The myth of generic DPA... and the magic of learning

Carolyn Whitnall[1], Elisabeth Oswald[1], and François-Xavier Standaert[2]

[1] University of Bristol, Department of Computer Science,
Merchant Venturers Building, Woodland Road, BS8 1UB, Bristol, UK.
{carolyn.whitnall, elisabeth.oswald}@bris.ac.uk
[2] Université catholique de Louvain, UCL Crypto Group
Place du Levant 3, B-1348 Louvain-la-Neuve, Belgium.
fstandae@uclouvain.be

Abstract. A prominent strand within the side-channel literature is the quest for *generic* attack strategies: methods by which data-dependent leakage measurements can be successfully analysed with 'no' a priori knowledge about the leakage characteristics. In this paper, we introduce a well-reasoned definition for what it means to have 'no' a priori insight (that is, to use a power model which approximates the device—up to nominality—by the equivalence classes associated with the *target* function), and use this to define generic DPA attacks. With these definitions we are able to clarify precise conditions (on the target function) under which generic attacks succeed. Doing so, we expose a rather limited range of vulnerable target functions, so that the 'myth' of the potential power of generic DPA is somewhat dispelled. We then shift focus onto linear regression-based attacks: linear regression can operate generically (as we explain) by 'fitting' the leakage measurements (differently for different key hypotheses) to a *full basis* of polynomial terms in the targeted bits.

Quite surprisingly, we show that even when the target function is not susceptible to generic DPA, applying some additional, non device-specific intuition to the different hypothesis-dependent models can in fact reveal the key. This intuition amounts to the idea that the estimated model coefficients associated with the correct key hypothesis ought to be 'more orderly', in some sense, provided the target function is sufficiently nonlinear (as is typically the case for cryptographic S-Boxes used in block ciphers).

To leverage this in a practical way we apply a model building technique called stepwise regression. Thus by 'emulating' a generic technique we can 'magically' produce successful attacks even when generic attacks applied in a conventional mode would fail.

1 Introduction

Ever since Kocher et al. showed that differential power analysis (DPA) could be successful even with very little information about the target implementation [18], the quest to find 'generic' methods for DPA attacks has been an ongoing endeavour for the research community. Informally, the appeal of a 'generic' method is the ability to successfully recover secret information even in the total absence of knowledge about the attacked device's data-dependent power consumption. Recent suggestions include mutual information analysis (MIA) using an identity power model [14], distinguishers based on the Kolmogorov-Smirnov (KS) two-sample test statistic [31,35] and the Cramer-von-Mises test [31], linear regression (LR)-based methods which can be seen as a sort of on-the-fly profiling [10,25], and an innovative approach using copulas [32].

However, all existing proposals share a common shortfall when applied to injective target functions: in order to distinguish between hypotheses the attacker must, after all, have some meaningful piece of knowledge by which to partition the measurements (in the case of MIA and KS-based DPA)[3] or select the appropriate set of covariates (in the case of LR-based DPA) [32]. Such attacks can no longer be considered 'generic', a description which is earned primarily by virtue of the non-reliance on *a priori knowledge* rather than the statistical methodology.

[3] A work-around frequently suggested in the literature is to group the predicted intermediate values according to their 7 least significant bits (sometimes called the 7LSB model). In fact does not circumvent this requirement: it does, sometimes, produce successful outcomes, but only when the leakage function is such that increasing noise distorts the trace measurements *towards* the model, rendering the seemingly arbitrary partition 'meaningful' after all [34].

The focus (in the above-mentioned papers) on defining universally-applicable distinguishers indicates a confusion about the role of the distinguisher and that of the power model in what has so far been only informally defined as 'generic' DPA. It also raises the fundamental question of whether *truly* 'generic' tools exist at all.

Establishing whether or not truly generic DPA attacks exist has fundamental consequences for the process of cryptographic device evaluation. The presence of generic attacks would imply that any device could potentially be attacked without any information about its internal functioning or leakage characteristics. Consequently, attacks based on profiling would only be 'better' in terms of efficiency (number of power traces needed)—not in terms of applicability. The absence of generic attacks would imply that there exist devices (leakage characteristics) which can only be evaluated soundly by performing profiled attacks—a practice which is not commonly undertaken at present. In the following, we tackle this important question in the practically relevant context of first-order DPA similar to the one investigated, e.g. in [10,14,18,25,31,35]. That is, we assume side-channel information such that the mean of the leakage distributions is key-dependent.

1.1 Our Contribution

We approach the problem of generic attacks by revisiting Stevens' 'levels of measurement' [29] to develop a theory of power models. Using this theory we first define what constitutes a *generic power model*. Next we investigate how different types of DPA distinguishers 'fit' to power models and derive a definition for a *generic-compatible distinguisher* accordingly. We then call the pairing of a *generic-compatible distinguisher* with the *generic power model* a first-order *generic strategy*. These definitions provide a basis for making conclusive general statements about generic DPA. We show that the noninjectivity of the target function is a prerequisite for any first-order generic strategy to succeed: in other words we prove the absence of a universally-applicable generic distinguisher in the context of first-order DPA! Note that generic higher-order DPA (i.e. exploiting the higher-moments of a leakage distribution) can only be more difficult to design. So the first-order context that we investigate in this work can be viewed as the most general to prove such negative conclusions regarding the possibility of generic attacks. We push our study further by observing that noninjectivity alone is still not sufficient for generic success, and we hence investigate additional requirements on the target function. It is already known that there is an inverse relationship between performance against certain S-Box criteria and susceptibility to DPA [22]; in this paper we demonstrate a sufficient condition for first-order generic success which is promoted (though not inevitably produced) by the design goal of *differential uniformity* [4].

Having thus dispelled the 'mythical' possibility for a universally-applicable generic distinguisher, but having gained more understanding, we make a fresh attempt at a distinguisher that operates without any *a-priori* assumptions about the leakage, but which produces results requiring only some non *device-specific* intuition for interpretation. A LR-based distinguisher is an ideal candidate; whilst it *can* operate from the starting point of a full basis of polynomial terms in the targeted bits (and thus, we show, qualifies as a generic-compatible distinguisher) the result will not only give a key ranking but also corresponding model estimates (for each key hypothesis). Of course, in case of injective target functions, keys will be indistinguishable in the ranking, which fact is consistent with the first half of our work and is already well-established in the literature [32]. However the resulting model estimates can also be used in the interpretation of results. The non device-specific intuition mentioned before in essence relates to the supposition that even if the leakage function is a high-degree polynomial of the target bits, it will remain 'simpler' or 'more orderly' for the correct key hypothesis than for an incorrect key hypothesis. This assumption can be exploited by the technique of *stepwise regression*, thus 'magically' producing successful outcomes even in scenarios where strictly generic strategies are known to fail! In practice, we observe that stepwise regression is best exploited in situations where the leakage function remains 'sufficiently simple' compared to the target function, which we argue is frequently observed in modern computing devices. Among others and for the first time, we perform successful key recoveries in the pathological case of Hamming weight leakages with an injective target function and generic power model, for which previ-

ously introduced generic-compatible distinguishers systematically failed. The experiments in Section 4.5 further describe successful attacks for a range of increasingly complex leakage functions.

As our investigations focus on the *feasibility* of generic success (irrespective of efficiency), we quantify the *asymptotic* capabilities of LR-based distinguishers in a range of practically meaningful first-order scenarios, targeting functions relating to DES, AES, and PRESENT block ciphers.

2 Preliminaries

2.1 Differential power analysis

We consider a 'standard DPA attack' scenario as defined in [20], and briefly explain the underlying idea as well as introduce the necessary terminology here. We assume that the power consumption T of a cryptographic device depends on some internal value (or state) $F_{k^*}(X)$ which we call the *target*: a function $F_{k^*} : \mathcal{X} \to \mathcal{Z}$ of some part of the known plaintext—a random variable $X \overset{R}{\in} \mathcal{X}$—which is dependent on some part of the secret key $k^* \in \mathcal{K}$. Consequently, we have that $T = L \circ F_{k^*}(X) + \varepsilon$, where $L : \mathcal{Z} \to \mathbb{R}$ describes the data-dependent component and ε comprises the remaining power consumption which can be modeled as independent random noise (this simplifying assumption is common in the literature—see, again, [20]). The attacker has N power measurements corresponding to encryptions of N known plaintexts $x_i \in \mathcal{X}$, $i = 1, \ldots, N$ and wishes to recover the secret key k^*. The attacker can accurately compute the internal values as they would be under each key hypothesis $\{F_k(x_i)\}_{i=1}^N$, $k \in \mathcal{K}$ and uses whatever information he possesses about the true leakage function L to construct a prediction model $M : \mathcal{Z} \to \mathcal{M}$.

DPA is motivated by the intuition that the model predictions under the correct key hypothesis should give more information about the true trace measurements than the model predictions under an incorrect key hypothesis. A distinguisher D is some function which can be applied to the measurements and the hypothesis-dependent predictions in order to quantify the correspondence between them. For a given such comparison statistic, D, the *theoretic* attack vector is $\mathbf{D} = \{D(L \circ F_{k^*}(X) + \varepsilon, M \circ F_k(X))\}_{k \in \mathcal{K}}$, and the *estimated* vector from a practical instantiation of the attack is $\hat{\mathbf{D}}_N = \{\hat{D}_N(L \circ F_{k^*}(\mathbf{x}) + \mathbf{e}, M \circ F_k(\mathbf{x}))\}_{k \in \mathcal{K}}$ (where $\mathbf{x} = \{x_i\}_{i=1}^N$ are the known inputs and $\mathbf{e} = \{e_i\}_{i=1}^N$ is the observed noise). Then the attack is *o-th order theoretically successful* if $\#\{k \in \mathcal{K} : \mathbf{D}[k^*] \leq \mathbf{D}[k]\} \leq o$ and *o-th order successful* if $\#\{k \in \mathcal{K} : \hat{\mathbf{D}}_N[k^*] \leq \hat{\mathbf{D}}_N[k]\} \leq o$.[4]

Definition 1 *A practical instantiation of a* standard univariate DPA attack *computes, given a set of power traces* $\mathbf{T}$*, a prediction model* M*, a set of inputs* $\mathbf{X}$*, and a comparison statistic* D*, the distinguishing vector* $\hat{\mathbf{D}}_N = \{\hat{D}_N(L \circ F_{k^*}(\mathbf{x}) + \mathbf{e}, M \circ F_k(\mathbf{x}))\}_{k \in \mathcal{K}}$*. A practical instantiation is said to be* o-th order successful *if* $\#\{k \in \mathcal{K} : \hat{\mathbf{D}}_N[k^*] \leq \hat{\mathbf{D}}_N[k]\} \leq o$*.*

2.2 Measuring DPA outcomes

Metrics to compare the *efficiency* of DPA attacks include the *(o-th order) success rate* and the *guessing entropy* of [28]—defined respectively as the probability of o-th order success and the expected number of key hypotheses remaining to test after a practical attack on a given number of traces. However, in the evaluation of generic strategies, the question of asymptotic feasibility takes precedence over that of efficiency. By the law of large numbers $\frac{1}{N}\sum_{i=1}^N L \circ F_{k^*}(x) + e_i \to L \circ F_{k^*}(x)$ as $N \to \infty$ (as long as the samples are independent and

[4] Note that standard DPA attacks do not include collision-based attacks [26], which exploit information from several leakage points per observation, and do not require a power model at all.

identically distributed). We can therefore discuss feasibility from the perspective of the *ideal* distinguishing vector $\mathbf{D}_{IDEAL} = \{D(L \circ F_{k^*}(X), M \circ F_k(X))\}_{k \in \mathcal{K}}$, noting that this no longer depends on the noise but only on the hypothesis-dependent power models relative to the true leakage. Indeed, averaging the trace measurements conditioned on the inputs is a popular pre-processing step in practice as it strips out irrelevant variance and reduces the dimensionality of the computations (see, for example, [1]); it is a sound approach as long as the side-channel information to be exploited originates in differences between the mean values of the leakage distributions, which *is* the case in our standard DPA scenario.

For the purposes of evaluating generic strategies, we will focus on first-order asymptotic success, as captured by the (ideal) nearest-rival distinguishing margin (see [33,34]): $NRMarg(\mathbf{D}_{IDEAL}) = \mathbf{D}_{IDEAL}[k^*] - \max\{\mathbf{D}_{IDEAL}[k] | k \neq k^*\}$.

2.3 Boolean vectorial functions

We are often interested in the special case that the key-indexed functions F_k can be expressed as $F_k(X) = F(k * X)$ where $F : \mathbb{F}_2^n \to \mathbb{F}_2^m$ is an $(n\text{-}m)$ Boolean vectorial function and $*$ denotes the *key combining* operator (for example, XOR). This case is particularly relevant to the study of block ciphers, which in general can be decomposed into combinations of such functions. Those components which are especially designed to introduce confusion into the system are known as S-Boxes (where 'S' stands for 'substitution').

Certain algebraic properties of such functions are known to be particularly important to the *cryptanalytic* robustness of a cipher system. We (very) briefly introduce such concepts as will play a role in our later analysis; for a good basic introduction see [16] or, for a more comprehensive explanation, [7,8].

F is *affine* if it can be expressed as a linear map followed by a translation—that is, if there exists a matrix $M \in \mathbb{F}_2^{m \times n}$ and a vector $v \in \mathbb{F}_2^m$ such that $F(x) = Mx \oplus v$. Such functions are known to be cryptanalytically vulnerable, and one of the aims in designing an S-Box is that any nonzero linear combination of the coordinate functions of F be *as far away as possible* from the set of all Boolean affine functions, in order to defend against *linear cryptanalysis* [21]. Thus the *nonlinearity* is defined: $N_F = \min_{u \in \mathbb{F}_2^n, v \in \mathbb{F}_2^m \setminus \{\mathbf{0}\}} \sum_{x \in \mathbb{F}_2^n} u \cdot x \oplus v \cdot F(x)$.

F is *balanced* if the preimages in F of all singleton subsets of $\mathbb{F}_2^m$ are uniformly sized: that is, $\forall y \in \mathbb{F}_2^m$, $\#\{x \in \mathbb{F}_2^n | F(x) = y\} = 2^{n-m}$. This property applies to many functions used in block ciphers, particularly S-Boxes [36] where any bias on the unobserved inputs is extremely undesirable.

The key property providing resistance to *differential cryptanalysis* as introduced by Biham and Shamir [4] is the notion of *differential uniformity*. This means that the derivatives of F with respect to $a \in \mathbb{F}_2^n$, defined as $D_a F(x) = F(x) \oplus F(x \oplus a)$, must be *as uniform as possible*. If there exists a vector $a \in \mathbb{F}_2^n$ such that $D_a F(x)$ is constant over $\mathbb{F}_2^n$ then a is called a *linear structure* of F and (as per [11]) can be exploited by a cryptanalyst. The space $\{a \in \mathbb{F}_2^n | D_a F = cst\}$ is the *linear space* of F and the larger it is, the more susceptible to cryptanalysis.

3 The 'myth' of generic DPA

What does it mean for an attack to be 'generic'? The discussion in the literature has focused on appropriating, as distinguishers, statistics which 'require few distributional assumptions'—trawling the statistical literature for nonparametric, distribution-comparing procedures such as the Kullback-Leibler divergence (a.k.a. Mutual Information Analysis) [14], the Kolmogorov-Smirnov [31,35] and Cramer-von-Mises [31] tests, and copulas [32]. However, the emphasis on finding 'distribution-free' statistics for use as distinguishers somewhat distracts from the essential defining feature of generic DPA which is that *no assumptions have been made about the device*

leakage. Clearly, the (fairly common) practice of combining such distinguishers with an informed prior model (i.e. *some* prior information) does not produce a generic attack. We hence also need to pay attention to what 'type' of power model is used in conjunction with a distinguisher; evidently we are seeking to define what constitutes a 'generic' power model.

This section concentrates first on the different types of model used in DPA attacks in Sect. 3.1, and which distinguishers are suitable in each instance. Based on this we subsequently give definitions for what is a generic power model, a generic-compatible distinguisher, and a generic DPA strategy in Sect. 3.2. These definitions then form the basis for a number of propositions that clarify the cases in which any generic strategy is bound to fail (we spell out necessary conditions for success and discuss further the feasibility of generic DPA).

3.1 Delineating leakage assumptions

Firstly we must distinguish between assumptions about the *data-dependent* leakage, as captured by the power model, and assumptions about the *distribution of the noise*—which in most cases play a less visible role, but can affect how accurately or efficiently certain statistics may be estimated. Figure 1 visualises this two-dimensional continuum, and indicates the suitability of some popular distinguishers as assumptions vary.

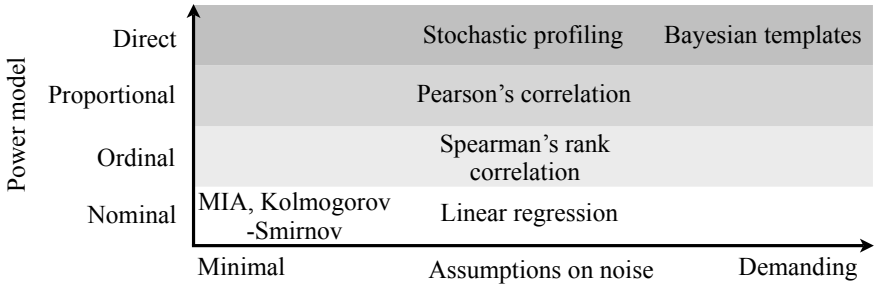

Fig. 1. Types of leakage model and the assumptions required by common distinguishers.

Assumptions about the noise range from *fully characterised distributions* as exploited (e.g.) by Bayesian template attacks, down to *no knowledge whatsoever*, when the robustness of nonparametric statistics such as mutual information and the Kolmogorov-Smirnov test may come in handy. Fortunately, the often reasonable assumption of approximate normality opens up a broad range of (semi-)parametric options, which are to be preferred as they are inherently less costly to estimate.

We now consider the *nature* of the power model, with which this paper is primarily concerned. Previous studies have talked about 'good' power models, in an arbitrary sense, and most have missed the very material distinction between different *levels* of model. As hinted towards in [2,12], the widely-accepted 'levels of measurement' laid out in [29] present a natural framework for delineation. It is important to understand the appropriate (type-specific) notion of accuracy for a given model, and to select a compatible distinguisher; that is, one which (implicitly) interprets the model according to the correct type.

The type of power model exploited by profiled attacks (e.g. Bayesian templates [9] and stochastic profiling [25]) amounts to a *direct approximation* of the actual power consumed by processing the data, in contribution to the overall consumption. This requirement is the most demanding possible, expressed as $M \approx L$ (c.f. the 'ratio scale' of [29]). The outcome of an attack will depend on how accurately the templates approximate the actual data-dependent consumption (as well as the noise distribution). The error sum-of-squares is a natural way of quantifying the appropriate notion of accuracy.

Less demanding is the requirement that the attacker has a power model which is a good approximation for L *up to proportionality*: $M \approx \alpha L$ (c.f. the 'interval scale' of [29]). Pearson's correlation coefficient provides a natural way to quantify accuracy and can be directly adapted for use as distinguisher [5] (a popular strategy since, as a simple, moment-based statistic, it can usually be estimated very efficiently with respect to the number of trace measurements required).

Less demanding again is the requirement that M approximates L *up to ordinality*: $\{z|M(z) < M(z')\} \approx \{z|L(z) < L(z')\}\ \forall z \in \mathcal{Z}$ (c.f. the 'ordinal scale' of [29]). Such a model could be exploited via a variant of correlation DPA using Spearman's rank correlation coefficient, as proposed in [2]. And, again, the accuracy of the model can be quantified via the rank correlation itself.

The least demanding requirement to place on a model is that it approximates the leakage function *up to nominality* only: $\{z|M(z) = M(z')\} \approx \{z|L(z) = L(z')\}\ \forall z \in \mathcal{Z}$ (c.f. the 'nominal scale' of [29]). As ever, such a model must be paired with a statistic which interprets the values appropriately: that is to say, as arbitrary labels only. In fact, these correspond to the 'partition-based' distinguishers of [27]. Typical examples include statistics which are used to compare arbitrary distributions, such as mutual information [14] and the Kolmogorov-Smirnov test statistic [31,35].

Appropriate notions of accuracy for a nominal model are drawn from classification theory. *Precision* is the probability that items grouped according to the model really do belong together, whilst *recall* is the probability that items which belong together are identified as such (see, e.g. [19]).[5]

$$\begin{aligned} Precision(M) &= \mathbb{P}(L(z) = L(z')|M(z) = M(z')), \\ Recall(M) &= \mathbb{P}(M(z) = M(z')|L(z) = L(z')). \end{aligned}$$

3.2 Defining 'genericity'

We are now in a position to discuss the generic power model: what, in practice, does it mean to make *no* assumptions about the data-dependent leakage? Essentially, that we do no more than to assign a distinct label to each value in the range of the target function. These labels can be seen to correspond to the key-dependent equivalence classes produced by the preimages of F_k: $[x]_k = F_k^{-1}[F_k(x)]\ \forall x \in \mathcal{X}$.

Definition 2 *The* generic power model *associated with key hypothesis $k \in \mathcal{K}$ is the nominal mapping to the equivalence classes induced by the key-hypothesised target function F_k.*

The 'identity' power model emphasised in previous literature is fine for this purpose as long as it is understood that the identity mapping is simply a convenient labelling system and should be interpreted *nominally* only. It is immediately clear that the *generic-compatible distinguishers* are precisely those (described in Section 3.1 above) which interpret hypothesis-dependent predictions as an approximation up to nominality of the data-dependent leakage.

Definition 3 *A distinguisher is* generic-compatible *if it is built from a statistic which operates on nominal scale measurements.*

[5] The classification theory literature more frequently states these definitions in terms of ratios of counts—practically convenient but less directly translatable across contexts. See [15] for a more explicit probabilistic interpretation; though in our case we are, of course, averaging over multiple classes.

This provides valuable clarification on previous work such as [3], which demonstrated successful attacks against Hamming weight leakage using *correlation* DPA with an 'identity' power model. The authors rightly remarked that this was possible precisely because, over $\mathbb{F}_2^4$, the identity is sufficiently accurate as a *proportional* approximation of the Hamming weight to produce a successful correlation attack. Far from operating generically, the identity mapping in such a strategy is interpreted as an *interval scale* model—not a perfect approximation but *adequate* in the specific case that L can be well-approximated by the *Hamming weight*. And even in this restricted case it is not, of course, invariant to permutation of the 'identity' labels.

Definitions 2 and 3 together give rise to a natural notion of a 'generic strategy':

Definition 4 *A* generic strategy *performs a standard univariate DPA attack using the generic power model paired with a generic-compatible distinguisher.*

However, as previous work on 'partition-based' distinguishers (separately, e.g. [14,32,35], and collectively [27]) has consistently noted, not all (indeed, not many) scenarios are suited to a generic strategy.

3.3 Conditions for a generic strategy to succeed

All distinguishers operate by identifying the key hypotheses producing the most accurate model predictions for the actual measurements, according to the appropriate notion of accuracy for the model type (some are able to perform this comparison more effectively or from fewer trace measurements). In the generic setting each key hypothesis $k \in \mathcal{K}$ gives rise to a model M_k s.t. $M_k^{-1}[z] = F_k^{-1}[z]\ \forall z \in F_k(\mathcal{X})$. Key-recovery will be possible *precisely* when the model produced by the correct key hypothesis is a better nominal approximation for the true leakage than those produced by any of the alternatives. We can therefore explore the conditions necessary for a successful attack—independently of any particular distinguisher—by reasoning directly about the accuracy of F_{k^*} and F_k, $\forall k \in \mathcal{K} \setminus \{k^*\}$ as nominal approximations for $L \circ F_{k^*}$. Recall the precision and recall measures introduced in Section 3.1 (with $\mathbb{E}$ to denote expectation):

$$\begin{aligned}
Precision(M_k) &= \mathbb{P}(L \circ F_{k^*}(x) = L \circ F_{k^*}(x') | F_k(x) = F_k(x')) \\
&= \mathbb{E}_{x \in \mathcal{X}} \left[\frac{\#F_{k^*}^{-1}[L^{-1}[L \circ F_{k^*}(x)]] \cap F_k^{-1}[F_k(x)]}{\#F_k^{-1}[F_k(x)]} \right] \\
Recall(M_k) &= \mathbb{P}(F_k(x) = F_k(x') | L \circ F_{k^*}(x) = L \circ F_{k^*}(x')) \\
&= \mathbb{E}_{x \in \mathcal{X}} \left[\frac{\#F_{k^*}^{-1}[L^{-1}[L \circ F_{k^*}(x)]] \cap F_k^{-1}[F_k(x)]}{\#F_{k^*}^{-1}[L^{-1}[L \circ F_{k^*}(x)]]} \right]
\end{aligned}$$

Trivially, the precision of the generic model under the correct hypothesis is always maximal (the leakage preimage must contain the function preimage). By contrast, the recall depends additionally on the true leakage function, so that even under the correct hypothesis we do not get perfect recall unless it happens that L is also injective. The ability of a strategy to reject an *incorrect* alternative requires the corresponding model to be of inferior quality; whether this is so depends on features of F_k and L. An immediate and quite restrictive pre-requisite arises from the inherent nature of the generic power model:

Proposition 1. *No generic strategy is able to distinguish the correct key k^* from an alternative hypothesis k if F_{k^*} and F_k are injective.*

Proof. If F_{k^*}, F_k are injective then $\forall x \in \mathcal{X}$, $F_k^{-1}[F_k(x)] = F_{k^*}^{-1}[F_{k^*}(x)] = \{x\}$. Each hypothesis produces models of equivalent nominal accuracy—no generic-compatible distinguisher can be expected to separate the candidates.

Indeed, all of the known generic-compatible distinguishers, from the seminal CHES '08 paper on MIA [14] to the recent copula-based method presented at Crypto '11 [32], have individually been shown to fail whenever the *composition* of the target function and the power model is injective; the same observation was made for the entire class of 'partition-based' distinguishers described in [27]. The authors duly noted that some restriction was required on the power model in order for these distinguishers to operate against an injective target, but left as an open question the *existence* (or demonstrable non-existence) of an as-yet undiscovered method which would somehow circumvent this requirement. Demonstrating that the limitation is attributable directly to the generic power model rules out this possibility.

Noninjectivity is therefore a necessary condition, but not, as we next establish, a sufficient one. In the general case it is rather difficult to formulate useful, concrete observations so we will henceforth narrow down to the restricted but highly relevant case that F is a *balanced* (n-m) function and k is introduced by key addition (as described in Section 2.3). It then becomes fairly straightforward to draw out such function characteristics as will obstruct a generic strategy.

Proposition 2. *Suppose F is a balanced, non-injective (n-m) function, with k introduced by (XOR) key addition, i.e. $F_k(x) = F(x \oplus k)$. Then:*

(a) If F is affine then no generic strategy is able to distinguish the correct key k^* from any *$k \in \mathcal{K} \setminus \{k^*\}$.*
(b) If $a \in \mathbb{F}_2^n$ is a linear structure of F then no generic strategy is able to distinguish between k^* and $k^* \oplus a$.
(c) If, for some $a \in \mathbb{F}_2^n$ we have that $D_aF(x)$ depends on x only via *$F(x)$, then no generic strategy is able to distinguish between k^* and $k^* \oplus a$.*

The proof of Proposition 2 can be found in Appendix A. Part (a) arises from the fact that all key hypotheses produce indistinguishably 'good' models for the leakage; the distinguishing vector produced by such an attack would be flat and maximal across all hypotheses.

The implication of 2(b) is that $k^* \oplus a$ cannot be rejected if the derivative of F with respect to a is *constant* over the *domain* of F, i.e $\#D_aF(\mathbb{F}_2^n) = 1$. In such a case we would expect a practical attack to exhibit a *ghost peak* at $k^* \oplus a$ [5]; [22], notes a corresponding phenomenon for correlation DPA.

Part (c) can be otherwise expressed as the fact that $k^* \oplus a$ cannot be rejected if the derivative of F with respect to a is *constant* over *each singleton preimage* of F, i.e. $\#D_aF(F^{-1}[F(x)]) = 1 \ \forall x \in \mathbb{F}_2^n$. We have actually observed this property in the fourth DES S-Box, for the key-offset $a = 47_{(10)} = 101111_{(2)}$: in consequence, $k^* \oplus 47$ produces a 'ghost peak' in the distinguishing vector, with a nonetheless substantial margin between these *two* and the remaining hypotheses—a good example of an attack scenario with a low first-order, but high second-order, success rate [28]. Our observation is consistent with, and further illuminates, past works such as [6] which recognised the unusual operation of DPA distinguishers confronted with this particular S-Box/offset combination.

Thus emerges a minimal requirement for k^* to be distinguished from k:

Proposition 3. *Suppose F is a balanced, noninjective n-m function, with k introduced by (XOR) key-addition. A necessary condition for a generic strategy to distinguish k^* from k is: $\exists x \in \mathbb{F}_2^n$ such that $\#D_{k^* \oplus k}F(F^{-1}[F(x)]) \neq 1$. If L is injective then this becomes a sufficient condition.*

This is informally expressed as the requirement that there is at least one (singleton) preimage over which the derivative with respect to $k^* \oplus k$ is *not* constant. The proof follows from our reasoning in support of Proposition 2 and can be found in Appendix A along with a toy example to demonstrate that we can no longer claim sufficiency if L is noninjective.

It is an explicit design goal that S-Boxes should have high differential entropy [4]; affine functions or functions with non-null linear spaces represent the extreme in terms of cryptanalytic vulnerability. The pursuit of this criteria does not guarantee the minimal condition above, as even a perfectly balanced derivative could be so arranged as to be constant over the singleton preimages (which are of cardinality 2^{n-m} since F is also balanced). However, it would certainly seem to increase the chance that the condition be met for a given key-offset, as the more finely D_aF partitions $\mathbb{F}_2^n$, the fewer the possible *refinements* into 2^m (balanced) parts. Therefore, among the (already restricted) class of noninjective S-Boxes we would expect ghost peaks and indistinguishable keys to be a rarity—even more so as the size of the S-Box increases.

4 The 'magic' of learning

The first part of this paper makes it clear that any generic-compatible distinguisher which returns only some 'classification accuracy' for the key hypotheses will fail against injective target functions. Intuitively, then, we looking for distinguishers that compute something 'more': linear regression-based methods are an immediate candidate because they *can* be used with a generic power model (equivalent, as we show, to providing a full basis of polynomial terms in the targeted bits)—in which case the distinguishing vector of goodness-of-fit values will be unable to discriminate between key hypotheses if the target is injective—but in the process they also return the estimated power model coefficients. Examining these one can readily observe that the fitted models for different key hypotheses are different: we will show that the application of some simple, non device-specific intuition to these models can in fact reveal the correct key hypothesis. This process can be automated straightforwardly by using LR in a stepwise mode.

We begin by introducing (standard) LR-based DPA, explaining the mechanism by which it distinguishes the correct key, and demonstrating that it is among the class of generic-compatible distinguishers. We then present the 'generic-emulating' stepwise regression-inspired variant which exploits the non device-specific intuition to successfully attack injective targets even with 'no' (other) prior knowledge. We finally demonstrate the (asymptotic) effectiveness of these distinguishers against well-known (injective and noninjective) S-Boxes, as the level of prior knowledge available varies from 'complete' to 'none'.

4.1 Introduction to linear regression-based DPA

The motivation for a linear regression-based approach begins with the observation that $L : \mathbb{F}_2^m \to \mathbb{R}$ can be viewed as a pseudo-Boolean vectorial function with a unique expression in numerical normal form [7]. That is to say, there exists coefficients $\alpha_u \in \mathbb{R}$ such that $L(z) = \sum_{u \in \mathbb{F}_2^m} \alpha_u z^u$, $\forall z \in \mathbb{F}_2^m$ (z^u denotes the monomial $\prod_{i=1}^m z_i^{u_i}$ where z_i is the i^{th} bit of z). Finding those coefficients amounts to finding a power model for L in polynomial function of the coordinate functions of F.

'Stochastic attacks' [25], proposed as an alternative to Bayesian template attacks, used linear regression in a profiling stage to estimate those coefficients from data collected on a controlled device. The authors also observed—though it was only later demonstrated [10]—that the procedure could be adapted to non-profiled key-recovery, in which the true leakage function is estimated 'on-the-fly' and recovered synchronously with the true key.

Appendix B provides background on linear regression; in short, the LR-based attack uses ordinary least squares to estimate, for each $k \in \mathcal{K}$, the parameters of the model $L_{k^*}(X) + \varepsilon = \alpha_{\mathbf{0}} + \sum_{u \in \mathcal{U}} F_k(X)^u \alpha_u$ where $\mathcal{U} \subseteq \mathbb{F}_2^m \setminus \{\mathbf{0}\}$. The distinguishing vector comprises the R^2 measure of fit from each of these models: $D_{\mathrm{LR}}(k) = \rho(L_{k^*}(X) + \varepsilon, \hat{\alpha}_{k,0} + \sum_{u \in \mathcal{U}} F_k(X)^u \hat{\alpha}_{k,u})^2$ (where ρ denotes Pearson's correlation coefficient). It can be viewed as a generalisation of correlation DPA, where the power model M is known *a priori*: $D_\rho(k) = \rho(L_{k^*}(X) + \varepsilon, M \circ F_k(X))$. In each case, the value of k which produces the *largest* distinguisher value is selected as the key guess.

4.2 Linear regression is generic-compatible

In the way the distinguisher is most naturally presented, the attacker's prior knowledge is contained within $\mathcal{U}$ and it is not immediately obvious exactly what *is* the power model, or where it fits alongside the various types presented in Section 3.1. In fact, each $u \in \mathcal{U}$ could be seen to represent a *separate* power model which divides the traces into two nominal classes: $\{x \in \mathbb{F}_2^n | F_k(x)^u = 1\}$ and $\{x \in \mathbb{F}_2^n | F_k(x)^u = 0\}$.[6] Intuitively, as long as the power consumption really *does* differ systematically according to the bit-interaction term represented by u, then this 'approximation' has low precision but high recall under the correct key hypothesis, and loses accuracy under an incorrect hypothesis as long as the function F is such that changes to the input produce nonuniform changes to the output. Such, in fact, is the mechanism by which the original difference-of-means distinguisher of [18] operates!

So the linear regression distinguisher could be viewed as an extension of difference-of-means DPA—a means of exploiting *multiple* (overlapping) nominal approximations, each of low precision (and therefore weak as standalone models) but in conjunction providing a refined description of the leakage. Note that any $u \in \mathcal{U}$ which does *not* induce a 'meaningful' partition will detract from the overall distinguishing power—hence the motivation to use any prior knowledge available to refine $\mathcal{U}$.

Intuitively, the generic instantiation should correspond to $\mathcal{U} = \mathbb{F}_2^m \setminus \{\mathbf{0}\}$, which is equivalent to imposing no restrictions on the form of the leakage. But our previous reasoning about the operation of generic strategies supposed a *single* power model (F_k, interpreted nominally) and it is hard to see how we might begin to reason about the impact of multiple power models. Fortunately, in the $\mathcal{U} = \mathbb{F}_2^m \setminus \{\mathbf{0}\}$ case *only*, the operation of the distinguisher *can* be re-framed in terms of the generic power model as defined above—implying that all of our prior reasoning applies.

Proposition 4. *The linear regression-based DPA attack with a full set of covariates $\mathcal{U} = \mathbb{F}_2^m \setminus \{\mathbf{0}\}$ constitutes a generic strategy.*

We sketch a proof as follows: If M_k is an arbitrary labeling on F_k, we can always map bijectively to $\mathbb{F}_2^m$ to acquire an arbitrary *permutation* of the function outputs $M'_k(x) = p \circ F_k(x)$. For each $u \in \mathbb{F}_2^m$, the associated monomial $M'_k(x)^u$ has a unique expression in *numerical* normal form $M'_k(x)^u = \sum_{v \in \mathbb{F}_2^m} b_v F_k(x)^v$, $b_v \in \mathbb{R}$ [7]. So the system of equations relating to an incorrect hypothesis k can be re-written in function of $F_k(x)$ by substituting in these expressions, expanding out and collecting up the terms. Note that we end up with different values of α_u, $u \in \mathbb{F}_2^m$ whenever we reparametrise in this way: but, crucially, the terms in the equation *collectively* explain the measured traces equally well—and it is in this sense that linear regression DPA is *invariant to re-labeling* and therefore can be discussed alongside other generic-compatible strategies (though it is not usually used in this way—particularly as restrictions on $\mathcal{U}$ contribute to efficiency gains in the estimation stage).

'Fixing' generic linear regression-based DPA for injective targets? By the above argument, we expect that linear regression-based DPA with $\mathcal{U} = \mathbb{F}_2^m \setminus \{\mathbf{0}\}$ to fail for injective targets. However, we want to discuss briefly how this failure manifests itself, as it gives a first intuition regarding how that linear regression might be adapted to work even for injective targets. We focus, for clarity, on the asymptotic case (see Section 4.5).

When the target function is injective, the data-dependent part of the power consumption can be expressed as a system of 2^n equations (in function of $F_k(x)$) with 2^n unknowns. Because this system is fully-determined and consistent under any key hypothesis it *always* has a perfect solution, so as to produce a flat distinguishing vector

[6] Note that the labeling is irrelevant since they are represented in the regression equation by dummy variables: the 1/0 assignment is arbitrary and will impact only the estimated coefficients, not the R^2.

of maximal R^2s. The attack can be made effective by introducing some prior information about the true form of the data dependent leakage; that is, by dropping known-to-be-redundant terms from the model equation for L so as to produce an *over*-determined system. By way of simple example, supposing bit contributions to be *independent* (a slight generalisation of a Hamming weight leakage assumption), equates to taking $\mathcal{U} = \bigcup_{i=0}^{m-1}\{2^i\}$. As long as the discarded terms include some which do not contribute to L but *do* contribute to $L \circ F_k \circ F_{k^*}^{-1}$ then k is distinguished from k^*. The point, though, is that this is no longer a 'generic' strategy as it relies on some minimal insights about the device. It also assumes that the prediction 'labels' *are* the function outputs $M_k = F_k$ (in order to correctly interpret $\mathcal{U} \subset \mathbb{F}_2^m \setminus \{\mathbf{0}\}$).

4.3 Exploiting non device-specific intuition

We have just argued in the previous section that if an attacker was able to make a meaningful restriction on the model terms, linear regression could be made to work—that is, to return R^2s which allow us to distinguish between key hypotheses (producing a successful, but non-generic attack). However we now return our attention to operating linear regression generically. In this case we still have that for $\mathcal{U} = \mathbb{F}_2^m \setminus \{\mathbf{0}\}$, the derived models give an equally good fit. However, they will only give the correct expression for L in function of the output bits when $k = k^*$: the rest of the time, they will give an expression for $L \circ F_k \circ F_{k^*}^{-1}$. So, in a sense, the per-key parameter estimates contain all the information required to indicate the secret key, if only the attacker knew how to recognise the correct expression for L.

Thus motivated, we examine the correct and incorrect expressions for L in the case that the target function is an injective S-Box (of size 8 bits in the case of AES, or 4 bits in the case of PRESENT) and that the true form of the leakage is the Hamming weight: $L(z) = \sum_{i=0}^{m} z^{2^i}$. Figure 2 shows the coefficients, in the polynomial expression for L, on the covariates as produced by the true key (in black) and on those as produced under an incorrect hypothesis (in grey). The high nonlinearity of the AES, PRESENT (and DES) S-Boxes ensure that, when viewed as a polynomial in $F_k(X)$ rather than $F_{k^*}(X)$, the leakage function L is also highly nonlinear in form.

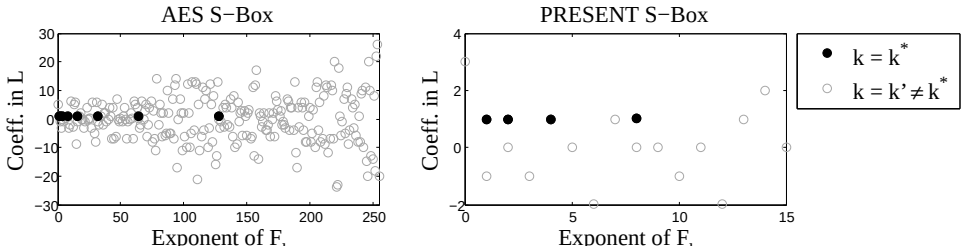

Fig. 2. Coefficients, in the polynomial expression for L, on the covariates as predicted under the correct and an incorrect key hypothesis.

An attacker confronted with such evidence would be justified in favouring hypothesis k^* over k: intuitively, it just seems more likely (especially given the known high nonlinearity of F) that the 'simpler' expression is the correct one. In order to exploit the information represented by the model estimates, we therefore need to trust this intuition (which implicitly also assumes that $M_k = F_k$—but since the target function is known to us this is completely reasonable). In that sense, we have taken a step away from the generic strategy—but since the intuition is not specific to any particular device it appears to be a very small step. That is, we just need to assume that the leakage function is 'sufficiently simple' compared to the target function. This is justified for a wide range of devices manufactured in CMOS technologies, including advanced 65-nanometer processes [24]. In fact, even for protected logic styles such as introduced by Tiri and Verbauwhede [30], it turns out that the

ensuring a complex (e.g. highly nonlinear) leakage function is a challenging task [23]. Besides, the results in Section 4.5 will also demonstrate that this 'simplicity constraint' on the leakage function can be quite relaxed.

Of course, comparing graphs is not ideal from a practical perspective, besides which the true leakage function may not always have so simple a form as to be visibly discernible: we would like to encapsulate the underlying reasoning into an automated and systematic procedure for testing hypotheses. In the next section we introduce a learning technique from data mining which uses our non device-specific intuition about 'what the leakage should look like' to produce, in a wide range of leakage scenarios, asymptotically successful key recovery against injective targets *even when provided with the full set of covariates* $\mathcal{U} = \mathbb{F}_2^m \setminus \{\mathbf{0}\}$. Such a strategy, whilst not *generic*, may reasonably be described as *generic-emulating*.

4.4 A stepwise regression-based distinguisher

Stepwise regression [17] is a model-building tool whereby potential explanatory variables are iteratively added and removed depending on whether they contribute sufficient explanatory power to meet certain threshold criteria (see Appendix C for full details). The resulting regression model should therefore exclude 'unimportant' terms whilst retaining all of the 'significant' terms. In the context of linear regression DPA this equates to testing each of the multiple binary models represented by $u \in \mathcal{U}$ separately (conditioned on current model) and then privileging those which appear most meaningful.

Under a correct key hypothesis, and *beginning with a full basis* $\mathcal{U} = \mathbb{F}_2^n \setminus \{\mathbf{0}\}$ we would expect to obtain a 'good' regression model which explains most of the variance in L, although with some minor terms absent if they do not meet our threshold criteria for statistical significance. The example depicted in Figure 2 above justifies the hope that the model produced under an incorrect hypothesis might be 'less good': with the explanatory power being so much more dispersed, the contribution of any individual term decreases. These small contributions are prejudiced against in the model building process (depending on the threshold criteria) but their *actual* contributions are real and so, therefore, is the loss in excluding them. If the aggregate loss is sufficient then the resulting R^2 will be enough reduced relative to the true key R^2 to distinguish between the two.

Figure 2 also reinforces the intuition (discussed in Section 3.3) that S-Box vulnerability increases with size: the extent to which explanatory power can be dispersed among the covariates under wrong-key reparametrisation is restricted by the number of covariates, $2^m - 1$. So whilst the potentially distinguishing feature can still be observed for the PRESENT S-Box it is less marked than in the case of AES, and we might expect key-recovery to be more difficult.

Of course, our chosen example scenario relates to a very simple leakage function. For the attack to be successful in more complex scenarios the target function F would have to introduce sufficient additional 'dispersion' of the explanatory power for the model quality to be affected. Moreover, the stepwise regression algorithm is sensitive to the threshold criteria set by the user and even to the order in which the variables are introduced. The optimal p-value at which to reject or accept candidate terms will vary depending on the properties of the target function and the size of its domain (i.e. the number of equations in the system). Careful tuning may be necessary before an attack becomes successful, and there is no prior guarantee that it can be so. Such decisions implicitly form part of the 'intuition' the attacker has about the true leakage. However, as previously mentioned this intuition is reasonable for a wide range of technologies. Furthermore, the next section demonstrates that—even in the 'uninformed' case that $\mathcal{U} = \mathbb{F}_2^n \setminus \{\mathbf{0}\}$—the stepwise approach *can* succeed (asymptotically) against injective targets, regardless of the degree of the leakage function, and can improve the efficiency of attacks against noninjective targets by widening the margins by which the true key is distinguished from the alternatives.

4.5 Asymptotic attack outcomes as knowledge on the power model varies

Figure 3 shows the distinguishing margins achieved (asymptotically) by linear regression-related attacks against AES, PRESENT and DES, for leakage polynomials of increasing degree, and for differing levels of prior knowledge on the data-dependent leakage, as follows:

- A perfectly characterised power model, which can be exploited straightforwardly in a correlation attack. (Corresponds to the line labelled 'Perfect model' in Figure 3).
- Knowledge of the degree d of the leakage polynomial, which can be exploited via LR and stepwise LR attacks using a restricted covariate set (or initial covariate set in the case of stepwise LR) i.e., $\mathcal{U} = \{u \in \mathbb{F}_2^n \setminus \{\mathbf{0}\} | HW(u) <= d\}$. (Labelled in Figure 3 as 'Max degree LR' and 'Max degree SW' respectively).
- No knowledge about the leakage polynomial, i.e., $\mathcal{U} = \mathbb{F}_2^n \setminus \{\mathbf{0}\}$, in which case we are interested in the performance of generic LR-based DPA and uniformed stepwise LR-based DPA. (Labelled in Figure 3 as 'Generic LR' and 'Uninformed SW' respectively).

As expected, the attacks exploiting a perfect characterisation perform best and are not penalised as the degree increases. LR with a known degree is decreasingly effective against injective targets (AES and PRESENT S-Boxes) as enlarging the set of included covariates increases the amount of variance that can be explained under a *rival* hypothesis—thus narrowing the distinguishing margins. It eventually fails altogether as it coincides with the generic distinguisher. However, as conjectured, stepwise linear regression *does* succeed, even against high degree leakage, and when combined with knowledge of the degree it is the best performing of all the linear regression-based attacks. The third panel nicely illustrates the robustness of a generic strategy against a noninjective target (namely, the first DES S-Box), whilst emphasising the efficiency gains available from both information on the degree and the use of stepwise regression.[7]

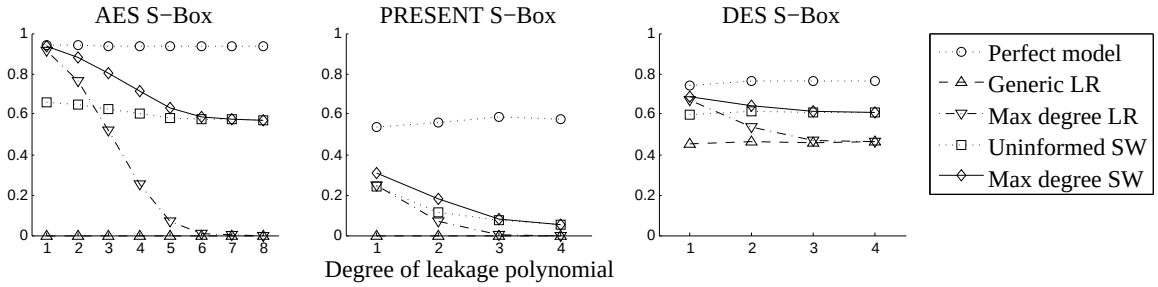

Fig. 3. Median distinguishing margins of attacks against AES, PRESENT and DES S-Boxes as the leakage degree increases (500 experiments with uniformly random coefficients between -10 and 10).

This final experiment allows us to come back on the requirement of 'sufficiently simple' leakage function expressed in Section 4.3. Namely, Figure 2 suggests that it is the comparative 'complexity' (in some sense) of $L \circ F_k \circ F_{k^*}^{-1}$ relative to L which is exploited by stepwise regression in order to recover k^*. We might naturally conjecture that the approach will eventually fail as long as L is sufficiently *nonlinear*, as there would no longer be any clue by which to distinguish the true key. Somewhat surprisingly, the example attacks above *do* succeed even when the leakage degree is maximal, so high polynomial degree is obviously not the relevant criteria to predict attack failure, at least asymptotically. Yet, we are quite able to artificially construct leakage functions against which stepwise regression does *not* succeed: for example, random permutations over $\{0, \ldots, 2^m - 1\}$. This would

[7] The asymptotic outcomes appear to be reliably consistent over the 500 repeated experiments—see Appendix D for more information.

indicate that the leakage functions for which the stepwise regression fails are the ones that achieve a high *cryptographic* nonlinearity when interpreted as functions over $\mathbb{F}_2^m$—at least when m is large enough. However, we leave as an open question the precise properties of L which will cause such failures.

5 Conclusion

We have differentiated between the *types* of prior knowledge which an attacker might have about the device leakage, and clarified the characteristics which appropriately define a power model as 'generic'. We have subsequently explored the distinguishers and scenarios which are compatible with such a construction. In particular, we have been able to conclusively demonstrate that noninjectivity of the target function is an inescapable prerequisite for such attacks to succeed, and moreover that this is not a sufficient condition but must be accompanied by certain other algebraic properties. The types of properties which *do* enable generic key recovery can be reasonably expected of many functions used in cryptography thanks to the particular aims of S-Box design criteria, but are not inevitably produced by the criteria. Our results show formally that the gap between generic strategies and so-called profiled attacks, taking advantage of prior knowledge on the leakage distribution, is not only a matter of efficiency (in terms of data and trace complexity to recover a key), but also of effectiveness (that is, there exist situations in which only the latter can succeed). Hence, although the gap can be small in certain scenarios (e.g. the first-order DPA in this paper), profiled attacks such as [9] are the only way to perform sound worst-case security evaluations [28]. We have then explained and developed the theory around linear regression-based first-order DPA and argued for its use as an efficient but flexible strategy. We have finally demonstrated how, by applying additional non device-specific intuition about the form of the leakage function, a data-mining technique known as 'stepwise regression' can produce (asymptotically) successful attacks even against *injective* targets, and can, moreover, enhance the trace efficiency of linear regression first-order DPA with or without prior knowledge. We describe this strategy as 'generic-emulating'. In this respect, an important scope of further research is to extend such generic-emulating strategies in the context of higher-order DPA and implementations protected with countermeasures. That is, can we also relax the limitations of generic-compatible distinguishers such as [13,32] in this more challenging context, exploiting non device-specific intuitions as in this work.

References

1. The DPA Contest. `http://www.dpacontest.org/`.
2. L. Batina, B. Gierlichs, and K. Lemke-Rust. Comparative Evaluation of Rank Correlation Based DPA on an AES Prototype Chip. In T.-C. Wu, C.-L. Lei, V. Rijmen, and D.-T. Lee, editors, *Information Security*, volume 5222 of *LNCS*, pages 341–354. Springer Berlin / Heidelberg, 2008.
3. L. Batina, B. Gierlichs, E. Prouff, M. Rivain, F.-X. Standaert, and N. Veyrat-Charvillon. Mutual Information Analysis: A Comprehensive Study. *Journal of Cryptology*, 24:269–291, 2011.
4. E. Biham and A. Shamir. Differential Cryptanalysis of DES-like Cryptosystems. In *Proceedings of the 10th Annual International Cryptology Conference on Advances in Cryptology*, CRYPTO '90, pages 2–21, London, UK, UK, 1991. Springer-Verlag.
5. E. Brier, C. Clavier, and F. Olivier. Correlation Power Analysis with a Leakage Model. In M. Joye and J.-J. Quisquater, editors, *Proceedings of CHES 2004*, volume 3156 of *LNCS*, pages 135–152. Springer Berlin / Heidelberg, 2004.
6. C. Canovas and J. Clediere. What Do S-boxes Say in Differential Side Channel Attacks? Cryptology ePrint Archive, Report 2005/311, 2005.
7. C. Carlet. *Boolean Models and Methods in Mathematics, Computer Science, and Engineering*, chapter Boolean Functions for Cryptography and Error Correcting Codes, pages 257–397. Cambridge University Press, New York, NY, USA, 1st edition, 2010.
8. C. Carlet. *Boolean Models and Methods in Mathematics, Computer Science, and Engineering*, chapter Vectorial Boolean Functions for Cryptography, pages 398–469. Cambridge University Press, New York, NY, USA, 1st edition, 2010.
9. S. Chari, J. Rao, and P. Rohatgi. Template Attacks. In B. Kaliski, Ç. Koç, and C. Paar, editors, *Proceedings of CHES 2002*, volume 2523 of *LNCS*, pages 51–62. Springer Berlin / Heidelberg, 2003.
10. J. Doget, E. Prouff, M. Rivain, and F.-X. Standaert. Univariate Side Channel Attacks and Leakage Modeling. *J. Cryptographic Engineering*, 1(2):123–144, 2011.

11. J.-H. Evertse. Linear Structures in Blockciphers. In D. Chaum and W. L. Price, editors, *Advances in Cryptology - Eurocrypt '87*, volume 304 of *LNCS*, pages 249–266. Springer, 1987.
12. B. Gierlichs. *Statistical and Information-Theoretic Methods for Power Analysis on Embedded Cryptography.* PhD thesis, Katholieke Universiteit Leuven, Faculty of Engineering, 2011.
13. B. Gierlichs, L. Batina, B. Preneel, and I. Verbauwhede. Revisiting Higher-Order DPA Attacks: Multivariate Mutual Information Analysis. In J. Pieprzyk, editor, *Topics in Cryptology – CT-RSA 2010*, volume 5985 of *LNCS*, pages 221–234, San Francisco, CA, USA, 2010. Springer-Verlag.
14. B. Gierlichs, L. Batina, P. Tuyls, and B. Preneel. Mutual Information Analysis: A Generic Side-Channel Distinguisher. In E. Oswald and P. Rohatgi, editors, *Proceedings of CHES 2008*, volume 5154 of *LNCS*, pages 426–442. Springer-Verlag Berlin, 2008.
15. C. Goutte and É. Gaussier. A Probabilistic Interpretation of Precision, Recall and *F*-Score, with Implication for Evaluation. In D. E. Losada and J. M. Fernández-Luna, editors, *Advances in Information Retrieval, 27th European Conference on IR Research, ECIR 2005*, volume 3408 of *LNCS*, pages 345–359. Springer, 2005.
16. H. M. Heys. A tutorial on linear and differential cryptanalysis. *Cryptologia*, 26:189–221, July 2002.
17. R. R. Hocking. The Analysis and Selection of Variables in Linear Regression. *Biometrics*, 32(1):1–49, 1976.
18. P. C. Kocher, J. Jaffe, and B. Jun. Differential Power Analysis. In *Proceedings of CRYPTO 1999*, pages 388–397, London, UK, 1999. Springer-Verlag.
19. G. Kowalski. *Information retrieval architecture and algorithms.* Springer, New York, 2011.
20. S. Mangard, E. Oswald, and F.-X. Standaert. One for All – All for One: Unifying Standard DPA Attacks. *IET Information Security*, 5(2):100–110, 2011.
21. M. Matsui. Linear Cryptanalysis Method for DES Cipher. In T. Helleseth, editor, *Advances in Cryptology – Eurocrypt '93*, volume 765 of *Lecture Notes in Computer Science*, pages 386–397. Springer, 1993.
22. E. Prouff. DPA Attacks and S-Boxes. In H. Gilbert and H. Handschuh, editors, *Fast Software Encryption*, volume 3557 of *LNCS*, pages 424–441. Springer Berlin / Heidelberg, 2005.
23. M. Renauld, D. Kamel, F.-X. Standaert, and D. Flandre. Information theoretic and security analysis of a 65-nanometer ddsll aes s-box. In B. Preneel and T. Takagi, editors, *CHES*, volume 6917 of *Lecture Notes in Computer Science*, pages 223–239. Springer, 2011.
24. M. Renauld, F.-X. Standaert, N. Veyrat-Charvillon, D. Kamel, and D. Flandre. A formal study of power variability issues and side-channel attacks for nanoscale devices. In K. G. Paterson, editor, *EUROCRYPT*, volume 6632 of *Lecture Notes in Computer Science*, pages 109–128. Springer, 2011.
25. W. Schindler, K. Lemke, and C. Paar. A Stochastic Model for Differential Side Channel Cryptanalysis. In J. Rao and B. Sunar, editors, *Proceedings of CHES 2005*, volume 3659 of *LNCS*, pages 30–46. Springer Berlin / Heidelberg, 2005.
26. K. Schramm, T. J. Wollinger, and C. Paar. A new class of collision attacks and its application to des. In T. Johansson, editor, *FSE*, volume 2887 of *LNCS*, pages 206–222. Springer, 2003.
27. F.-X. Standaert, B. Gierlichs, and I. Verbauwhede. Partition vs. Comparison Side-Channel Distinguishers: An Empirical Evaluation of Statistical Tests for Univariate Side-Channel Attacks against Two Unprotected CMOS Devices. In P. Lee and J. Cheon, editors, *ICISC 2008*, volume 5461 of *LNCS*, pages 253–267. Springer Berlin / Heidelberg, 2009.
28. F.-X. Standaert, T. G. Malkin, and M. Yung. A Unified Framework for the Analysis of Side-Channel Key Recovery Attacks. In A. Joux, editor, *Advances in Cryptology, Proceedings of EUROCRYPT 2009*, volume 5479 of *LNCS*, pages 443–461, Berlin, Heidelberg, 2009. Springer-Verlag.
29. S. S. Stevens. On the theory of scales of measurement. *Science*, 103:677–680, 1946.
30. K. Tiri and I. Verbauwhede. Securing encryption algorithms against dpa at the logic level: Next generation smart card technology. In C. D. Walter, Çetin Kaya Koç, and C. Paar, editors, *CHES*, volume 2779 of *Lecture Notes in Computer Science*, pages 125–136. Springer, 2003.
31. N. Veyrat-Charvillon and F.-X. Standaert. Mutual Information Analysis: How, When and Why? In C. Clavier and K. Gaj, editors, *Proceedings of CHES 2009*, volume 5747 of *LNCS*, pages 429–443. Springer Berlin / Heidelberg, 2009.
32. N. Veyrat-Charvillon and F.-X. Standaert. Generic side-channel distinguishers: Improvements and limitations. In P. Rogaway, editor, *Advances in Cryptology – CRYPTO 2011*, volume 6841 of *LNCS*, pages 354–372. Springer Berlin / Heidelberg, 2011.
33. C. Whitnall and E. Oswald. A Comprehensive Evaluation of Mutual Information Analysis Using a Fair Evaluation Framework. In P. Rogaway, editor, *Advances in Cryptology – CRYPTO 2011*, LNCS. Springer Berlin / Heidelberg, 2011.
34. C. Whitnall and E. Oswald. A Fair Evaluation Framework for Comparing Side-Channel Distinguishers. *Journal of Cryptographic Engineering*, 1(2):145–160, August 2011.
35. C. Whitnall, E. Oswald, and L. Mather. An Exploration of the Kolmogorov-Smirnov Test as Competitor to Mutual Information Analysis. Cryptology ePrint Archive, Report 2011/380, 2011. `http://eprint.iacr.org/`.
36. A. M. Youssef and S. E. Tavares. Resistance of Balanced S-Boxes to Linear and Differential Cryptanalysis. *Inf. Process. Lett.*, 56:249–252, December 1995.

A Conditions for a generic strategy to succeed

Here we provide simple proofs for the claims stated in Section 3.3. For conciseness we first prove Proposition 2 part (c) and then show that parts (a) and (b) are covered as special cases.

Proof. (Of 2(c)). Ultimately, k^* is indistinguishable from k if $F_k^{-1}[F_k(x)] \subseteq F_{k^*}^{-1}[L^{-1}[L \circ F_{k^*}(x)]]\ \forall x \in \mathbb{F}_2^n$ as this implies that F_k is just as accurate a model for $L \circ F_{k^*}$ as F_{k^*} (that is $Precision(F_k) = Precision(F_{k^*}) = 1$ and $Recall(F_k) = Recall(F_{k^*})$ as follows directly from the formulae).

It is sufficient to show that $\forall x \in \mathbb{F}_2^n$, $x' \in F_k^{-1}[F_k(x)] \Rightarrow x' \in F_{k^*}^{-1}[F_{k^*}(x)]$, since, trivially, $F_{k^*}^{-1}[F_{k^*}(x)] \subseteq F_{k^*}^{-1}[L^{-1}[L \circ F_{k^*}(x)]]$.

If $D_aF(x)$ depends on x only via $F(x)$ we can write $D_aF(x) = c(F(x))$ for some function $c : \mathbb{F}_2^m \to \mathbb{F}_2^m$.

It thus follows that $F_{k^*}(x) = F(x \oplus k^* \oplus a \oplus a) = D_aF(x \oplus k^* \oplus a) \oplus F(x \oplus k^* \oplus a) = c(F(x \oplus k^* \oplus a)) \oplus F(x \oplus k^* \oplus a) = c(F_{k^* \oplus a}(x)) \oplus F_{k^* \oplus a}(x)$.

So if $x' \in F_{k^* \oplus a}^{-1}[F_{k^* \oplus a}(x)]$ then:

$$\begin{aligned} F_{k^*}(x') &= c(F_{k^* \oplus a}(x')) \oplus F_{k^* \oplus a}(x') \\ &= c(F_{k^* \oplus a}(x)) \oplus F_{k^* \oplus a}(x) \\ &= F_{k^*}(x). \end{aligned}$$

I.e. $x' \in F_{k^*}^{-1}[F_{k^*}(x)]$ and thus $F_{k^* \oplus a}^{-1}[F_{k^* \oplus a}(x)] \subseteq F_{k^*}^{-1}[F_{k^*}(x)] \subseteq F_{k^*}^{-1}[L^{-1}[L \circ F_{k^*}(x)]]$.

Part (b) follows trivially once we notice that, if $a \in \mathbb{F}_2^n$ is a linear structure of F, we can replace $c(F(x))$ in the above argument with c for some $c \in \mathbb{F}_2^m$ constant over all x.

Part (a) follows from the observation that if F is affine, the linear space of F is the whole of $\mathbb{F}_2^n$ so that k^* is indistinguishable from $k = k' \oplus a$ for *all* $a \in \mathbb{F}_2^n \setminus \{\mathbf{0}\}$ (and thus for all $k \in \mathcal{K} \setminus \{k^*\} \subseteq \mathbb{F}_2^n$) by the same argument.

Proof. (Of Proposition 3). That the condition is necessary follows directly from Proposition **??**. Now suppose that, additionally, L is injective.

Choose $x' \in \mathbb{F}_2^n$ such that $\#D_{k^* \oplus k}F(F^{-1}[F(x' \oplus k)]) \neq 1$—which can be re-written as $\#D_{k^* \oplus k}F(F_k^{-1}[F_k(x')]) \neq 1$.

Thus $\exists x'' \in F_k^{-1}[F_k(x')]$ such that:

$$\begin{aligned} D_{k^* \oplus k}F(x' \oplus k) \neq D_{k^* \oplus k}F(x'' \oplus k) &\Rightarrow F(x' \oplus k \oplus k^* \oplus k) \oplus F(x' \oplus k) \neq F(x'' \oplus k \oplus k^* \oplus k) \oplus F(x'' \oplus k) \\ &\Rightarrow F(x' \oplus k^*) \oplus F(x' \oplus k) \neq F(x'' \oplus k^*) \oplus F(x'' \oplus k) \\ &\Rightarrow F_{k^*}(x') \oplus F_k(x') \neq F_{k^*}(x'') \oplus F_k(x'') \\ &\Rightarrow F_{k^*}(x') \neq F_{k^*}(x'') \qquad (\text{since } x'' \in F_k^{-1}[F_k(x')]) \\ &\Rightarrow x'' \notin F_{k^*}^{-1}[F_{k^*}(x')] \\ &\Rightarrow F_{k^*}^{-1}[F_{k^*}(x')] \neq F_k^{-1}[F_k(x')] \end{aligned}$$

Now we look at what this does to the precision and recall of F_k as a nominal model for F_{k^*}, beginning with the summands in the numerator of both expressions:

$$\#F_{k^*}^{-1}[L^{-1}[L \circ F_{k^*}(x)]] \cap F_k^{-1}[F_k(x)] = \#F_{k^*}^{-1}[F_{k^*}(x)] \cap F_k^{-1}[F_k(x)] \begin{cases} < 2^{n-m}, & \text{if } x = x' \\ \leq 2^{n-m}, & \text{if } x \neq x'. \end{cases}$$

By the balancedness of F and the injectivity of L the denominator summands in the precision and recall expressions always take the value 2^{n-m}. In this case, then, we get that $Precision(F_{k^*}) = Recall(F_{k^*}) = 1$ whilst $Precision(F_k) = Recall(F_k) < 1$, so that a sufficiently sensitive generic-compatible distinguisher will be able to reject the hypothesis k.

It only remains to show that sufficiency cannot be claimed when L is noninjective, which we do with a simple illustrative example:

Define $F : \mathbb{F}_2^3 \to \mathbb{F}_2^2$ and $L : \mathbb{F}_2^2 \to \{1, 2\}$ such that:

$$F(x) = \begin{cases} 0, & x \in \{0, 3\} \\ 1, & x \in \{1, 2\} \\ 2, & x \in \{4, 5\} \\ 3, & x \in \{6, 7\}, \end{cases} \qquad L(z) = \begin{cases} 1, & z \in \{0, 1\} \\ 2, & z \in \{2, 3\}. \end{cases}$$

$$\text{So } F_0(x) = F(x \oplus 0) = F(x)$$

$$\text{and } F_4(x) = F(x \oplus 4) = \begin{cases} 0, & x \in \{4, 7\} \\ 1, & x \in \{5, 6\} \\ 2, & x \in \{0, 1\} \\ 3, & x \in \{2, 3\}. \end{cases}$$

Then (for example) $F_0^{-1}[F_0(0)] = \{0, 3\} \neq \{0, 1\} = F_4^{-1}[F_4(0)]$, but nonetheless $F_0^{-1}[L^{-1}[L \circ F_0(0)]] = \{0, 1, 2, 3\} = F_4^{-1}[L^{-1}[L \circ F_4(0)]] \supset F_4^{-1}[F_4(0)]$ and in fact it can be checked that $F_4^{-1}[F_4(x)] \subset F_0^{-1}[L^{-1}[L \circ F_0(x)]]$ $\forall x \in \mathbb{F}_2^3$ so that $Precision(M_4) = Precision(M_0) = 1$ and $Recall(M_4) = Recall(M_0)$, implying that key candidates 0 and 4 cannot be distinguished from one another.

B Linear regression

Linear regression is a statistical method for modelling the relationship between a single dependent variable Y and one or more explanatory variables Z. It operates by finding a least-squares solution $\hat{\beta}$ to the system of linear equations $Y = Z\beta + \varepsilon$, where Y is an N-dimensional vector of measured outcomes, Z is an N-by-p matrix of p measured 'covariates', β is the p-dimensional vector of unknown parameters, and ε is the noise or error term, that is, all remaining variation in Y which is *not* caused by Z. Once the model has been estimated, the goodness-of-fit can be measured (for example) by the 'coefficient of determination', R^2, which quantifies the proportion of variance explained by the model: $R^2 = 1 - \frac{SS_{\text{error}}}{SS_{\text{total}}}$, where $SS_{\text{total}} = \sum_{i=1}^{N} (Y_i - \frac{1}{N} \sum_{i=1}^{N} Y_i)^2$ is the total sum of squares and $SS_{\text{error}} = \sum_{i=1}^{N} (Y_i - Z_i \hat{\beta})^2$ is the error sum of squares.

In the case that Z includes a constant term (the associated parameter estimate is called the intercept), the coefficient of determination is the square of the correlation coefficient between the outcomes and their predicted values: $R^2 = \rho(Z\hat{\beta}, Y)^2$. It is appealing as an attack distinguisher by virtue of this close relationship with correlation, coupled with the fact that it requires far less knowledge about the true form of the leakage to succeed. In correlation DPA the attacker has *prior knowledge* of a power model M and the distinguishing vector takes the form $D_\rho(k) = \rho(L_{k^*}(X) + \varepsilon, M \circ F_k(X))$. In linear regression DPA the challenge is to *simultaneously recover the true power model* along with the correct key as follows:

- Model the measured traces in function of the predicted coordinate function outputs and such higher-order interactions as you believe to be influential.
- Estimate the parameters and compute the resulting R^2 under each possible key hypothesis.
- If the largest R^2 is produced by the predictions relating to the correct key hypothesis then the attack has succeeded.

The LR-based distinguishing vector is thus: $D_{\mathrm{LR}}(k) = \rho(L_{k^*}(X) + \varepsilon, \hat{\alpha}_{k,0} + \sum_{u \in \mathcal{U}} F_k(X)^u \hat{\alpha}_{k,u})^2$, where ρ is Pearson's correlation coefficient, defined for two random variables A, B as $\rho(A,B) = \frac{\mathrm{Cov}(A,B)}{\sqrt{\mathrm{Var}(A)\mathrm{Var}(B)}}$.

C Stepwise regression

The inputs to the procedure are an $N \times 1$ vector Y containing observations of the dependent variable, p $N \times 1$ vectors $\{Z_i\}_{i=1}^{p}$ for each of the candidate explanatory variables, a set of indices indicating terms to be included regardless of explanatory power $I_{fix} \subset \{1, \ldots, p\}$ and a set of indices indicating *additional* terms to include in the initial model $I_{initial} \subseteq \{1, \ldots, p\}$ (s.t. $I_{fix} \cap I_{initial} = \emptyset$).

1. Set $I_{in} = I_{initial}$. Set $I_{test} = \{1, \ldots, p\} \setminus \{I_{in} \cup I_{fix}\}$.
2. For all $j \in I_{test}$ fit the model $Y = \beta_0 + \sum_{i \in I_{fix} \cup I_{in}} \beta_i Z_i + \beta_j Z_j + \varepsilon$ using least-squares regression and obtain the p-value on Z_j (call it $pval_j$).
3. If $\min_{j \in I_{test}} pval_j \leq pval_{add}$ then set $I_{in} = I_{in} \cup \mathrm{argmin}_{j \in I_{test}} pval_j$, $I_{test} = I_{test} \setminus \mathrm{argmin}_{j \in I_{test}} pval_j$ and repeat from step 2.
4. Else fit the model $Y = \beta_0 + \sum_{i \in I_{fix} \cup I_{in}} \beta_i Z_i + \varepsilon$ using least-squares regression and obtain $\{pval_i\}_{i \in I_{in}}$.
5. If $\max_{i \in I_{in}} pval_i \geq pval_{rem}$ then set $I_{in} = I_{in} \setminus \mathrm{argmax}_{i \in I_{test}} pval_i$, $I_{test} = I_{test} \cup \mathrm{argmax}_{i \in I_{test}} pval_i$ and return to step 2.
6. Else return I_{in}.

Note that the p-values on included terms change when other terms are added or removed—hence the need for an iterative procedure that re-tests the significance of included terms to identify candidates for removal. The threshold p-values for model entry and removal, $pval_{add}$ and $pval_{rem}$, are user-determined and will influence the resulting model. The terms included in the initial model will also influence the result. The MatLab defaults are $pval_{add} = 0.05$, $pval_{rem} = 0.1$ and $I_{initial} = I_{fix} = \emptyset$.

D Variability of measured outcomes

The asymptotic outcomes reported in Section 4.5 are based on 500 different leakage functions constructed to have uniformly random coefficients between -10 and 10. Figure 3 displays the medians but provide a reliable indication of the behaviour over the whole sample as the variance is moderate, at least in the case of AES and DES S-Boxes. By way of illustration, Figure 4 below shows the 1^{st} percentiles of the measured outcomes observed. Successful outcomes against AES and DES are preserved (although diminished); there are more failure cases against the PRESENT S-Box, which we conjecture is due to its smaller size, which restricts the degree of cryptographic nonlinearity attainable. It should, of course, be noted that these attacks use *fixed* stepwise inclusion/exclusion thresholds, and that the failure cases may respond to more sensitive tuning.

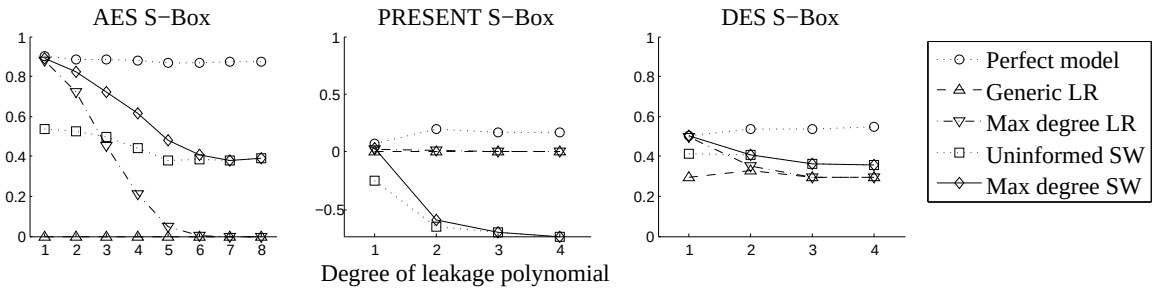

Fig. 4. First percentile of the distinguishing margins of attacks against AES, PRESENT and DES S-Boxes as the actual degree of the leakage polynomial increases (500 experiments with uniformly random coefficients between -10 and 10).