Revisiting Higher-Order DPA Attacks:
Multivariate Mutual Information Analysis

Benedikt Gierlichs, Lejla Batina, and Ingrid Verbauwhede

K.U. Leuven, ESAT/SCD-COSIC and IBBT
Kasteelpark Arenberg 10, B-3001 Leuven-Heverlee, Belgium
firstname.lastname@esat.kuleuven.be

Abstract. Security devices are vulnerable to side-channel attacks that perform statistical analysis on data leaked from cryptographic computations. Higher-order (HO) attacks are a powerful approach to break protected implementations. They inherently demand multivariate statistics because multiple aspects of signals have to be analyzed jointly. However, all published works on HO attacks follow the approach to first apply a pre-processing function to map the multivariate problem to a univariate problem and then to apply established 1^{st} order techniques. We propose a novel and different approach to HO attacks, Multivariate Mutual Information Analysis (MMIA), that allows to directly evaluate joint statistics without pre-processing. While this approach can benefit from a good power model, it also works without an assumption. A thorough empirical evaluation of MMIA and established HO attacks confirms the overwhelming advantage of the new approach: MMIA is more efficient and less affected by noise. Most important and opposed to all published approaches, MMIA's measurement cost grows sub-exponentially with the attack order. As a consequence, the security provided by the masking countermeasure needs to be reconsidered as 3^{rd} and higher order attacks become very practical.

1 Introduction

Embedded devices such as smart cards, mobile phones, RFID tags are becoming increasingly pervasive. In order to secure the applications, these devices execute cryptographic algorithms and protocols to authenticate data and entities and to protect the confidentiality of sensitive data. An embedded device is by definition physically accessible and it is very likely that the device falls into the hands of a malicious user. The physical accessibility has led to a number of very powerful attacks that include physical tampering and side-channels. A typical example is Differential Power Analysis (DPA) [10]. The technique explores weaknesses of implementations rather than algorithms, allowing an attacker to extract the secret of a device by monitoring its power dissipation, if no special countermeasures are taken. A successful DPA attack is subject to two conditions: i) there exists an intermediate variable in the implementation that is correlated with the power consumption and ii) this variable exclusively depends on the plaintext (or ciphertext) and a small part of the key.

In order to protect devices against DPA, one can get rid of the second condition by data randomization or masking [4]. The idea is to conceal intermediate values through addition or multiplication with random values, which makes it impossible to correctly predict the intermediate variable.

However, this so-called 1^{st} order masking succumbs to higher-order DPA attacks (HODPA) as originally proposed by Messerges [12] and Chari et al. [2]. These HODPA attacks are based on the joint statistical properties of multiple aspects of the signal, typically joint analysis of the power consumption at two (or more) points in time. In this case one would think of multivariate analysis but all established techniques [12, 2, 14, 19, 9] rely on a pre-processing step to map the multivariate problem to a univariate problem before attacking the result with a standard DPA attack.

HODPA attacks imply higher costs in terms of number of samples and computational complexity. In addition, the identification of points in time at which to take the signals is a hard problem. Another common issue is the pre-processing step: while this problem has been studied by many authors, finding the optimal transformation is still an open problem. Furthermore, it is evident that none of the solutions is generic as each pre-processing is tightly linked to a leakage model, that is not always met in practice [15]. Eventually, it is unclear how the pre-processing functions can be generalized for attacks of order higher than two without accepting enormous drawbacks.

Our contribution solves all but one of the aforementioned problems. At CHES 2008 a side-channel distinguisher called Mutual Information Analysis (MIA) was introduced [7]. This 1^{st} order attack is effective without any knowledge or assumption about the particular dependencies between processed data and observable power consumption. Our proposal, multivariate MIA (MMIA), inherits this important property. Also the issue of pre-processing unravels because MMIA is explicitly multivariate which furthermore allows to easily adapt it to attacks of order higher than two. What remains is the problem of identifying the points in time at which to take the signals. Our theory is confirmed by an extensive empirical evaluation of MMIA and established HODPA attacks against 1^{st} and 2^{nd} order masked software implementations of a DES like mini-cipher.

This paper is organized as follows. In Sect. 2 we summarize related work on MIA, the concept of masking and in particular HODPA attacks. In Sect. 3 we discuss our motivation and formulate a generic 2^{nd} order attack problem, for which we present a sound solution in Sect. 4. Section 5 deals with the empirical evaluation of our proposal and HODPA attacks and it gives empirical evidence for the advantage of MMIA in HO attack scenarios. We conclude our work in Sect. 6.

2 Related Work

2.1 Notation and Mutual Information Analysis

A device performs a cryptographic computation $E_k(x)$ under some key k from a keyspace $\mathcal{K} = \{0,1\}^m$. Since the key is unknown it is modeled as a random variable (RV) $\mathbf{K}$ on $\mathcal{K}$ with a priori uniform probability mass function (PMF). The information leakage of the device, due to its physical properties, is modeled by the side channel leakage function L. The values of L depend on state transitions in the device which are created by a word (w) being processed. Since the word of interest typically depends on k it is a priori unknown and hence another RV $\mathbf{W}_k$ on a space $\mathcal{W} = \{0,1\}^n$ where n is the word length of the device. As the input of L depends on the key, so does the output and hence we model the values of the leakage function as $\mathbf{L}_k$ on $\mathcal{L} = \{0, ..., l\}$ with $l \leq 2^n$. An adversary observes $\mathbf{L}_k$ by measuring a physical observable (here power consumption) which is modeled as RV $\mathbf{O}$ on a space $\mathcal{O}$. To remind the reader that the observation $\mathbf{O}$ depends on the actual key used in the device and on the device's leakage function L we write $\mathbf{O}_{\mathsf{L},k}$.

MIA is a generic but straight-forward variant of 1^{st} order DPA. The central problem is to decide whether two RVs with certain PMFs are correlated (in a statistical sense). The RV $\mathbf{O}_{\mathsf{L},k}$ is the measurement of the power consumption. The other RV $\mathbf{L}_k$ is the leakage of a predictable intermediate result of the computation that depends on a (small) part of the key. DPA applies a statistical test $T(\mathbf{O}_{\mathsf{L},k}, \mathbf{L}_{k'})$ for all key guesses k' that measures whether the RVs are correlated or not. The value of k' that leads to the highest correlation is an adversary's best guess. In this context, statistical tests are frequently called distinguishers in the literature [18].

MIA's core is the mutual information based distinguisher. Let $\mathbf{I}(\mathbf{L}'_k; \mathbf{O}_{\mathsf{L},k})$ denote the mutual information

$$\mathbf{I}(\mathbf{L}_{k'}; \mathbf{O}_{\mathsf{L},k}) = \mathsf{H}(\mathbf{O}_{\mathsf{L},k}) - \mathsf{H}(\mathbf{O}_{\mathsf{L},k}|\mathbf{L}_{k'}), \tag{1}$$

between $\mathbf{L}'_k$ and $\mathbf{O}_{\mathsf{L},k}$, where $\mathsf{H}(\cdot)$ denotes Shannon entropy. MIA evaluates $\mathbf{I}(\mathbf{L}_{k'}; \mathbf{O}_{\mathsf{L},k})$ for all key hypotheses k' and the value of k' that leads to the highest mutual information is an adversary's best guess.

Details. Typically the exact leakage behavior of the device, represented by L, is unknown to the adversary and has to be estimated. In real-world scenarios, the PMFs of the RVs involved are unknown but have been sampled. To estimate the PMFs we follow the approach of [7] and use histograms with l bins. Typically the observations $\mathbf{O}_{\mathsf{L},k}$ are power *traces* $\mathbf{O}_{\mathsf{L},k}(t)$. The above holds for the interesting point(s) in time $t = \tau_j$ at which the device actually processes the targeted words (or correlated values).

Simplifications. For most of this paper we make three (non-restrictive) assumptions that simplify the analysis and allow us to keep the notation simple, which helps the reader to focus on the content:

- The adversary can estimate the leakage function L reasonably well. This allows us to continue writing L.
- The measurement channel is ideal, i.e. a bijective map $\mathbf{L}_k \Leftrightarrow \mathbf{O}_{\mathsf{L},k}$. As bijective maps are transparent to MIA [7], we can model the adversary with direct access to $\mathbf{L}_k$ instead of the physical measurement.
- Sampling of PMFs is sufficient and entropy estimations are good. This allows us to continue writing H.

Under these assumptions Eq. (1) becomes $\mathbf{I}(\mathbf{L}_{k'}; \mathbf{L}_k) = \mathsf{H}(\mathbf{L}_k) - \mathsf{H}(\mathbf{L}_k|\mathbf{L}_{k'})$. Note that all assumptions are to the adversary's advantage and security observations therefore also hold for weaker adversaries. We revisit these practical issues later in Sect. 4.

2.2 Masking

Masking is usually implemented by combining the intermediate values with random data and by adapting the algorithm accordingly. The effect of masking is that each intermediate value that is predictable by an attacker is pairwise uncorrelated to the masked intermediate values that are actually processed. As the dynamic instantaneous power consumption of a device depends on the data that it processes, it depends on the masked intermediate values and is thus uncorrelated to predictable (unmasked) values.

In the following we formalize these notions. Let $\mathbf{X}$ be the input of the algorithm and $\mathbf{M}$ denote a RV with uniform PMF. Masking is implemented by replacing the intermediate value $\mathbf{W}_k = f_k(\mathbf{X})$ by $\mathbf{W}_k \circ \mathbf{M} = f'_k(\mathbf{X}, \mathbf{M})$ where $\circ$ is a suitable operation. In masked block cipher implementations, for example, one often uses the exclusive-or operation and replaces $\mathbf{W}_k$ by $\mathbf{W}_k \oplus \mathbf{M}$. An S-box table lookup S-box$(\mathbf{W}_k^{in})$ can be implemented with a recomputed S-box such that $\mathbf{W}_k^{out} = \text{S-box'}(\mathbf{W}_k^{in} \oplus \mathbf{M}) = \text{S-box}(\mathbf{W}_k^{in}) \oplus \mathbf{M}$.

If $\mathbf{M}$ is an RV with uniform PMF, intermediate results $\mathbf{W}_k$ predictable by an adversary (i.e. $k' = k$) are not correlated to computed intermediate results $\mathbf{W}_k \circ \mathbf{M}$. It follows that $\mathbf{I}(\mathsf{L}(\mathbf{W}_k); \mathsf{L}(\mathbf{W}_k \circ \mathbf{M})) = 0$.

2.3 Higher-Order Attacks

The mounting point for 2^{nd} order attacks is the fact that the side channel leakage $\mathsf{L}(\mathbf{W}_k \circ \mathbf{M})$ of a masked value depends on a predictable value $\mathbf{W}_k$ and an unpredictable value $\mathbf{M}$. The core idea is to jointly analyze the leakage of the masked value and the mask (or a second masked value masked with the same mask) to establish a relation to the predictable $\mathbf{W}_k$. In the example in Fig. 1 (left) one could combine the leakage $\mathsf{L}(\mathbf{W}_k \circ \mathbf{M})$ at time

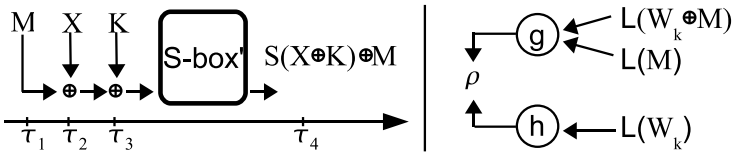

Fig. 1. Left: Masked S-box lookup with recomputed S-box. Right: Schematic of 2^{nd} order DPA with pre-processing functions g and h that output correlated values

τ_4 with the leakage at any other time, e.g. $\mathsf{L}(\mathbf{M})$ at time τ_1 which is the example we use from now on.

2^{nd} order DPA requires a suitable pre-processing that combines leakages of two masked RVs such that the leakage $\mathsf{L}(\mathbf{W}_k)$ of the targeted and unmasked intermediate result (or a function thereof) is revealed and can be attacked with 1^{st} order DPA. More formally that is: one looks for functions $g(\mathbf{L}_{\tau_1}, \mathbf{L}_{\tau_4})$ and $h(\mathsf{L}(\mathbf{W}_k))$ that yield highly correlated values [13], see Fig. 1 (right). These values can be attacked with 1^{st} order DPA.

Early proposals for the pre-processing did not consider the function h but focused on a function g whose outputs are correlated with $\mathsf{L}(\mathbf{W}_k)$ and mentioned two essential options: the product of the two leaked values and the absolute value of their difference.

The first work showing a practical higher-order attack to defeat the masking countermeasure came from Messerges [12]. He assumed that the device leaks the Hamming weight (HW) of intermediate values (i.e. $\mathsf{L}(\cdot) = \mathrm{HW}(\cdot)$) and proposed to compute the absolute difference $|\mathrm{HW}(\mathbf{W}_k \oplus \mathbf{M}) - \mathrm{HW}(\mathbf{M})|$ in the pre-processing (abs-diff-DPA). Messerges showed that, when focusing on a single bit,

$$|\mathrm{HW}(\mathbf{W}_k \oplus \mathbf{M}) - \mathrm{HW}(\mathbf{M})| = \mathrm{HW}(\mathbf{W}_k \oplus \mathbf{M} \oplus \mathbf{M}) = \mathrm{HW}(\mathbf{W}_k)\,. \quad (2)$$

Thus, the pre-processing reveals the unmasked $\mathrm{HW}(\mathbf{W}_k)$ which can be attacked with 1^{st} order DPA. If one wants to attack more than a single bit simultaneously Eq. (2) changes to

$$\mathrm{HW}(\mathbf{W}_k) = \mathrm{HW}(\mathbf{W}_k \oplus \mathbf{M}) + \mathrm{HW}(\mathbf{M}) - 2 \cdot \mathrm{HW}((\mathbf{W}_k \oplus \mathbf{M}) \wedge \mathbf{M}) \quad (3)$$

where $\wedge$ denotes bitwise AND. However, an adversary cannot evaluate this function because $\mathbf{W}_k \oplus \mathbf{M}$ and $\mathbf{M}$ are unknown.

The Hamming weight assumption was also used by Oswald et al. in [14]. They showed that the idea of Eq. (2) can still be used to attack multiple bits although the equality no longer holds. For 8-bit variables, the Pearson correlation coefficient (ρ) between the predictable $\mathrm{HW}(\mathbf{W}_k)$ and the output of the abs-diff pre-processing decreases to 0.24.

Chari et al. [2] suggested to use the product $\mathrm{HW}(\mathbf{W}_k \oplus \mathbf{M}) \cdot \mathrm{HW}(\mathbf{M})$ in the pre-processing (product-DPA). Their technique does not require the

ideal Hamming weight model but still makes some restrictive assumptions about the leakage and power consumption behavior and is in practice more vulnerable to deviations from the model. Waddle and Wagner [19] were the first to clearly split higher-order attacks into pre-processing and attack step as we present them here. They proposed a few variants of product-DPA that differ in complexity. The work of Joye et al. [9] introduces a more theoretical approach to 2^{nd} order DPA. The authors analyzed single bit 2^{nd} order abs-diff-DPA in the Hamming weight model, as introduced by Messerges, and in the Hamming distance model. They suggest to use a power of the absolute difference in the pre-processing, which yields a slightly higher coefficient ρ [14].

3 Motivation

An adversary faces three essential problems when mounting a HO-DPA:[1]

1. How to identify the points of interest τ_j when the interesting intermediate values leak?
2. How to model the power consumption at these points in time? This question is particularly interesting as the power consumption model need not be the same at several instants, e.g. while a random number generator is active vs. during a table lookup.
3. How to choose the functions g and h for the pre-processing? This question is particularly interesting because the answer is tightly linked to the previous question.

In this paper we will not deal with the first problem but assume that the instants τ_j are known.[2] The same assumption is made in all related literature except for [14].

In our view previous work discusses 2^{nd} and HODPA in specific contexts and under restrictive assumptions, that are not always met in practice [14, 15] but the drawbacks have been accepted. In particular we point out that most contributions assume i) the linear Hamming weight or distance model and implicitly require them for the attack to work, due to the choice of g (and h where applicable); and ii) that the leakage functions associated to the computation of different intermediate results, where different parts of the device may be active, are the same or very similar.

It is surprising that, while DPA and its higher-order variants have been published more than 10 years ago, the problem of finding an optimal pre-processing (even for a specific context) remains unsolved. To the best of our knowledge, the practitioner's choice is Messerges' abs-diff-DPA either targeting a single bit, which is sound in the pre-processing but has to deal

[1] Note that, concerning items two and three, Messerges' "absolute difference" pre-processing is sound, i.e. flawless, for 1-bit values.

[2] Some advice can be found in Appendix A.

with algorithmic noise, or targeting multiple bits, where algorithmic noise is reduced but the pre-processing suboptimal.

A weakness of earlier work is that tight link between leakage model(s) and pre-processing. It is evident that a pre-processing tailored to specific leakage functions looses all meaning if the leakage models are not met. Note that such practical issues were even mentioned in original papers [2] and the authors suggest pre-pre-processing steps to attempt to fix them.

The only works that relax the above mentioned assumptions or that could be accordingly adapted [1, 13] deal with variants of template attacks [3], which consider an adversary who is able to characterize the leakage function(s) of the target device and the implementation as well as the electrical properties of the measurement setup. The adversarial context of a profiled attack is, however, beyond the scope of this paper.

3.1 Problem Statement

We formulate a generic 2^{nd} order DPA problem that relaxes all (but one) of the above mentioned assumptions and requirements. Informally speaking, we ask "what is possible" if i) the power models at the two (known) instants are unknown and possibly substantially different, which implies ii) that the best choice for g and h in the pre-processing is *a priori* unknown. Further, the sought method should naturally extend to attacks of order greater than 2. A sound solution would be a powerful tool that allows successful HO attacks in a range of scenarios otherwise resistant or inaccessible to standard attacks due to intrinsic errors introduced in the pre-processing [15, 12, 2, 14, 19, 9].

Formally, let $\mathsf{L}_{\tau_1}(\mathbf{M})$ denote the leakage at time τ_1 and $\mathsf{L}_{\tau_2}(\mathbf{W}_k \circ \mathbf{M})$ denote the leakage at instant τ_2. Further, let L_{τ_1} and L_{τ_2} be arbitrary surjective mappings. Given leakage $(\mathsf{L}_{\tau_1}(\mathbf{M})\ ,\ \mathsf{L}_{\tau_2}(\mathbf{W}_k \circ \mathbf{M}))$ determine k with non-negligible advantage over a random guess. Note that solving the problem does not necessarily require a transformation step (i.e. a pre-processing).

4 Extending MIA to Multivariate Analysis

It is natural to look for methods that can solve the above stated problems without making possibly wrong assumptions. In [7] Gierlichs et al. showed that MIA is a 1^{st} order attack that works without (restrictive) assumptions about the leakage function. We can thus use MIA to solve the problem of unknown leakage functions L_{τ_1} and L_{τ_2}. Since MIA is well suited to exploit dependencies between RVs without making an assumption about *how* the random variables are related, it appears natural to also use this technique to solve the second problem, i.e. joining the information contained in $\mathsf{L}_{\tau_1}(\mathbf{M})$ and $\mathsf{L}_{\tau_2}(\mathbf{W}_k \circ \mathbf{M})$.

The extension of MIA to a multivariate scenario is straight forward: one merely computes the mutual information of three RVs

$$\mathbf{I}(\mathsf{L}(\mathbf{W}_{k'}); \mathsf{L}_{\tau_2}(\mathbf{W}_k \circ \mathbf{M}); \mathsf{L}_{\tau_1}(\mathbf{M}))\,. \tag{4}$$

In [11, 6, 5] it is shown that

$$\mathbf{I}(\mathbf{X}; \mathbf{Y}; \mathbf{Z}) = \mathbf{I}(\mathbf{X}; \mathbf{Y}) - \mathbf{I}(\mathbf{X}; \mathbf{Y}|\mathbf{Z})\,. \tag{5}$$

Depending on the source, Eq. (5) is either called multivariate mutual information or mutual interaction. It is clear that Eq. (5) can have positive and negative values depending on the relation between the arguments. For example, if $\mathbf{X}$ and $\mathbf{Y}$ are independent but possibly related through $\mathbf{Z}$ as in our context, then

$$\mathbf{I}(\mathbf{X}; \mathbf{Y}; \mathbf{Z}) = \mathbf{I}(\mathbf{X}; \mathbf{Y}) - \mathbf{I}(\mathbf{X}; \mathbf{Y}|\mathbf{Z}) = 0 - \mathbf{I}(\mathbf{X}; \mathbf{Y}|\mathbf{Z}) \leq 0$$

and one says that $\mathbf{Z}$ explains the correlation between $\mathbf{X}$ and $\mathbf{Y}$. Note that the choice of how to substitute the arguments is arbitrary, any combination works. The MMIA key recovery attack decides for the key hypothesis k' that minimizes expression (4). For the more general case of n^{th} order MIA attacks one computes

$$\mathbf{I}(\mathbf{X}_1; \ldots; \mathbf{X}_{N+1}) = \mathbf{I}(\mathbf{X}_1; \ldots; \mathbf{X}_N) - \mathbf{I}(\mathbf{X}_1; \ldots; \mathbf{X}_N|\mathbf{X}_{N+1})\,.$$

We want to emphasize that our proposal has one clear advantage: there is no need to assume leakage functions neither to choose the functions g and h for the pre-processing. This makes it generic and applicable in virtually any scenario as long as their exist a chain of information channels all the way from a processed word down to the physical observable.

4.1 Further Applications

MMIA can also perform n^{th} order attacks using more than n side-channel signals. The simplest example is a 1^{st} order attack using two or more instants τ_j simultaneously. Since each intermediate value of a cryptographic computation leaks at least twice (first it is computed, then it is used at least once), joint analysis of the corresponding instants τ_j is an advantage. This idea corresponds directly to using multiple points of interest in template attacks [3].

4.2 Relaxing Assumptions

Before we test our theory in practice, we need to address the simplifying assumptions, formulated in Sect. 2, that we made up to now.

The idealized measurement channel. Following [7] we assume that, given the values of the leakage functions L_{τ_1} and L_{τ_2}, the link between leaked values and measurement of the observables $\mathbf{O}_{\mathsf{L}_{\tau_1},\mathbf{M}}$ and $\mathbf{O}_{\mathsf{L}_{\tau_2},k}$ can be approximated as a bijective relation. In practice this relation is possibly disturbed by the impact of noise and inaccuracy of the measurement, which implies a loss of information. Shannon's noisy-channel coding theorem [16, 17] suggests that the loss be canceled by repeated observation.

The hypothetical leakage function(s). We assumed that the adversary can estimate the leakage function(s) well. In some cases this might be difficult and our solution should not rely on a good estimation. By assumption, the leakage behavior is a deterministic map $\mathsf{L} : \mathcal{W} \mapsto \mathcal{L}$. Following [7] we assume all L to be bijective maps as this guarantees that no information is lost, possibly at the cost of a limited decrease in efficiency. We denote a hypothetical leakage function by L' and its output by $\mathbf{L}'$. Estimating L' well improves the performance of the attack because not only is no information lost but also the information is exploited more efficiently.

Sampling of PMFs. We cannot guarantee a sufficient sampling quality and good entropy estimators. The reader be reminded that, in the remainder of the paper, all information theoretic quantities $\mathbf{I}$ and H are estimates. We chose not to introduce a new symbol (e.g. $\tilde{\mathbf{I}}$) to keep the notation clear. Using Eq. (5) and substituting the leakage with the respective observation, expression (4) becomes

$$\mathbf{I}(\mathbf{L}'_{k'}; \mathbf{O}(\tau_1); \mathbf{O}(\tau_2)) = \mathbf{I}(\mathbf{O}(\tau_1); \mathbf{O}(\tau_2)) - \mathbf{I}(\mathbf{O}(\tau_1); \mathbf{O}(\tau_2) | \mathbf{L}'_{k'}) \,. \qquad (6)$$

5 Reality Check

We study our approach and confront its performance with established HODPA attacks in three scenarios with varying level of difficulty. In each scenario, we apply MMIA with and without the assumption of a Hamming weight leakage function as well as abs-diff-DPA (targeting one and multiple bits) and product-DPA targeting multiple bits.

We consider a Boolean masking scheme as for example described in [8] for DES or triple-DES. For simplicity we focus on a representative mini-cipher consisting of data masking, key addition and a single S-box lookup of the first S-box (S_1). This reflects the setting depicted in Fig. 1. In practice, we pre-compute the values $\mathbf{M} = m_i$ and $\mathbf{W}_k \circ \mathbf{M} = S_1(x_i \oplus k) \oplus m_i$ for each encryption i on a desktop computer and send them to the card, which successively moves the values over its data-bus at τ_1 and τ_2. The data bus is reset to $0x00$ before and after each memory access. All attacks are provided with the physical measurements $\mathbf{O}(\tau_1)$ and $\mathbf{O}(\tau_2)$. Note that

unmasked values are never processed by the card and that $\mathbf{M}$ as well as $\mathbf{W}_k \circ \mathbf{M}$ exist on $\{0,1\}^4$.

For our experiments we use an 8-bit RISC microcontroller based smart-card. The power measurements represent the voltage drop over a 10Ω resistor inserted in the smart-card's GND. We developed implementations for three different scenarios: A) 1^{st} order masking exactly as described above; B) 1^{st} order masking where the side-channel leakage is affected by unpredictable algorithmic noise; C) 2^{nd} order masking (without algorithmic noise).

Following the recent advances concerning the comparison of univariate side-channel distinguishers [18] we apply the first-order success rate to assess the performance of the attacks. The first-order success rate expresses the probability that, given n measurements, the attack's best guess is the correct key. For each scenario, we acquired a set of 100 000 power curves using random masks and plaintexts. To evaluate one scenario, we split the set into 100 packs v_i $(i = 1, \ldots, 100)$ of 1000 curves and do the following:

`for` n `:=` 25 `to` 1000
$counter \leftarrow 0$
(a) `for` i `:=` 1 `to` 100
  i. `select the first` n `curves from set` v_i
  ii. `run the attack for` $k' \in \{0,1\}^6$
  iii. `increase` $counter$ `if attack successful`
(b) `compute success rate for` n `curves as` $counter/100$

For MMIA the best key guess minimizes Eq. (6). For all other methods the best key guess gives rise to the highest correlation coefficient (in absolute terms).

For MMIA we assume L_{τ_1} and L_{τ_2} to be bijective mappings (in our setting 4bit to 4bit). This choice affects the number of bins that one should use for the histograms because it defines the size of the spaces $\mathbf{L}_{\tau_1}$ and $\mathbf{L}_{\tau_2}$ exist on. As we assume that $\mathbf{L} \Leftrightarrow \mathbf{O}$ holds at both time instants, we use 16 bins for the histograms of $\mathbf{O}(\tau_1)$ and $\mathbf{O}(\tau_2)$. We also use 16 bins for the histograms of the leakage $\mathbf{L}'_{k'}$ of the predictable (unmasked) S-box output $S_1(x_i \oplus k)$. For MMIA using the Hamming weight assumption, we have $\mathsf{L}_{\tau_j}(\cdot) = \mathrm{HW}(\cdot)$ and we use 5 bins for all of the histograms.

For all other attacks with explicit pre-processing we assume the linear Hamming weight model at both instants. The pre-processing functions are

$$\breve{\mathbf{O}} = g(\mathbf{O}(\tau_1), \mathbf{O}(\tau_2)) = |\mathbf{O}(\tau_1) - \mathbf{O}(\tau_2)| \quad \text{abs-diff-DPA}$$
$$\breve{\mathbf{O}} = g(\mathbf{O}(\tau_1), \mathbf{O}(\tau_2)) = \mathbf{O}(\tau_1) \cdot \mathbf{O}(\tau_2) \quad \text{product-DPA}\,.$$

All attacks target the same unmasked intermediate result $\mathbf{W}_k = S_1(\mathbf{X} \oplus k)$ which does not give rise to 1^{st} order leakage. For abs-diff-DPA targeting a single bit we predict only the least significant bit of $\mathbf{W}_k$. The attacks evaluate $\rho(\breve{\mathbf{O}}, \mathrm{HW}(\mathbf{W}_{k'}))$.

In scenario A the leakage of the device at τ_1 and τ_2 is very close to the linear Hamming weight model. Given the mask values m_i, we obtain at τ_1

and τ_2 a Pearson $\rho > 0.99$. Figure 2 shows the results for the five attacks in scenario A.

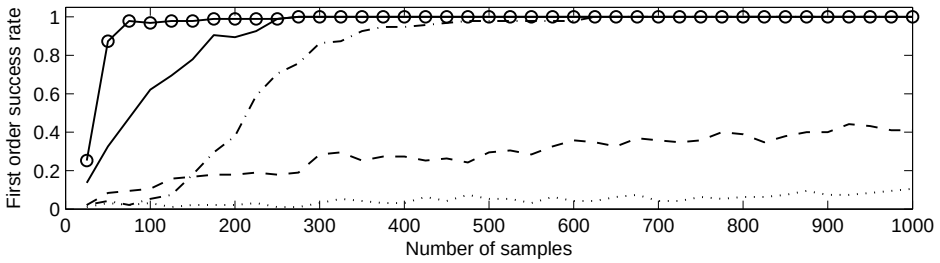

Fig. 2. First order success rates, 2^{nd} order DPA: abs-diff-DPA (solid), 1-bit abs-diff-DPA (dashed), product-DPA (dotted), generic MMIA (dash-dotted), MMIA with HW assumption (solid and $\circ$)

We can see that abs-diff-DPA (solid) performs well in this scenario. About 80 curves suffice to achieve a success rate of 50% and starting from about 275 curves the attack reveals the correct key with success rate 1. Single-bit abs-diff-DPA (dashed) and in particular product-DPA (dotted) perform much worse. The success rates stay well below 50% even when using 1000 measurements. Generic MMIA (dash-dotted) is, as one can expect, less efficient than abs-diff-DPA in this scenario but eventually reaches success rate 1 using 630 curves. About 220 curves are required to achieve 50% success. MMIA using the HW assumption (solid and $\circ$) outperforms all other attacks. The attack needs about 35 curves to achieve 50% success rate, about 80 curves to achieve success rates $> 95\%$ and just like abs-diff-DPA about 275 curves to be reliable. This attack is particularly interesting as it takes uncertainty about the leakage functions out of the equation and shows the impact of sound joint statistics. Both attacks, this variant of MMIA and abs-diff-DPA, use the corret power model. The advantage of MMIA is entirely due to the way it combines the side-channel information. We conclude that the attack that uses the correct power model and sound joint statistics is most efficient. However, we remind that scenario A is a somewhat easy target.

In Appendix B we present results of 2^{nd} order attacks in the slightly more challenging scenario B.

In scenario C we wanted to find out how MMIA scales with respect to attacks of order greater than 2 and extended scenario A to 2^{nd} order masking. The S-box output values are concealed by two independent random masks $\mathbf{M}_1$ and $\mathbf{M}_2$. The masks leak at τ_1 and τ_2, the double-masked S-box output at τ_3. The attack approach remains mostly the same and we compute $\mathbf{I}(\mathbf{L}'_{k'}; \mathbf{O}(\tau_1); \mathbf{O}(\tau_2); \mathbf{O}(\tau_3))$.

Note, the leakage behavior of the card is again very close to the Hamming weight model, which allows us to compare our results to simulation based results[3] of 3^{rd} order DPA attacks in the literature. We also present experimental results for 3^{rd} order product-DPA while for abs-diff-DPA we could not even find a reference on how to compute it.[4]

Figure 3 shows our experimental results for generic MMIA, MMIA using the Hamming weight assumption, and product-DPA. We can observe that

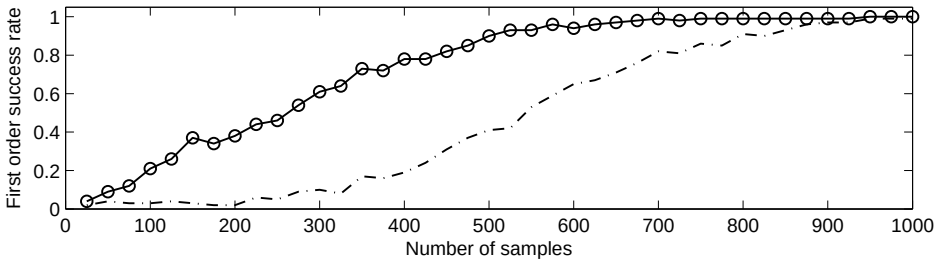

Fig. 3. First order success rates for 3^{rd} order MMIA: generic MMIA (dash-dotted), MMIA with HW assumption (solid and ∘), product-DPA(dotted, permanently at 0)

both MMIA variants require less than 1000 measurements to reach a success rate 1 while the success rate of product-DPA is permanently 0. The figure shows the clear advantage of MMIA in this challenging scenario. We also tested product-DPA using all 100 000 measurements for scenario C, but the attack did not detect the correct value of the key.

According to [15] 3^{rd} order product-DPA requires more than 400 000 measurements in noise free and Hamming weight model based simulations and more than 3.5 million measurements in more realistic simulations. These values were, however, derived for attacks against 8-bit intermediate results. According to our own simulation of scenario C (no noise and Hamming weight model) 3^{rd} order product-DPA requires more than 50 000 measurement samples to reach success rate 1. Our observations go along with the prevailing opinion that, with respect to HODPA attacks, the measurement cost grows exponentially with the order of the attack [15, 2, 14]. Concerning MMIA this is clearly not the case.

6 Conclusion

Confronted with a new problem, one typically first tries to transform it into another problem for which one knows the solution. HODPA attacks seem to

[3] We are not aware of published experimental results for 3^{rd} order attacks, suggestions by the reviewers are most welcome.

[4] Again, suggestions by the reviewers are most welcome.

be such a problem. They inherently demand multivariate statistics because multiple aspects of signals have to be analyzed jointly. However, all publications on HO-attacks follow the approach to first apply a pre-processing function to map the multivariate problem to a univariate problem and then to apply established 1^{st} order DPA techniques. All proposed pre-processing functions have drawbacks that are accepted at the price of an exponential growth of the measurement cost with the attack order. We propose a novel and different approach for HO attacks that does not suffer from intrinsic errors. The empirical evidence confirms the overwhelming advantage of MMIA over established HODPA attacks in easy and more challenging 2^{nd} order attack scenarios but particularly concerning attacks of order greater than 2. MMIA's measurement cost grows sub-exponentially with the attack order. As a consequence, the security provided by the masking countermeasure needs to be reconsidered as 3^{rd} and higher-order attacks become very practical. The typically implemented combination of masking and temporal randomization should render MMIA attacks more difficult.

A Identifying the Points of Interest

One approach towards identifying these instants may be to examine the empirical variance of several power traces obtained during processing of constant input data. In this case, the variance in the power traces is mostly caused by the masking and thus reveals the points in time when masked values are processed. Another approach is to select a small time window based on an educated guess and to perform an exhaustive search over all pairs of time instants [14].

B A slightly more Challenging 2^{nd} Order Attack Scenario

Scenario B is slightly more challenging than scenario A as the side-channel leakage is affected by unpredictable algorithmic noise. Still, the predictable leakage follows the Hamming weight model reasonably well. Given the mask values m_i we obtain at τ_1 and τ_2 a Pearson $\rho \sim 0.75$. Figure 4 shows the results for the five attacks in scenario B. We can see that the performance of all attacks is affected by the noise. In particular, the success rates of all 2^{nd} order attacks with pre-processing decreases drastically. None of them reaches a success rate of 10% even when using 1000 measurements. This decrease is caused by the deviation from the HW model. Comparing the performance of both MMIA variants to scenario A, we make a couple of observations. MMIA using the HW assumption is more affected than generic MMIA. The amount of measurments required for it to reach a success rate of 1 is almost doubled ($\approx 275 \rightarrow \approx 450$). This is a direct consequence of the Hamming weight assumption, which is no longer valid. Still, the assumption is meaningful to some extent and this attack remains the most efficient.

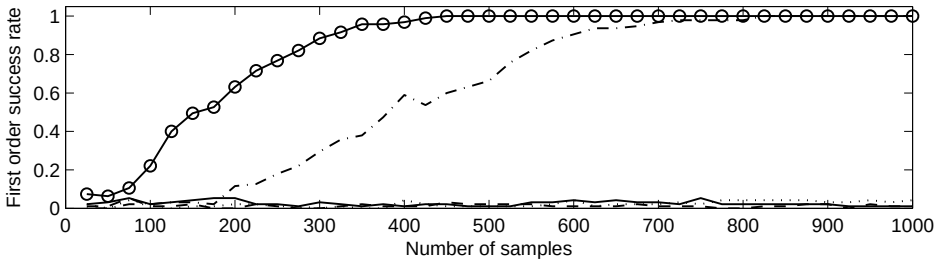

Fig. 4. First order success rates: DPA with abs. diff. (solid), DPA with prod. (dotted), 1-bit DPA with abs. diff. (dashed), generic second-order MIA (dash-dotted), second-order MIA with HW (solid and ∘)

Generic multivariate MIA is less affected. The number of measurements required to reach a success rate 1 increases by roughly 30% from ≈ 630 to ≈ 830. This decrease can not be caused by a wrong model assumption and we thus assign it to the algorithmic noise in the measurements.

References

1. D. Agrawal, J. R. Rao, P. Rohatgi, and K. Schramm. Templates as master keys. In Josyula R. Rao and Berk Sunar, editors, *Cryptographic Hardware and Embedded Systems - CHES 2005, 7th International Workshop, Edinburgh, UK, August 29 - September 1, 2005, Proceedings*, number 3659 in Lecture Notes in Computer Science, pages 15–29. Springer-Verlag, 2005.
2. S. Chari, C.S. Jutla, J.R. Rao, and P. Rohatgi. Towards sound approaches to counteract power-analysis attacks. In M. Wiener, editor, *Advances in Cryptology: Proceedings of CRYPTO'99*, number 1666 in Lecture Notes in Computer Science, pages 398–412. Springer-Verlag, 1999.
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. R. M. Fano. Transmission of Information: A Statistical Theory of Communications. MIT Press, Cambridge, MA, USA, 1961.
7. B. Gierlichs, L. Batina, P. Tuyls, and B. Preneel. Mutual information analysis - a generic side-channel distinguisher. In Elisabeth Oswald and Pankaj Rohatgi, editors, *Cryptographic Hardware and Embedded Systems - CHES 2008*, volume 5154 of *Lecture Notes in Computer Science*, pages 426–442. Springer-Verlag, 2008.
8. L. Goubin and J. Patarin. DES and Differential Power Analysis (The "Duplication" Method). In Ç.K. Koç and C. Paar, editors, *Proceedings of 1st International Workshop on Cryptographic Hardware and Embedded Systems (CHES)*, number 1717 in Lecture Notes in Computer Science, pages 158–172. Springer-Verlag, 1999.

9. M. Joye, P. Paillier, and B. Schoenmakers. On second-order differential power analysis. In Josyula R. Rao and Berk Sunar, editors, *Cryptographic Hardware and Embedded Systems - CHES 2005, 7th International Workshop, Edinburgh, UK, August 29 - September 1, 2005, Proceedings*, number 3659 in Lecture Notes in Computer Science, pages 293–308. Springer-Verlag, 2005.
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. W. J. McGill. Multivariate information transmission. *Psychometrika*, (19):97–116, 1954.
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. E. Oswald and S. Mangard. Template attacks on masking - resistance is futile. In Masayuki Abe, editor, *Topics in Cryptology - CT-RSA 2006, The Cryptographers' Track at the RSA Conference 2006, San Jose, CA, USA, February 13-17, 2006, Proceedings*, volume 4377 of *Lecture Notes in Computer Science*, pages 243–256. Springer-Verlag, 2006.
14. E. Oswald, S. Mangard, C. Herbst, and S. Tillich. Practical Second-Order DPA Attacks for Masked Smart Card Implementations of Block Ciphers. In David Pointcheval, editor, *Topics in Cryptology - CT-RSA 2006, The Cryptographers' Track at the RSA Conference 2006, San Jose, CA, USA, February 13-17, 2006, Proceedings*, volume 3860 of *Lecture Notes in Computer Science*, pages 192–207. Springer, 2006.
15. K. Schramm and C. Paar. Higher Order Masking of the AES. In David Pointcheval, editor, *Topics in Cryptology - CT-RSA 2006, The Cryptographers' Track at the RSA Conference 2006, San Jose, CA, USA, February 13-17, 2006, Proceedings*, volume 3860 of *Lecture Notes in Computer Science*, pages 208–225. Springer, 2006.
16. C. E. Shannon. A Mathematical Theory of Communication. Bell System Technical Journal, volume 27, 1948.
17. C. E. Shannon. Communication in the Presence of Noise. Proceedings of the IRE, vol. 37, no. 1, pp. 1021, reprinted in Proceedings if the IEEE, vol. 86, no. 2, 1998, 1949.
18. Franois-Xavier Standaert, Benedikt Gierlichs, and Ingrid Verbauwhede. Partition vs. comparison side-channel distinguishers. In *Information Security and Cryptology - ICISC 2008: 11th International Conference*, Lecture Notes in Computer Science, page 16, Seoul,KR, 2008. Springer-Verlag.
19. J. Waddle and D. Wagner. Towards efficient second-order power analysis. In Marc Joye and Jean-Jacques Quisquater, editors, *Proceedings of 6th International Workshop on Cryptographic Hardware and Embedded Systems (CHES)*, number 3156 in Lecture Notes in Computer Science, pages 1–15. Springer-Verlag.