}$ for uniform noise and $\mathcal{D}_{opt}^{M,L}$ for Laplacian noise require less traces to succeed than $\mathcal{D}_{opt}^{M,G}$ does for Gaussian noise. More precisely, for $\sigma = 6$, $\mathcal{D}_{opt}^{M,U}$ requires only 28 traces to reach $\hat{\mathbb{P}}_s \geq 90\%$, $\mathcal{D}_{opt}^{M,L}$ requires 200 traces, whereas $\mathcal{D}_{opt}^{M,G}$ in case of Gaussian noise needs 300 measurements. This is in keeping with the known information-theoretic fact that detection (or decoding) in Gaussian noise is harder than in any other type of noise.

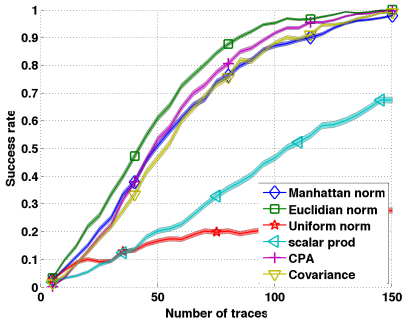
5.2 Unknown Model: *Epistemic* and *Stochastic* Noise

To account for a partially unknown model, we choose $Y_j = [\texttt{Sbox}[T \oplus k]]_j$ for $j = 1, \ldots, 8$ and $X = \sum_{j=1}^{8} \alpha_j Y_j(k^\star) + N$, where $\alpha_j \sim \mathcal{N}(1, \sigma_\alpha)$ are unknown and changing for each experiment. Note that in this scenario $\mathbf{Y}(k)$ is a *column* and not a value as in the previous subsection. Figure 5 shows typical values for $\sigma_\alpha \in \{2, 4\}$, showing that the assumption about $\boldsymbol{\alpha}$ is realistic (see e.g., [7]). We compare our new optimal distinguisher with the linear regression analysis (LRA) [4], which is a *non-profiling* variant of the stochastic approach [19] and

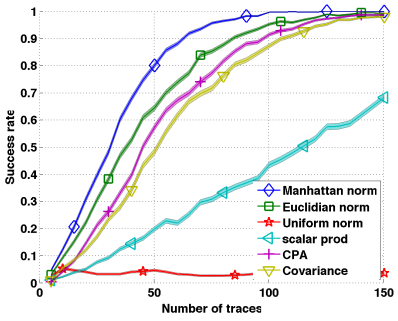


(a) Gaussian Noise, $\sigma = 1$

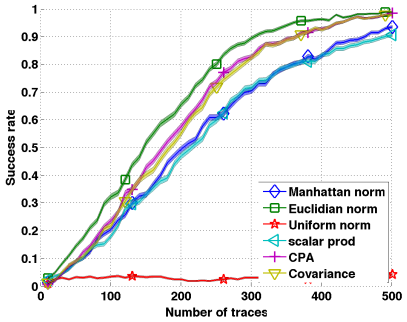
(b) Laplacian Noise, $\sigma = 1$



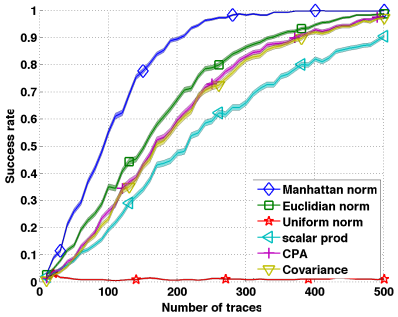
(c) Gaussian Noise, $\sigma = 3$



(d) Laplacian Noise, $\sigma = 3$



(e) Gaussian Noise, $\sigma = 6$

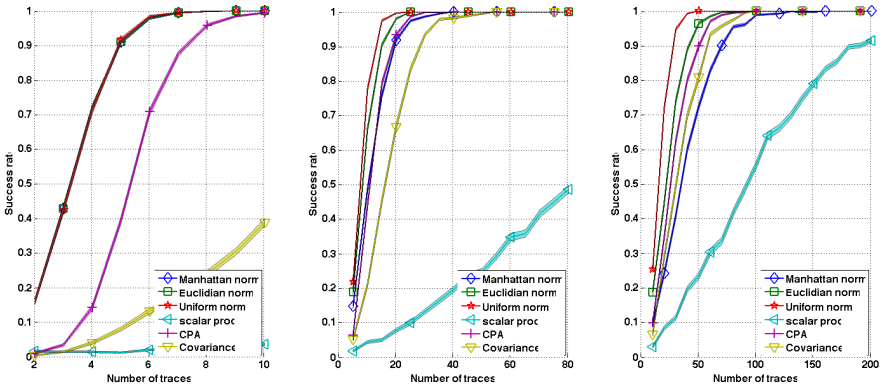


(f) Laplacian Noise, $\sigma = 6$

Fig. 3: Success rate various σ, with a known model

the most efficient attack so far in the case where the model drifts away from the Hamming weight model [4, 9]. LRA is defined as

$$\mathcal{D}_{LRA}(\mathbf{x}, \mathbf{t}) = \arg\min_{k} \frac{\|\mathbf{x} - \mathbf{y}'(k) \cdot \boldsymbol{\beta}(k)\|_2^2}{\|\mathbf{x} - \overline{\mathbf{x}}\|_2^2}, \tag{35}$$



(a) $\sigma = 1$ (b) $\sigma = 3$ (c) $\sigma = 6$

Fig. 4: Success rate for a uniform noise distribution, with a known model

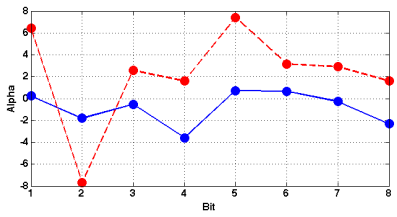
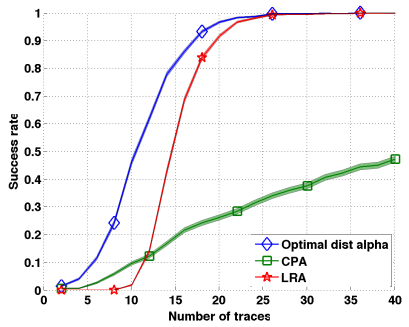


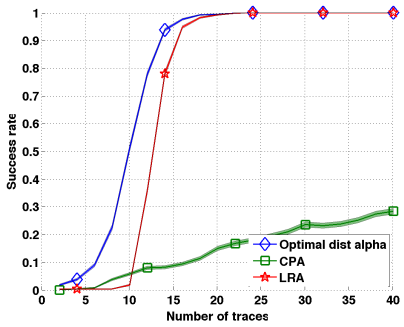
Fig. 5: Exemplary values of $\boldsymbol{\alpha}$ for $\sigma_\alpha = 2$ (blue) and $\sigma_\alpha = 4$ (red dashed)

where $\mathbf{y}'(k) = (\mathbf{1}, \mathbf{y}_1(k), \mathbf{y}_2(k), \ldots, \mathbf{y}_8(k))$ is an $m \times 9$ matrix and $\boldsymbol{\beta}(k) = (\beta_1(k), \ldots, \beta_9(k))$ are the regression coefficients $\boldsymbol{\beta}(k) = (\mathbf{y}'(k)^t \cdot \mathbf{y}'(k))^{-1}\mathbf{y}'(k)^t\mathbf{x}$. Criterion (35) is also known as the *coefficient of determination* [6]. We compared the optimal distinguisher to LRA and CPA, for which we used $Y = \mathsf{HW}[\mathtt{Sbox}[T \oplus k]]$. Apart from this we used the same experimental setup as above.

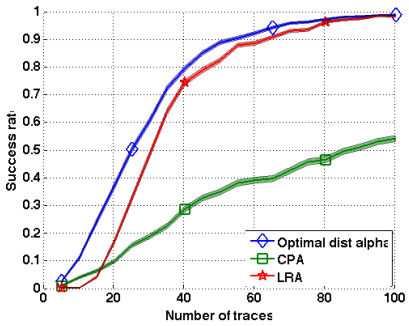
Figure 6 displays the success rate for $\sigma \in \{1, 3, 6\}$ and $\sigma_\alpha \in \{2, 4\}$. As expected CPA is performing worse than both other attacks. Remarkably, in all scenarios $\mathcal{D}_{opt}^{\alpha,G}$ (labeled Optimal dist alpha) is more efficient than LRA. This is perhaps not surprising as regression analysis involves mean squared minimization rather than direct success probability maximization as $\mathcal{D}_{opt}^{\alpha,G}$ does. As already observed in [4], LRA needs a large enough number of traces for estimation, that is why $\hat{\mathbb{P}}_s$ stays low until around 10 traces (Fig. 6a and 6b). One can observe that both distinguishers perform better for $\sigma_\alpha = 4$ (Figures 6b, 6d and 6f) than for $\sigma_\alpha = 2$ (Figures 6a, 6c and 6e). This can be explained by the improved distinguishability through the *distinct influence of each bit*. On the contrary, $\mathcal{D}_{CPA}$ becomes worse when σ_α increases, because the model drifts father away from the Hamming weight model.



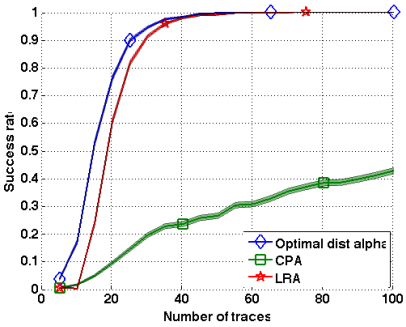
(a) $\sigma_\alpha = 2, \sigma = 1$



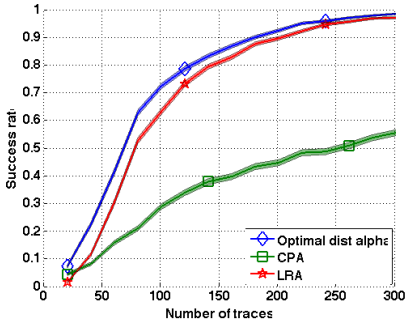
(b) $\sigma_\alpha = 4, \sigma = 1$



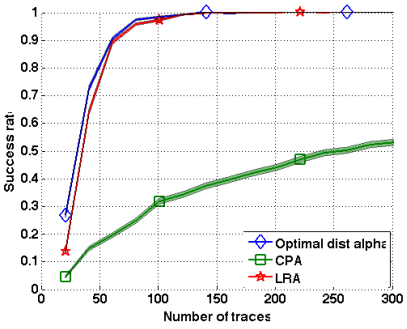
(c) $\sigma_\alpha = 2, \sigma = 3$



(d) $\sigma_\alpha = 4, \sigma = 3$



(e) $\sigma_\alpha = 2, \sigma = 6$



(f) $\sigma_\alpha = 4, \sigma = 6$

Fig. 6: Success rate for various ESNRs, unknown model

6 Conclusion

We examined the *key extraction* problem in *a side-channel context* as a *digital communication* problem. Following the reasoning used in digital communication theory, we derived the *optimal distinguisher* (called *optimal decoder* in digital communication theory). It is a formula that takes as input a multiplicity of pairs of side-channel leakage measurements and corresponding text chunks, and that returns the key guess that maximizes the success probability. The methodical derivation of distinguishers yields an estimator that can be directly computed from the measured data.

In the case where no information is known about the channel (Sect. 2.2), we recovered the template attack. When the leakage function is known (Sect. 3), the approach yields a different distinguisher for each noise distribution. For the classical case of additive Gaussian noise, the optimal distinguisher cannot be interpreted as a covariance nor as a correlation, albeit very close for low SNR. In addition, when the leakage model is known only on a proportional scale we recover CPA exactly. When the noise is non-Gaussian, the optimal distinguishers are very different from CPA or correlation and each optimal distinguisher is the most efficient in its scenario. When the leakage model is partially unknown (Sect. 4) and modeled as an unevenly weighted sum of bits with unknown weights, our method outperforms the non-profiled version of the stochastic approach (LRA).

This study suggests that a mathematical study of distinguishers should prevail in the field of side-channel analysis. As a perspective, our optimal distinguishers may be tested on real measurements. This should include a preliminary step to determine the underlying scenario as precisely and efficiently as possible in terms of the number of traces. Especially, the determination of the noise distribution is a notoriously difficult problem. Moreover, the extension of our work to higher-order attacks (when the noise distribution might differ from Gaussian) seems promising.

References

1. Mehdi-Laurent Akkar, Régis Bevan, Paul Dischamp, and Didier Moyart. Power analysis, what is now possible... In *ASIACRYPT*, volume 1976 of *Lecture Notes in Computer Science*, pages 489–502. Springer, 2000.
2. Suresh Chari, Josyula R. Rao, and Pankaj Rohatgi. Template Attacks. In *CHES*, volume 2523 of *LNCS*, pages 13–28. Springer, August 2002. San Francisco Bay (Redwood City), USA.
3. Jean-Sébastien Coron, Paul C. Kocher, and David Naccache. Statistics and Secret Leakage. In *Financial Cryptography*, volume 1962 of *Lecture Notes in Computer Science*, pages 157–173. Springer, February 20-24 2000. Anguilla, British West Indies.
4. Julien Doget, Emmanuel Prouff, Matthieu Rivain, and François-Xavier Standaert. Univariate side channel attacks and leakage modeling. *J. Cryptographic Engineering*, 1(2):123–144, 2011.
5. Robert G. Gallager. *Information theory and reliable communication*. Wiley, 1968.

6. O.J.W.F. Kardaun. *Classical Methods of Statistics*. Springer, 2005.
7. Michael Kasper, Werner Schindler, and Marc Stöttinger. A stochastic method for security evaluation of cryptographic FPGA implementations. In Jinian Bian, Qiang Zhou, Peter Athanas, Yajun Ha, and Kang Zhao, editors, *FPT*, pages 146–153. IEEE, 2010.
8. Paul C. Kocher, Joshua Jaffe, and Benjamin Jun. Differential Power Analysis. In *Proceedings of CRYPTO'99*, volume 1666 of *LNCS*, pages 388–397. Springer-Verlag, 1999.
9. Victor Lomné, Emmanuel Prouff, and Thomas Roche. Behind the scene of side channel attacks. In Kazue Sako and Palash Sarkar, editors, *ASIACRYPT*, volume 8269 of *Lecture Notes in Computer Science*, pages 506–525. Springer, 2013.
10. Houssem Maghrebi, Olivier Rioul, Sylvain Guilley, and Jean-Luc Danger. Comparison between Side Channel Analysis Distinguishers. In Tat Wing Chim and Tsz Hon Yuen, editors, *ICICS*, volume 7618 of *LNCS*, pages 331–340. Springer, October 29-31 2012. Hong Kong.
11. Stefan Mangard, Elisabeth Oswald, and François-Xavier Standaert. One for All - All for One: Unifying Standard DPA Attacks. *Information Security, IET*, 5(2):100–111, 2011.
12. Judy H. Moore and Gustavus J. Simmons. Cycle Structures of the DES with Weak and Semi-Weak Keys. In Andrew M. Odlyzko, editor, *CRYPTO*, volume 263 of *Lecture Notes in Computer Science*, pages 9–32. Springer, 1986.
13. Amir Moradi, Nima Mousavi, Christof Paar, and Mahmoud Salmasizadeh. A Comparative Study of Mutual Information Analysis under a Gaussian Assumption. In *WISA (Information Security Applications, 10th International Workshop)*, volume 5932 of *Lecture Notes in Computer Science*, pages 193–205. Springer, August 25-27 2009. Busan, Korea.
14. Saralees Nadarajah. A generalized normal distribution. *Journal of Applied Statistics*, 32(7):685–694, 2005.
15. Marcel J.M. Pelgrom, Aad C.J. Duinmaijer, and Anton P.G. Welbers. Matching properties of MOS transistors. *IEEE Journal of Solid State Circuits*, 24(5):1433–1439, 1989.
16. Emmanuel Prouff and Matthieu Rivain. Theoretical and practical aspects of mutual information-based side channel analysis. *International Journal of Applied Cryptography (IJACT)*, 2(2):121–138, 2010.
17. Oscar Reparaz, Benedikt Gierlichs, and Ingrid Verbauwhede. A note on the use of margins to compare distinguishers. In *COSADE (*to appear*)*, Lecture Notes in Computer Science. Springer, April 14-15 2014. Paris, France.
18. 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.
19. Werner Schindler, Kerstin Lemke, and Christof Paar. A Stochastic Model for Differential Side Channel Cryptanalysis. In LNCS, editor, *CHES*, volume 3659 of *LNCS*, pages 30–46. Springer, Sept 2005. Edinburgh, Scotland, UK.
20. Youssef Souissi, Nicolas Debande, Sami Mekki, Sylvain Guilley, Ali Maalaoui, and Jean-Luc Danger. On the Optimality of Correlation Power Attack on Embedded Cryptographic Systems. In Ioannis G. Askoxylakis, Henrich Christopher Pöhls, and Joachim Posegga, editors, *WISTP*, volume 7322 of *Lecture Notes in Computer Science*, pages 169–178. Springer, June 20-22 2012.
21. François-Xavier Standaert, François Koeune, and Werner Schindler. How to Compare Profiled Side-Channel Attacks? In Springer, editor, *ACNS*, volume 5536 of *LNCS*, pages 485–498, June 2-5 2009. Paris-Rocquencourt, France.

22. 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.
23. Andrew J. Viterbi and Jim K. Omura. *Principles of digital communication and coding*. McGraw-Hill series in electrical engineering, 2007.
24. Carolyn Whitnall and Elisabeth Oswald. A Fair Evaluation Framework for Comparing Side-Channel Distinguishers. *J. Cryptographic Engineering*, 1(2):145–160, 2011.
25. Carolyn Whitnall, Elisabeth Oswald, and François-Xavier Standaert. The Myth of Generic DPA...and the Magic of Learning. In Josh Benaloh, editor, *CT-RSA*, volume 8366 of *Lecture Notes in Computer Science*, pages 183–205. Springer, 2014.

A Optimal Mono-Bit Distinguisher for Known model and Gaussian Noise

In the mono-bit case, every $Y_i(k)$ $(0 \leq i < m)$ takes only two different values. W.l.o.g., let us assume $Y_i(k) = \pm 1$. Then, $\|\mathbf{y}(k)\|_2^2 = m$ and is thus independent on the key. Thus,

$$\mathcal{D}^{M,G}_{opt(1\text{ bit})}(\mathbf{x}, \mathbf{t}) = \arg\max_k \ \sum_{i|y_i(k)=1} x_i - \sum_{i|y_i(k)=-1} x_i. \tag{36}$$

Surprisingly, this distinguisher *is not* any variant of DoM presented in the seminal paper [8] by Kocher, Jaffe and Jun ($\mathcal{D}^{M,G}_{\text{KJJ}}$) nor in the alleged t-test improvement [3] by Coron, Kocher and Naccache ($\mathcal{D}^{M,G}_{\text{CKN}}$). In particular, $\mathcal{D}^{M,G}_{\text{KJJ}}(\mathbf{x}, \mathbf{t}) = \arg\max_k \ \overline{\mathbf{x}_{+1}} - \overline{\mathbf{x}_{-1}}$ and $\mathcal{D}^{M,G}_{\text{CKN}}(\mathbf{x}, \mathbf{t}) = \arg\max_k \ (\overline{\mathbf{x}_{+1}} - \overline{\mathbf{x}_{-1}}) \Big/ \sqrt{\frac{\sigma^2_{\mathbf{x}_{+1}}}{n_{+1}} + \frac{\sigma^2_{\mathbf{x}_{-1}}}{n_{-1}}}$ where $n_{\pm 1} = \sum_{i|y_i(k)=\pm 1} 1$, $\sigma^2_{\mathbf{x}_{\pm 1}} = \frac{1}{n_{\pm 1}-1} \sum_{i|y_i(k)=\pm 1} (x_i - \overline{\mathbf{x}_{\pm 1}})^2$ and $\overline{\mathbf{x}_{\pm 1}} = \frac{1}{n_{\pm 1}} \sum_{i|y_i(k)=\pm 1} x_i$. However, when m is large, the two distinguishers $\mathcal{D}^{M,G}_{opt(1\text{ bit})}$ and $\mathcal{D}^{M,G}_{\text{KJJ}}$ become equivalent, as $n_{\pm 1} \approx m/2$ (independently of k, using an argument similar to that of Footnote 6). But even in this case, $\mathcal{D}^{M,G}_{\text{CKN}}$ is non-equivalent with them. We notice that the normalization $\mathcal{D}^{M,G}_{\text{CKN}}$ is useful when there are many samples, since it normalizes the difference between $Y(k) = -1$ and $Y(k) = +1$ (hence avoid ghost peaks), but this consideration is out of the scope of this paper.

The success rate of all three attacks for $\sigma = 1$ is displayed in Fig. 7 showing that the optimal distinguishing rule (Eq. (36)) is the most efficient to reach a empirical success rate $\hat{\mathbb{P}}_s = 90\%$. For $\sigma > 1$ all 3 distinguishers were found almost equivalent, which is reasonable. Those results highlight that *intuitive* distinguishers (such as $\mathcal{D}^{M,G}_{\text{KJJ}}$, that aims at showing a difference of leakage) or *classic* (such as $\mathcal{D}^{M,G}_{\text{CKN}}$, based on the well-established t-test) distinguishers are not necessarily the best.

B Noise Distribution Resulting from Multiplication

When combining two leakage samples multiplicatively in case of Gaussian noise, the noise distribution is no longer following a Gaussian distribution. More precisely,

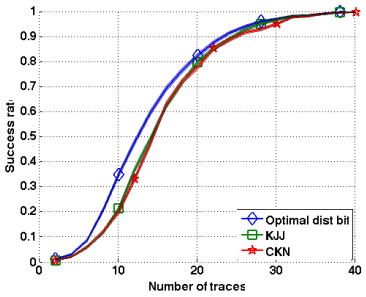


Fig. 7: Success rate for one-bit attacks

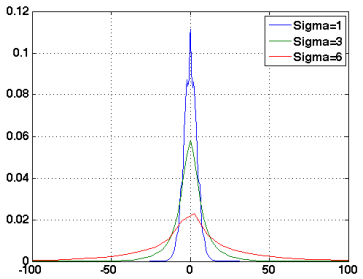


Fig. 8: Empirical distribution of X_1X_2

let us assume we have two leakages $X_1 = Y_1(k^\star) + N_1$ and $X_2 = Y_2(k^\star) + N_2$ that are multiplied, then $X_1X_2 = (Y_1(k^\star) + N_1) \cdot (Y_2(k^\star) + N_2) = Y_1(k^\star) \cdot Y_2(k^\star) + Y_2(k^\star) \cdot N_2 + Y_2(k^\star) \cdot N_1 + N_1 \cdot N_2$. Due to the product, the distribution of X_1X_2 is no longer Gaussian. Figure 8 displays the empirical distribution in this case, which looks similar to a Laplacian distribution (compare to Fig. 2b).