Mutual Information Analysis
A Universal Differential Side-Channel Attack

Benedikt Gierlichs[1], Lejla Batina[1], and Pim Tuyls[1,2]

[1] K.U. Leuven, ESAT/SCD-COSIC
Kasteelpark Arenberg 10, B-3001 Leuven-Heverlee, Belgium
{bgierlic,lbatina,ptuyls}@esat.kuleuven.be

[2] Philips Research
Eindhoven, The Netherlands

Abstract. In this paper, we develop an information theoretic differential side-channel attack. An embedded device containing a secret key is modeled as a black box with a leakage function whose output is captured by an adversary through the noisy measurement of a physical observable *e.g.* the power consumed by the device. We assume only that the measured values depend somehow on the leakage and thus on the word being processed by the device. Without any knowledge on the particular dependency, this fact is exploited to mount a side-channel attack. We build a distinguisher which uses the Mutual Information between the observed and the leaked values as a statistical test. The Mutual Information is maximal when the hypothetical key guessed by the attacker equals the key in the device. Our approach is confirmed by experimental results. We perform power analysis on an embedded device using our Mutual Information based distinguisher and show that the correct key is clearly distinguishable. Finally, our approach allows to compute a good estimate of the minimal number of traces required to perform a successful attack and gives an upper bound on the information leakage in a single observation.

Keywords: Differential Side Channel Analysis (DSCA), Information Theory, Mutual Information

1 Introduction

Currently, embedded devices form the major part of the CPU market [16]. It seems that the vision of Pervasive Computing or Ambient Intelligence is being realised since we are more and more surrounded by devices such as smart cards, mobile phones, PDAs and more recently RFIDs and sensor nodes. These devices typically operate in hostile environments under control of a potential attacker and hence the data contained in them might be relatively easy compromised. Due to the tight constraints on their resources such as memory, number of gates, power *etc.*, it is a very challenging task to protect the information they carry in an adequate way.

The physical accessibility has led to a number of new very powerful attacks that include physical tampering and side-channels. As an example we mention Differential Power Analysis (DPA) [10] which demonstrates that by monitoring the power line of a smart card reader, the cryptographic keys can be rather efficiently extracted if no special countermeasures are taken. More precisely, two types of power attacks were introduced; Simple Power Analysis (SPA) and Differential Power Analysis. The main difference is that SPA exploits the properties of a single (or a few averaged) power measurements, while DPA exploits the statistical differences in a large set of observations.

In the last decade many other side-channels have been described such as electromagnetic emanation [17], timing [9], acoustic [19] *etc.* Both, theory and practice have been developed and as a consequence several more advanced attacks have been proposed such as template [3] and higher-order attacks [12]. In parallel, a broad range of countermeasures has been put forward [4, 7, 8, 11]. For all side-channels we use the terminology Differential Side Channel Analysis (DSCA) when we refer to Differential Attacks.

DPA attacks as introduced by Kocher *et al.* [10] use a partitioning function to sort all power curves into two subsets. The partitioning function is defined by a special selection bit (*e.g.* the lsb) within an intermediate value of a (cryptographic) computation which can be predicted on the basis of a key hypothesis and a plaintext. The difference between the averages of the power consumption curves of these two subsets shows a clear peak for the correct key guess. To this distance of means test based on a partitioning of observed values we refer as a side-channel distinguisher. Other often used distinguishers are based on the Hamming weight or Hamming distance of some intermediate value which depends on a number of key and plaintext bits. In this paper we focus on univariate analysis, *i.e.* all functions considered are assumed to have time as at most one independent variable, but the work can be extended to the multivariate case.

Recently, a new research area appeared which deals with theoretical models for physical attacks in general and side-channel attacks in particular: Physical Observable Cryptography [14]. This line of research attempts to introduce the notion of provable security into cryptosystems that leak some side-channel information. Such attacks require new models and new definitions of an adversary. Micali and Reyzin have evaluated some basic theorems of traditional black box cryptography and they have shown that these results do not hold in this new setting. In their physical observable model, the assumptions they made were very strong and their adversary is the strongest possible. This is one of the reasons why their model is hard to work with in practice and difficult to apply to cryptographic primitives such as block ciphers, for which even black box security cannot be proven. This open question was the motivation for the work of Standaert *et al.* [20]. In their attempt to quantify the leakage, they restricted the most general assumptions from [14]. This led to a further refinement of the model and to the classification of adversaries (attacks) and leakage functions.

Our work is not following the same line of research although we also aim to introduce a more theoretical approach to side-channel analysis. More precisely, we follow the information theoretic approach in order to develop a more general attack than those previously known. The question we pose is whether one can perform a successful attack without incorporating any knowledge on the functional relationship between a physical observable, *e.g.* power consumption, and a leakage function, *e.g.* Hamming weight. We are interested in using information theoretic notions and insights to formalize side-channel concepts in such a way that each specific side-channel distinguisher can be seen as an instance of a Mutual Information based distinguisher. While previously used distinguishers are bound to specific classes of leakage functions, our Mutual Information based distinguisher is not.

We illustrate the new approach in a concrete situation and we show how to estimate the minimal number of samples of the physical observable, that are required.

Although information theoretical notions such as key entropy, increased entropy due to noise, *etc.* have been frequently used in DSCA related literature [1, 15, 6], their potential has not been fully exploited.

This paper is organized as follows. Section 2 introduces the basic notions of information theory. In Sect. 3 we introduce an information theoretic model for side-channel attacks and analysis. It leads to the construction of a distinguisher that allows to infer the secret key from observed values. Section 4 provides the theoretical justification of our approach and compares it to so far widely deployed methods. In Sect. 5 we provide empirical evidence for the correctness of our model and its practicability whereas in Sect. 6 we empirically compare it to DPA and Correlation Power Analysis (CPA) [2]. In Sect. 7 we exemplify the applications of Mutual Information beyond key recovery and we conclude our work in Sect. 8.

2 Information Theory

We introduce the basic notions of information theory. For more details we refer to [5].

2.1 Information Theory Preliminaries

Let $\mathbf{X}$ be a random variable on a (discrete) space $\mathcal{X}$ with probability distribution $\mathbb{P}_{\mathbf{X}}$. The uncertainty that one has about the value of such a random variable when an experiment is performed, is expressed by the Shannon entropy of $\mathbf{X}$ which is usually denoted by $\mathsf{H}(\mathbf{X})$ or $\mathsf{H}(\mathbb{P}_{\mathbf{X}})$. It is defined by the following equation

$$\mathsf{H}(X) = -\sum_{x \in \mathcal{X}} \mathbb{P}_{\mathbf{X}}[\mathbf{X} = x] \log_2 \mathbb{P}_{\mathbf{X}}[\mathbf{X} = x]\,. \tag{1}$$

$\mathsf{H}(\mathbf{X})$ expresses the uncertainty in bits. The entropy of the pair of random variables $(\mathbf{X}, \mathbf{Y})$ (where $\mathbf{Y}$ is a random variable on a space $\mathcal{Y}$) is denoted by $\mathsf{H}(\mathbf{X}, \mathbf{Y})$ and it expresses the uncertainty one has about both. We note that the entropy of two random variables is sub-additive *i.e.*

$$\mathsf{H}(\mathbf{X}, \mathbf{Y}) \leq \mathsf{H}(\mathbf{X}) + \mathsf{H}(\mathbf{Y})\,, \tag{2}$$

with equality if and only if $\mathbf{X}$ and $\mathbf{Y}$ are independent. Often one is interested in the uncertainty about $\mathbf{X}$ given that one has obtained the outcome of an experiment on a related random variable $\mathbf{Y}$ belonging to a possibly different space $\mathcal{Y}$. This is expressed by the conditional entropy $\mathsf{H}(\mathbf{X}|\mathbf{Y})$ which is defined as follows,

$$\mathsf{H}(\mathbf{X}|\mathbf{Y}) = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \mathbb{P}_{\mathbf{X},\mathbf{Y}}[\mathbf{X} = x, \mathbf{Y} = y] \log_2 \mathbb{P}_{\mathbf{X}|\mathbf{Y}}[\mathbf{X} = x|\mathbf{Y} = y], \tag{3}$$

where $\mathbb{P}_{\mathbf{X},\mathbf{Y}}$ denotes the joint probability distribution of $\mathbf{X}$ and $\mathbf{Y}$ and $\mathbb{P}_{\mathbf{X}|\mathbf{Y}}$ stands for the conditional probability distribution of $\mathbf{X}$ given $\mathbf{Y}$. When $\mathbf{Y}$ can be considered as an observation of $\mathbf{X}$ over a noisy channel, then one often characterizes the channel by its set of conditional distributions $\{\mathbb{P}_{\mathbf{Y}|\mathbf{X}=x}\}_{x \in \mathcal{X}}$. The reduction in uncertainty on $\mathbf{X}$ that is obtained by having observed $\mathbf{Y}$, is exactly equal to the information that one has obtained on $\mathbf{X}$ by having observed $\mathbf{Y}$. Hence the formula for the Mutual Information $\mathbf{I}(\mathbf{X};\mathbf{Y})$ is given by,

$$\mathbf{I}(\mathbf{X};\mathbf{Y}) = \mathsf{H}(\mathbf{X}) - \mathsf{H}(\mathbf{X}|\mathbf{Y}) = \mathsf{H}(\mathbf{X}) + \mathsf{H}(\mathbf{Y}) - \mathsf{H}(\mathbf{X},\mathbf{Y}) = \mathbf{I}(\mathbf{Y};\mathbf{X})\,. \tag{4}$$

The Mutual Information satisfies $0 \le \mathbf{I}(\mathbf{X};\mathbf{Y}) \le \mathsf{H}(\mathbf{X})$. The lower bound is reached if and only if $\mathbf{X}$ and $\mathbf{Y}$ are independent. The upper bound is achieved when $\mathbf{Y}$ uniquely determines $\mathbf{X}$. Hence, the larger the Mutual Information, the more close the relation between $\mathbf{X}$ and $\mathbf{Y}$ is to a one-to-one relation.

3 Side Channel Model

In this section, we describe a general model and attack methodology to exploit side-channel leakage of cryptographic devices with a minimal set of assumptions (in particular, we assume only that there is a functional relationship between the leaked and observed values). In Sect. 5 we illustrate the results of our model and method in a concrete situation.

3.1 Definitions and Notations

Let $A_1,\dots,A_l$ be a set of subsets of a space $\mathcal{X}$. The set $\mathcal{A} = \{A_1,\dots,A_l\}$ is a partition of $\mathcal{X}$ if and only if $A_i \cap A_j = \emptyset$ for all $i \neq j, i, j = 1,\dots,l$ and $\cup_i A_i = \mathcal{X}$. The elements $A_i, i = 1,\dots,l$ of $\mathcal{A}$ are called atoms.

We model a device (*e.g.* an IC) that carries out a cryptographic operation E_k depending on a secret key k, modeled as the random variable $\mathbf{K}$, as a physical computer $\mathcal{PC}$, *i.e.* an abstract computer $\mathcal{AC}$ with a side channel leakage function $\mathcal{L}$: $(\mathcal{AC},\mathcal{L})$ (*cf.* [14]). The leakage function $\mathcal{L}$ models the fact, that the adversary can observe (up to a certain extent) the internal state of $\mathcal{PC}$. We assume that $\mathcal{L}$ depends on time and on the word w being processed by $\mathcal{PC}$. We model the words w being processed as a random variable $\mathbf{W}$ on $\{0,1\}^n$. Hence the leakage function $\mathcal{L}$ contains information on $\mathbf{W}$. Therefore we model the output values of $\mathcal{L}$ as a discrete random variable $\mathbf{L}$ on a space $\mathsf{L} = \{0,\dots,l\}$. It is furthermore assumed that $l \le 2^n$.

The random variable $\mathbf{L}$[1] is observed by measuring a physical observable $\mathbf{O}$. The physical observable $\mathbf{O}$ is modeled as another random variable, on a continuous space O where $\mathsf{O} = \mathbb{R}$ models the most general case. Summarizing we have a model consisting of a cascade of two channels (cf. Fig.1):

1. $\mathbf{W} \to \mathbf{L}$: the leakage channel through which information on the word w is revealed at some time $t = \tau$.
2. $\mathbf{L} \to \mathbf{O}$: the measurement (observation) channel of the leakage through which $\mathbf{O}$ provides information on $\mathbf{L}$.

During an attack the attacker obtains $q > 0$ observations $o_i, i = 1,\dots,q$, of $\mathbf{O}$.

First, we consider the points of interest in time $t = \tau_i$ when the word w being processed is the result of a function $f_k : \{0,1\}^m \to \{0,1\}^n, x \mapsto f_k(x)$ applied on an

[1] We will simply speak about the random variable $\mathbf{L}$ when we mean the output value of the leakage function $\mathcal{L}$.

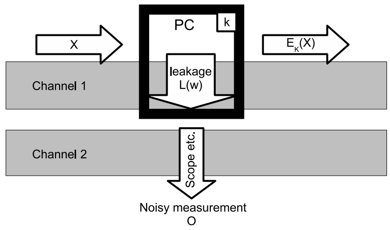

Fig. 1. Schematic illustration of the cascaded channels

input $\mathbf{X}$ (plaintext)[2]. We assume that the cryptographic primitive E_K is known to the attacker and that $f_k(\cdot)$ is an intermediate result of $E_K(\cdot)$. The secret key k is a random variable $\mathbf{K}$ on a key space $\{0,1\}^m$ which is uniformly random distributed. We will focus on a *known plaintext attack* [3] where plaintexts $\mathbf{X}$ are chosen uniformly random from $\{0,1\}^m$.

3.2 Side Channel Attack

We denote by $\mathcal{M} = \{o_{x_1}, \ldots, o_{x_q}\}$ the multi-set[4] of q measurements of the physical observable $\mathbf{O}$ when the (known) inputs $x_1, \ldots, x_q$ were processed by the device. A side channel attacker has to develop a distinguisher $\mathcal{D}$, which takes as input the measurements $o_{x_1}, \ldots, o_{x_q}$ and the plaintexts $x_1, \ldots, x_q$, and creates a non-negligible advantage for retrieving the key in the following experiment:

Experiment $\mathbf{Exp}_{\mathcal{L}}^{sc}$:
$\mathbf{K} \leftarrow_R \{0,1\}^m$
$x_1, \ldots, x_q \leftarrow_R \{0,1\}^m, \quad o_{x_1}, \ldots, o_{x_q} \leftarrow \mathbf{O}$
$k^* \leftarrow \mathcal{D}(o_{x_1}, \ldots, o_{x_q}; x_1, \ldots, x_q)$

The advantage $\mathbf{Adv}(o_{x_1}, \ldots, o_{x_q}; x_1, \ldots, x_q)$ is defined as
$\mathbf{Adv}(o_{x_1}, \ldots, o_{x_q}; x_1, \ldots, x_q) = \mathrm{Prob}[k^* = k]$.

3.3 Construction of an Information Based Distinguisher

To each possible key $k' \in \{0,1\}^m$, we associate a partition $\mathcal{H}_{k'} = \{H_0^{k'}, \ldots, H_l^{k'}\}$ on $\{0,1\}^m$ which is defined by

$$H_i^{k'} = \{x \in \{0,1\}^m \quad | \quad \mathbf{L}(f_{k'}(x)) = i\} \quad \text{for} \quad i = 0, \ldots, l.$$

The partition $\mathcal{H}_{k'}$ induces a subdivision[5] $\mathcal{G}_{k'} = \{G_0^{k'}, \ldots, G_l^{k'}\}$ of the measurement space O. The subdivision $\mathcal{G}_{k'}$ is defined by,

$$G_i^{k'} = \{o_x \in \mathsf{O} |\; x \in H_i^{k'}\}\,.$$

[2] For ease of notation we assume that the key space and the plaintext space are of equal size $\{0,1\}^m$, but generalizations are straightforward.
[3] Note that application to a known ciphertext scenario is straightforward.
[4] A multi-set is a set in which values can appear several times.
[5] In contrast to a partition, the atoms of a subdivision do not necessarily have an empty intersection.

Let $\mathbb{P}_{\mathbf{L}}$ and $\mathbb{P}_{\mathbf{O}}$ denote the probability distributions of the random variables $\mathbf{L}$ and $\mathbf{O}$ respectively. We note that for a given plaintext x, $\mathbf{O}$ depends on the actual key k used by the device while the value of $\mathbf{L}$ depends on the hypothetical key k' guessed by the attacker.

Given the multi-set of measurements $\mathcal{M} = \{o_{x_1}, \ldots, o_{x_q}\}$, and a subdivision $\mathcal{G}_{k'}$ on O, we define the following set of conditional distributions $\left\{\tilde{\mathbb{P}}^{k|k'}_{\mathbf{O}|\mathbf{L}_{k'}=i}(\mathbf{O}|\mathbf{L}=i)\right\}_{i=0}^{l}$. The distributions $\tilde{\mathbb{P}}^{k|k'}_{\mathbf{O}|\mathbf{L}=i}(\mathbf{O} = o|\mathbf{L} = i)$ describe the random variable $\mathbf{O}$ given that $\mathbf{L}(f_{k'}(x)) = i$ for a hypothetical key k'.

They represent a noisy observation channel $\mathbf{L} \to \mathbf{O}$ which depends on the hypothetical key k', the actual key k, and the physical properties of the device and the measurement setup. The distributions $\{\tilde{\mathbb{P}}^{k|k'}_{\mathbf{O}|\mathbf{L}=i}\}_{i=0}^{l}$ are determined empirically by generating the histograms (cf. Sect. 5.1) of the measurements $o_{x_1}, \ldots, o_{x_q}$ belonging to the atoms of $\mathcal{G}^{k'}$.

We define the Mutual Information $\mathbf{I}_{k'k}(\mathbf{L};\mathbf{O})$ under the key guess k' while the actual key is k as follows,

$$\mathbf{I}_{k'k}(\mathbf{L};\mathbf{O}) = \mathsf{H}(\mathbb{P}_{\mathbf{O}}) - \mathsf{H}(\mathbb{P}^{k|k'}_{\mathbf{O}|\mathbf{L}}) , \tag{5}$$

where $\mathbb{P}^{k|k'}_{\mathbf{O}|\mathbf{L}}$ denotes the empirical conditional distribution used in the computation of $\mathsf{H}^{k'}(\mathbf{O}|\mathbf{L})$.

We define our distinguisher $\mathcal{D} : \mathsf{O}^q \times \{0,1\}^m \to \{0,1\}^m$ by the following equation: given a multi-set $\mathcal{M} = \{o_{x_1}, \ldots, o_{x_q}\}$ of observations and the corresponding plaintexts $x_1, \ldots, x_q$,

$$\mathcal{D}(o_{x_1}, \ldots, o_{x_q}; x_1, \ldots, x_q) \mapsto k^* \quad \text{iff} \quad \mathbf{I}_{k^*k}(\mathbf{L};\mathbf{O}) = \max_{k'} \mathbf{I}_{k'k}(\mathbf{L};\mathbf{O}) . \tag{6}$$

The distinguisher $\mathcal{D}$ can be extended to retrieve also the point(s) in time $t = \tau_i$ when the targeted computation $f_k(\cdot)$ happens. Then it takes as input the multi-set of observed traces $\mathcal{M} = \{o_{x_1}(t), \ldots, o_{x_q}(t)\}$ and the plaintexts $x_1, \ldots, x_q$. The extended distinguisher is defined by,

$$\begin{aligned} \overline{\mathcal{D}}(o_{x_1}(t), \ldots, o_{x_q}(t); x_1, \ldots, x_q) &\mapsto (k^*, \tau) \quad \text{iff} \\ \mathbf{I}_{k^*k}(\mathbf{L};\mathbf{O}(\tau)) &= \max_{(k',t)} \mathbf{I}_{k'k}(\mathbf{L};\mathbf{O}(t)) \end{aligned} \tag{7}$$

4 Theoretical Considerations

4.1 Theoretical Justification

As mentioned above, entropy is the uncertainty about the measurements and quantifying this value gives us useful information such as the number of measurements required for a successful attack. Another interesting quantity is Mutual Information that measures the mutual (in)dependence between variables. An advantage is that it can also be applied to non-Gaussian distributions. Hence, at first we want to address side-channel leakage by measuring this value. Intuitively, considering this function of two random variables, the maximum should be obtained for the correct key guess.

As discussed in Sect. 3.1 a physical observable depends on the data words being processed by the device for all time instants t. A leakage function takes as an argument one specific data word that is an intermediate result of the computation. Therefore, it is evaluated only once by the attacker while the device handles the data word at instant(s) $t = \tau_i$. Thus, for all $t \neq \tau_i$ the device processes different words and in this case observables are independent of the leakage function. The issue of partial dependency will be addressed at the end of this section.

More precisely, we consider the Mutual Information between the output of a leakage function $\mathbf{L}$ and an observable $\mathbf{O}$, *i.e.* the reduction in the uncertainty on $\mathbf{L}$ due to the knowledge of $\mathbf{O}$ for a key candidate k', so $\mathbf{I}_{k'k}(\mathbf{L}; \mathbf{O})$ as defined in (5).

If the leakage function $\mathbf{L}$ is computed for the correct key k which corresponds to the observed values $\mathbf{O}$ then the Mutual Information $\mathbf{I}_{kk}(\mathbf{L}; \mathbf{O})$ will obtain the maximal value and for other values k' (incorrect key hypotheses) the value of the Mutual Information $\mathbf{I}_{k'k}(\mathbf{L}; \mathbf{O})$ will be lower.

We consider now $\mathbf{I}_{k'k}(\mathbf{L}; \mathbf{O})$ for incorrect key hypotheses in all time instants t where $t \neq \tau_i$.

$$\begin{aligned}
&\mathbf{I}_{k'k}(\mathbf{L}; \mathbf{O}) = \\
&= \mathsf{H}(\mathbf{L}(f(x_i, k')) + \mathsf{H}(\mathbf{O}(x_i, k)) - \mathsf{H}(\mathbf{L}(f(x_i, k')), \mathbf{O}(x_i, k)) = \\
&= \mathsf{H}(\mathbf{L}(f(x_i, k')) + \mathsf{H}(\mathbf{O}(x_i, k)) - (\mathsf{H}(\mathbf{L}(f(x_i, k')) + \mathsf{H}(\mathbf{O}(x_i, k))) = 0\,.
\end{aligned}$$

The first equality follows by the definition of Mutual Information and the second one from (2). On the other hand, for the correct key $k' = k$ the mutual information $\mathbf{I}_{kk}(\mathbf{L}; \mathbf{O})$ results in a strictly positive value. This follows directly from the non-negativity of Mutual Information and the fact that equality (to zero) holds if and only if two random variables are independent. So, at right time $t = \tau_i$, the correct key leads to the highest Mutual Information while at wrong time instants, incorrect key candidates result in a Mutual Information of zero. The existence of a maximum for Mutual Information as a function in time defined on a key space K is therefore proven. The uniqueness of the arguments follows directly from the assumptions on (in)dependency of a physical observable and the leakage function.

Hence, Mutual Information is theoretically equal to zero for all incorrect key guesses. This holds for all $t \neq \tau_i$ but in practice we also get "peaks" at other moments as it is shown in Sect. 5. The reason for this is that some data processed during the execution of the algorithm may be related with the intermediate data w in time τ_i. We also mention here that in practice we do not get zero but some values close to it as we are working with the noisy observation $\mathbf{O}$ of $\mathbf{L}$.

Some peaks also appear for wrong key guesses, which seems to contradict the theory. These so-called "ghost peaks" occur due to the properties of the leakage function. For example, the Hamming weight of the S-box output can still be partially correlated for two different key guesses. Similar observations with respect to CPA are mentioned in [2].

4.2 Comparison of Mutual Information, DPA and CPA

Here we discuss the comparison of the proposed Mutual Information based distinguisher and two other common DSCA distinguishers, *i.e.* the distance of means test and the correlation test. For the first one, we refer to the work done by Kocher *et al.* [10]. The second one, Correlation Power Analysis [2], estimates the linear correlation coefficient between the leaked and the observed values, which is slightly more computationally expensive than standard DPA, but often gives better results.

CPA can only detect linear correlations and is therefore limited to attack scenarios where a linear approximation is justified. DPA does not require any specific dependency between the target bit and the observable, but is limited to the distance of means test, which uses less information than available. As already mentioned, this fact allows to envision the Mutual Information distinguisher as a generalization of other methods. So, the results obtained by this distinguisher are platform independent and arbitrary relationships between the power consumption and the leaked value can be assumed. The result is always expressed in bits and we can also estimate the number of measurements required for key recovery.

We compare these distinguishers from the experimental point of view in more detail in Sect. 6.

5 Experimental Results for Mutual Information

In this section, we apply the theoretical framework from Sect. 3 and provide experimental results based on measurements from an ATMega163 micro controller (8-bit) performing AES-128 encryption in software.[6] The measurements $\mathbf{O}(t)$ represent the voltage drop over a 50Ω resistor inserted in the smart card reader's ground line. We sample the power consumption at $t = 1, \ldots, 20\,000$ during the first round of the AES-128 encryption of randomly chosen plaintexts[7] with a constant key. Our experiments focus on the first key byte denoted by $\mathbf{K}$ and the first plaintext byte denoted by $\mathbf{X}$, but application to the other bytes as well as to a known cipher text scenario are straightforward.

5.1 Mutual Information Applied to Side Channel Leakage

Our Mutual Information method does not require any assumptions on the dependency between the measured value and the leaked value, except for the one that is fundamental to DSCA: *the leakage and thus the power consumption of a device (partially) depends on the data w it is processing.* We empirically confirm this statement with the following experiment, for which we use a sample size of $q = 50\,000$ power curves $o_i(t)$ $(i = 1, ..., q)$. As leakage function $\mathcal{L}$, we use the value of the S-Box's outcome for the first byte during the first round. Hence, each $o_i(t)$ is assigned to an atom of $\mathcal{G}^{k'}$ by $l_i = \text{S-box}(x_i \oplus k), l_i \in \{0, ..., 255\}$.

First, we compute the sum of squared pairwise differences (*sosd*), *i.e.* we compute the means $m_j(t)$ of $\{o_i(t) \mid l_i = j\}$ $(j = 0, \ldots, 255)$ and sum up their squared pairwise differences (note that we omit the time parameter t):

$$m_j = \frac{1}{|\{o_i|l_i = j\}|} \sum_{o_i|l_i=j} o_i \qquad sosd = \sum_{j,n=0}^{255} (m_j - m_n)^2 \,.$$

Figure 3 shows the resulting *sosd* trace. The obvious peaks appear during the Initial Roundkey Addition, the jointly implemented SubBytes and ShiftRows transformations, and the MixColumn operation.

[6] We would like to point out that the AES encryption terminates in constant time.
[7] To model a known plaintext attack.

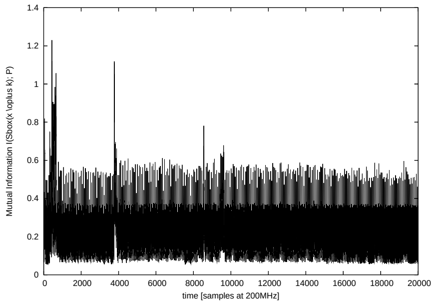

Fig. 2. Mutual Information of $q = 50\,000$ curves

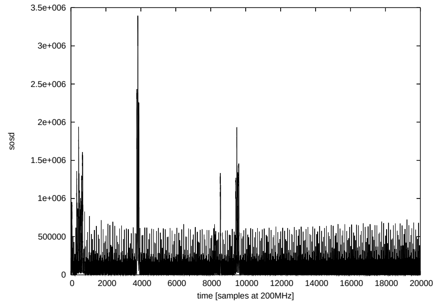

Fig. 3. The sum of squared differences of $q = 50\,000$ curves

Next, we compute the Mutual Information $\mathbf{I}_{kk}(\mathbf{L};\mathbf{O})$ of the output $\mathbf{L}$ of the leakage function and the observed power consumption $\mathbf{O}(t)$ according to (4, 5). The number of bins in the histogram can be chosen according to Scott's rule [18], see the appendix.

We use 256 bins for the histogram of l_i (since $l_i \in 0,\ldots,255$) and for the histogram of $o_i(t)$ (since the resolution of our oscilloscope is 8 bit). This choice allows the maximum precision but also requires a large amount of data. The resulting Mutual Information trace is depicted in Fig. 2. One observes that the peaks in the two plots are synchronous, hence the Mutual Information distinguisher is sensitive to differences in the power consumption. The applied leakage function $\mathcal{L}$ partitions the power curves into identical subsets, independently of the assumed key byte k', since the Initial RoundKey Addition with a constant key byte and the SubBytes substitution are bijections. Hence, for any guess of the key byte the resulting partitions are merely permuted and the statistical tests sosd and Mutual Information (incl. all intermediate entropy values) are independent of a key hypothesis. This means that this partitioning function $\mathcal{L}$ does not allow key recovery. However, while the sosd metric only allows to reveal the point(s) of interest $t = \tau_i$, the Mutual Information distinguisher additionally provides the adversary with an estimate of the maximum Mutual Information $\mathbf{I}_{kk}(\text{S-box}(x \oplus k);\mathbf{O})$ in *bits*. This important figure gives an upper bound on the number of (secret) bits an adversary can learn from a single curve, on average. Only such a partition can lead to this estimate of the maximum Mutual Information, because it treats each possible byte of the 8-bit implementation separately. Applications of this number will be discussed in Sect. 7.

5.2 Empirical Evidence

This section aims at providing empirical evidence that the requirements from Sect. 3 are fulfilled and hence at confirming the theoretical considerations of Sect. 4.

The Mutual Information metric, as most other statistical tests, is bound in its efficiency to recover keys by the leakage function $\mathcal{L}$. The closer the partitioning by $\mathbf{L}$ is to the unknown physical data-dependency inherent in $\mathbf{O}$, the more significant the outcome of the statistical test will be. Note, however, that knowledge of the dependency between the atoms of the partition and the atoms of the subdivison (CPA needs a linear dependency) is not required, as has been shown in Sect. 5.1.

In the next experiment, we apply the well-studied and widely agreed-on Hamming weight Model[8] combined with the Mutual Information distinguisher to a set of $q = 1\,000$ power curves $\mathbf{O}(t)$. We denote HW(w) as the number of bits set to "1" in the eight-bit word w, *i.e.* $\mathrm{HW(w)} = \sum_{i=1}^{8} w_i$, $\mathrm{HW(w)} \in \{0, \ldots, 8\}$. Based on a key guess $k' \in \{0, \ldots, 255\}$, each $o_i(t)$ is assigned to an atom of $\mathcal{G}^{k'}$ by $l_i = \mathrm{HW}(\text{S-box}(x_i \oplus k'))$.

We compute the Mutual Information of the distributions $\mathbf{L}$ and $\mathbf{O}(t)$ according to (4, 5). We set the number of bins for the histogram of l_i to 9 and as $o_i(t)$ ideally is a one-to-one function of l_i (if the Hamming Weight Model was correct), we also use 9 bins for the histogram of $o_i(t)$. Figure 4 depicts the resulting Mutual Information trace for the correct key guess $k' = k$. As can be seen when comparing to Fig. 2 and 3, the trace shows

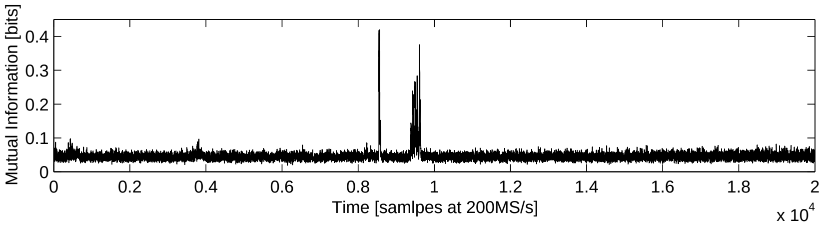

Fig. 4. Mutual Information over time for the correct key hypothesis

clear peaks at some[9] of the points of interest $t = \tau_i$. Hence, the first requirement from Sect. 3 is empirically confirmed. To verify whether the second requirement is fulfilled, we compute the same Mutual Information trace for all other key hypotheses k' and test, if the highest derived Mutual Information value for any wrong k' is lower than the one for $k' = k$. More formally that is: $\mathrm{argmax}_t\, \mathbf{I}_{kk}(\mathbf{L}, \mathbf{O}(t)) > \mathrm{argmax}_{t,k' \neq k}\, \mathbf{I}_{k'k}(\mathbf{L}, \mathbf{O}(t))$. Figure 5 shows the highest Mutual Information value (selected from the whole time frame) for every key hypothesis. The peak for the correct key hypothesis $k' = k$ is clearly distinguishable. Figure 6 shows the Mutual Information trace for the second best but wrong key hypothesis. The height of the visible "ghost peaks" is less than a third of the height of the peak for the correct hypothesis and they appear at different instants. Obviously, the second requirement from Sect. 3 is empirically confirmed. The maximum Mutual Information value achieved for $k' = k$ will be discussed in Sect. 7.

6 Comparison to DPA and CPA

In this section, we compare Mutual Information to other widely accepted and adopted statistical tests.

[8] Note that our experimental platform implements a Harvard architecture and pre-charges its bus to "0", so that the Hamming weight is equivalent to the bus' toggle count.

[9] The peaks appear during the MixColumns operation, when the S-box output leaks most due to our AES implementation. The fact that no peaks appear *e.g.* during the Initial RoundKey Addition is explained by the lack of partial dependencies due to the Hamming Weight Model.

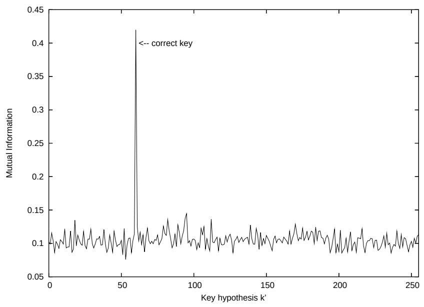

Fig. 5. Maximum Mutual Information per key hypothesis

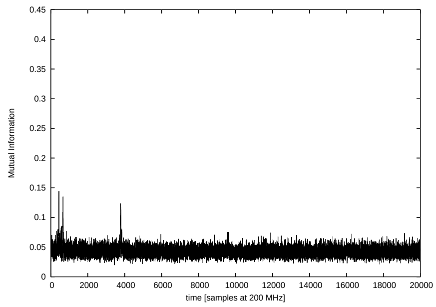

Fig. 6. Mutual Information over time for the second best key hypothesis

6.1 Comparison to Correlation Power Analysis

CPA as proposed in [2], estimates the Pearson Correlation Coefficient between a vector of observations $\mathbf{O}(t)$ and a vector of predictions $\mathbf{L}$

$$\rho_{\mathbf{LO}}(t) = \frac{q \sum o_i(t) l_i - \sum o_i(t) \sum l_i}{\sqrt{q \sum o_i(t)^2 - (\sum o_i(t))^2} \sqrt{q \sum l_i^2 - (\sum l_i)^2}} \,. \tag{8}$$

The summations are taken over the q measurements and the correlation coefficient has to be estimated for each time slice $t = 1, \ldots, T$ within the power curves $\mathbf{O}(t)$. We apply CPA to a set of power curves $o_i(t)$ $(i = 1, \ldots, q)$ and form the q predictions according to l_i= HW(S-box($x_i \oplus k'$)).[10] To show the impact of the population size, we use $q = 1, \ldots, 1000$. Fig. 7 shows the maximum correlation coefficient, *i.e.* the maximum from the overall time frame, for each key hypotheses k' on the vertical axis over the population size q on the horizontal axis. The plot shows that the correct key hypothesis

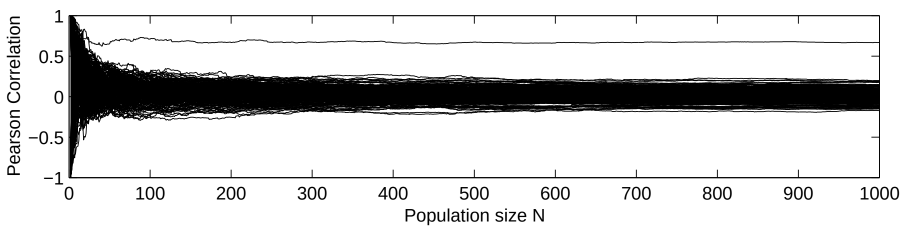

Fig. 7. Max. and min. correlation for each k' over the number of samples used

$k' = k$ is clearly distinguishable from about $q = 30$ upwards.

We now repeat the experiment but use the distinguisher Mutual Information instead of the correlation coefficient. More precisely, we use the same set of q power curves $o_i(t)$, and the same partitioning function l_i = HW(S-box($x_i \oplus k'$)) in order to compute the Mutual Information $\mathbf{I}_{kk'}(\mathbf{L}; \mathbf{O})$ according to (4,5) for each time slice t. Again we use 9 bins for the histograms. Figure 8 shows the result of this experiment in the same manner as used for Fig. 7. The plot shows that the correct key hypothesis $k' = k$ is

[10] Note that the 'reference state' mentioned in [2] is "0" in our scenario due to the pre-charged Harvard Architecture of our experimental platform. Therefore the proposed Hamming Distance is equal to the Hamming weight.

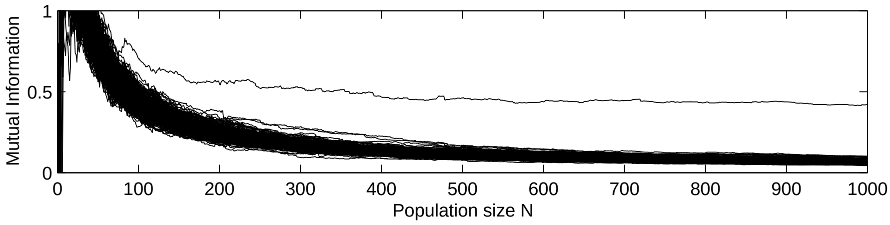

Fig. 8. Max. Mutual Information for each k' over the number of samples used

clearly distinguishable from approximately $q = 75$ upwards.

Summarizing the results: CPA is able to recover the correct key from a smaller population size than Mutual Information ($q \approx 30$ vs. $q \approx 75$). These results are explained if one considers the differences in the approaches, more precisely in the power models. CPA assumes the linear relation $o_i = a \cdot l_i + b$ between the Hamming weight l_i and the measured power consumption o_i and is therefore limited to finding *linear* correlations between $\mathbf{O}$ and $\mathbf{L}$ (cf. [2]). Mutual Information on the other hand makes no further assumption on the relation between the atoms H_i^k of a partition and their typical power consumption inherent in the atoms g_i^k of the subdivision, so that every possible power model (in this case in nine variables) is plausible. A linear dependency seems to be a good first approximation of our platform's power model and thus CPA needs less measurements. However, the true power model of our platform (based on the Hamming weight assumption) seems to be more complex since the correlation coefficient does not get close to its maximum value 1.

6.2 Comparison to Differential Power Analysis

(Single-bit) DPA as proposed in [10] computes the DPA bias signal

$$\Delta(t) = \frac{\sum_i o_i(t) l_i}{\sum_i l_i} - \frac{\sum o_i(t)(1 - l_i)}{\sum_i (1 - l_i)} \tag{9}$$

as the difference between the average of all measurements for which a so called target bit is 0 and the average of all measurements for which the target bit is 1. The summations are taken over the q samples and the bias signal has to be computed for each time slice within the power measurements $\mathbf{O}(t)$. We apply DPA to a set of power curves $o_i(t)$ $(i = 1, \ldots, q)$ and use the least significant bit of S-box$(x_i \oplus k')$ as the target bit l_i. Again, we use $q = 1, \ldots, 1\,000$ to show the impact of the population size. Figure 9 shows the maximum DPA bias for each key hypotheses k', *i.e.* the maximum from the overall time frame, on the vertical axis over the population size q on the horizontal axis. The plot shows that the correct key hypothesis $k' = k$ is clearly distinguishable from approximately $q = 490$ upwards. Comparing these results to Mutual Information: Mutual Information is able to reliably recover the key from a smaller population than single-bit DPA ($q \approx 75$ vs. $q \approx 490$). These results are explained if one considers the differences in the approaches, *i.e.* the power models, once again. Single-bit DPA considers the mean values of two sets of measurements for which the target bit is either 1 or 0 and assumes that these means must differ. Mutual Information on the other hand considers not only the mean value of each set, but its entropy and thus the distribution

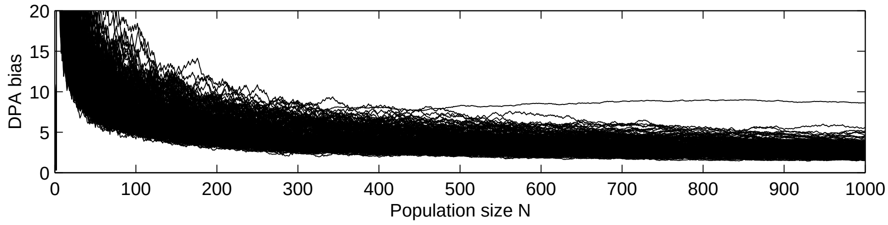

Fig. 9. Maximum single bit DPA difference per key hypothesis

of the values in the set. This explains why Mutual Information recovers the key from fewer measurements.

We also tried multi-bit DPA as proposed in [13] and used 2 to 8 bits for the target bit function. However, in line with the conclusions of [2] and [13] we observed that, in our setup, multi-bit DPA leads to worse results than single-bit DPA due to the small population size q.

7 Application of the Results

This section discusses applications of Mutual Information analysis beyond key recovery.

7.1 How Many Curves Do We Need on Average?

By combining the leakage model with the notions of Fig. 1 we estimate the minimum number of measurements needed on average for non-ambiguous key recovery. We exemplify the approach using the Hamming weight Model.

In the Hamming weight Model, the input to channel 1 is $\mathbf{W} = \text{S-box}(x \oplus k)$ and its output $\mathbf{L}$ is the Hamming weight of the input. The Mutual Information of channel 1 can be computed analytically. We assume that the inputs to channel 1, hence the plaintexts, originate from a uniformly random distribution. Let $w_i = (0, 1, 2, \ldots, 255)$ and $l_i \in \{0, \ldots, 8\}$ (the Hamming weights of $\mathbf{W}$) for all i. Then the Mutual Information of channel 1 is given by $\mathbf{I}(\mathbf{W}; \mathbf{L}) = 2.5442$ bits.

The input to channel 2 is $\mathbf{L}$ and at its output we get $\mathbf{O}$, a noisy observation of $\mathbf{L}$. From the experiment in Sect. 6 we estimate the Mutual Information of channel 2 as $\mathbf{I}(\mathbf{L}; \mathbf{O}) = 0.41$ bits.

The uncertainty on $\mathbf{L}$ is $\mathsf{H}(\mathbf{L}) = 2.5442$ bits when the keys and plaintexts are chosen uniformly at random. The uncertainty on $\mathbf{W}$ and hence the key $\mathbf{K}$ is $\mathsf{H}(\mathbf{K}) = \mathsf{H}(\mathbf{W}) = log_2(2^8) = 8$ bits.

The task of an attacker is to track bits backwards through both channels in order to learn $\mathbf{W}$ and thus the key. She needs to learn 2.5442 bits from channel 2 in order to know one Hamming weight l_i. This implies that on average she will need to observe $\mathbb{E}(o_i) = \frac{2.5442}{0.41} = 6.21$ measurements o_i from the same Hamming weight l_i to learn its value. Considering $\mathbf{L}$ as a random variable with probability distribution $\mathbb{P}_{\mathbf{L}} = (\frac{1}{256}, \frac{8}{256}, \ldots, \frac{1}{256})$ (where the probabilities are ordered according to increasing Hamming weights) which reflects uniformly distributed plain texts, we can compute the number v_i of measurements needed for each Hamming weight l_i as $v_i = \frac{\mathbb{E}(o_i)}{\mathbb{P}_{\mathbf{L}}(l_i)} = \frac{6.21}{\mathbb{P}_{\mathbf{L}}(l_i)}$.

Finally, the weighted mean of the numbers of required measurements for each Hamming weight, $\mathbb{E}(\mathbf{M}) = \sum_i \mathbb{P}_{\mathbf{L}}(l_i) * v_i$, estimates the minimum number of measurements that are required on average to learn one Hamming weight[11]. For our setup this number is 55 measurements. The fact that this number slightly deviates from our experimental analysis, where we need ≈ 75 measurements, can be explained by the impossibility of observing 256 different plaintexts as a uniform distribution in less than 256 experiments. Hence, for a more precise estimation one needs to know the probability distribution of the plain texts. Further, our estimation does not cover the occurrence of redundant information, *e.g.* multiple observations of the same plaintext.

7.2 Application of the Maximum Mutual Information

The maximum amount of possible information leakage is a major concern for all manufacturers of secure embedded devices. Usually this figure is unknown and the security of a device is evaluated by exposing it to efficient attacks. If the attacks are successful, the device is equipped with additional countermeasures, and the procedure is repeated until the desired security level is reached.

The results of a Mutual Information analysis as presented in Sect. 5 provides a manufacturer with a very good estimate of the maximal possible information leakage. Based on his knowledge about the efficiency of the available countermeasures, where efficiency denotes the increased uncertainty of an attacker, the manufacturer can directly choose an appropriate set of them and circumvent the costly and lengthy evaluation cycle.

8 Conclusion

We have introduced Information theoretical concepts to DSCA and constructed a side channel distinguisher based on Mutual Information, that efficiently and practically performs under relaxed assumptions and can be seen as a generalization of all previously applied statistical tests. In particular, we relax the assumption that a side channel adversary needs insight in the dependency of observations and leaked values. To carry out a successful Mutual Information based attack, the only requirement for an adversary is to know for *which words* processed the leakage differs. This means that the attacker does not need to know *how* the observation differs with the leakage.

We have also shown applications of Mutual Information in DSCA beyond key recovery. In short it allows to asses the maximal possible information leakage, which is an important figure to manufacturers of secure embedded devices, and to estimate the minimal number of observations needed.

References

1. L. Batina and C. Jansen. Side-Channel Entropy for Modular Exponentiation Algorithms. In L. Tolhuizen, editor, *Proceedings of the 24th Symposium on Information Theory in the Benelux*, pages 37–44, Veldhoven, The Netherlands, May 2003. Werkgemeeschap voor Informatie-en-Communicatietheorie, Enschede, The Netherlands.
2. E. Brier, C. Clavier, and F. Olivier. Correlation power analysis with a leakage model. In M. Joye and J.-J. Quisquater, editors, *Proceedings of 6th International Workshop on Cryptographic Hardware and Embedded Systems (CHES)*, number 3156 in Lecture Notes in Computer Science, pages 16–29. Springer-Verlag, 2004.

[11] We note that even if we have several known plaintexts for only one Hamming weight, the key can be recovered exactly.

3. S. Chari, J.R. Rao, and P. Rohatgi. Template attacks. In B.S. Kaliski Jr., Ç.K. Koç, and C. Paar, editors, *Proceedings of 4th International Workshop on Cryptographic Hardware and Embedded Systems (CHES)*, number 2523 in Lecture Notes in Computer Science, pages 172–186. Springer-Verlag, 2002.
4. J.-S. Coron and L. Goubin. On Boolean and Arithmetic Masking against Differential Power Analysis. In Ç.K. Koç and C. Paar, editors, *Proceedings of 2nd International Workshop on Cryptographic Hardware and Embedded Systems (CHES)*, number 1965 in Lecture Notes in Computer Science, pages 231–237. Springer-Verlag, 2000.
5. T.M. Cover and J.A. Thomas. Elements of Information Theory. John Wiley & Sons, 2006.
6. E. Dewitte, B. De Moor, and B. Preneel. Information theoretic measures applied to power and electromagnetic traces from an FPGA performing an elliptic curve point multiplication. In B.S. Kaliski Jr., Ç.K. Koç, and C. Paar, editors, *Proceedings of 25th Symposium on Information Theory in the Benelux*, 2004.
7. J.D. Golić and C. Tymen. Multiplicative masking and power anaylsis of AES. In B.S. Kaliski Jr., Ç.K. Koç, and C. Paar, editors, *Proceedings of 4th International Workshop on Cryptographic Hardware and Embedded Systems (CHES)*, number 2535 in Lecture Notes in Computer Science, pages 198–212. Springer-Verlag, 2002.
8. L. Goubin. A sound method for switching between boolean and arithmetic masking. In Ç.K. Koç, D. Naccache, and C. Paar, editors, *Proceedings of 3rd International Workshop on Cryptographic Hardware and Embedded Systems (CHES)*, number 2162 in Lecture Notes in Computer Science, pages 3–15. Springer-Verlag, 2001.
9. P. Kocher. Timing attacks on implementations of Diffie-Hellman, RSA, DSS and other systems. In N. Koblitz, editor, *Advances in Cryptology: Proceedings of CRYPTO'96*, number 1109 in Lecture Notes in Computer Science, pages 104–113. Springer-Verlag, 1996.
10. P. Kocher, J. Jaffe, and B. Jun. Differential power analysis. In M. Wiener, editor, *Advances in Cryptology: Proceedings of CRYPTO'99*, number 1666 in Lecture Notes in Computer Science, pages 388–397. Springer-Verlag, 1999.
11. T.S. Messerges. Securing the AES finalists against power analysis attacks. In B. Schneier, editor, *In Proceedings of 7th International Workshop on Fast Software Encryption Workshop (FSE)*, number 1978 in Lecture Notes in Computer Science. Springer-Verlag, 2000.
12. T.S. Messerges. Using second-order power analysis to attack DPA resistant software. In Ç.K. Koç and C. Paar, editors, *Proceedings of the 2nd International Workshop on Cryptograpic Hardware and Embedded Systems (CHES)*, volume 1965 of *Lecture Notes in Computer Science*, pages 238–251. Springer-Verlag, 2000.
13. T.S. Messerges, E.A. Dabbish, and R.H. Sloan. Examining smart-card security under the threat of power analysis attacks. *IEEE Trans. Comput.*, 51(5):541–552, 2002.
14. S. Micali and L. Reyzin. Physically observable cryptography. In M. Naor, editor, *In Proceedings of 1st Theory of Cryptography Conference TCC*, number 2951 in Lecture Notes in Computer Science, pages 278–296. Springer-Verlag, 2004.
15. S. Nikova, C. Rechberger, and V. Rijmen. Threshold implementations against side-channel attacks and glitches. In P. Ning, S. Qing, and N. Li, editors, *Proceedings of Information and Communications Security, 8th International Conference, ICICS 2006*, number 4307 in Lecture Notes in Computer Science, pages 529–545. Springer-Verlag, 2006.
16. C. Paar. The Next Five Years of Embedded Security: Ad-hoc Networks and BMWs. In ECRYPT workshop - Cryptographic Advances in Secure Hardware - CRASH 2005, September 6-7 2005. Invited talk.
17. J.-J. Quisquater and D. Samyde. ElectroMagnetic Analysis (EMA): Measures and Couter-Measures for Smard Cards. In I. Attali and T. P. Jensen, editors, *Smart Card Programming and Security (E-smart 2001)*, volume 2140 of *Lecture Notes in Computer Science*, pages 200–210. Springer-Verlag, 2001.
18. D. Scott. On optimal and data-based histograms. Biometrika, Vol. 66, No. 3., pp. 605–610, 1979.
19. A. Shamir and E. Tromer. Acoustic cryptanalysis. `http://theory.csail.mit.edu/~tromer/acoustic/`.
20. F.-X. Standaert, T.G. Malkin, and M. Yung. A formal practice-oriented model for the analysis of side-channel attacks. Cryptology ePrint Archive, Report 2006/139, 2006. `http://eprint.iacr.org/`.

Appendix

Scott's rule [18] defines the optimum bin width b_n^* as

$$b_n^* = \left(\frac{6}{\int_{-\infty}^{\infty} f'(x)^2 dx} \right)^{\frac{1}{3}} n^{-\frac{1}{3}}$$

where n is the number of measurements and f is the underlying probability distribution. For Gaussian distributions $b_n^* = 3.49 s n^{-\frac{1}{3}}$, where s is the empirical standard deviation. If one assumes Gaussian noise in the observations, he derives the number of bins as $\frac{\max(\mathsf{O}) - \min(\mathsf{O})}{b_n^*}$.