Statistical methods for non-profiled differential side-channel analysis: Theory and evaluation

CAROLYN WHITNALL

University of BRISTOL

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Engineering.

September 2012

c. 49,000 words
Abstract

Differential side-channel analysis (DSCA) aims at recovering cryptographically-secured secret information by exploiting the relationship between the physically-observable characteristics of a device and the data manipulated inside it. Prior knowledge about this relationship (obtained, perhaps, by detailed examination of an equivalent device) is known to greatly enhance attack success. What may be achieved with little or no prior knowledge at all is less clear. Strategies designed on such a basis have been loosely termed ‘generic’, but the scenarios in which these are possible without some meaningful knowledge on the leakage appear rare.

In this thesis we formalise the notion of ‘generic DSCA’ in order to understand it better and to make concrete statements about when and in what sense it is possible. We confirm that the range of scenarios to which it may be applied truly is limited—requiring that the device at some stage implements a predictable function which is non-injective and sufficiently nonlinear (e.g. the DES S-Box transformations).

We explore popular proposals based on mutual information and other non-parametric statistics. To facilitate meaningful comparisons we first introduce a theoretic evaluation framework to enable like-for-like comparisons between different methods and avoid the pit-falls of (necessarily estimator-dependent) empirical comparisons. One of the lessons learned by employing this framework is that mutual information is indeed optimal in some information-theoretic sense (as was initially supposed) and that it is the added burden of estimation which makes it a poor choice in all but the most unusual of leakage scenarios.

We also analyse linear regression-based methods and their use as ‘generic’ strategies. Applied in this way, they are restricted to the same limited scope as any other such strategy. However, we identify a unique feature of the way they operate which allows them to be adapted to incorporate some additional non-device-specific information in such a way as to produce successful outcomes in a far broader range of scenarios.
Acknowledgements

My PhD experience has been a resoundingly positive one; I am full of gratitude for the opportunity, the EPSRC funding, and the many and varied forms of support which have conspired to get me to this point. With so many wonderful people contributing in such wildly diverse ways it is definitely an impossible task to individually name and elaborate on each within a reasonable page limit. What follows is my best attempt to précis—full version available on request!

My first thanks must go to Elisabeth Oswald for undertaking to supervise my research and for guiding and supporting me through it with wonderful clarity. I have also enjoyed—and benefited greatly from—our opportunities to collaborate with François-Xavier Standaert.

Other people who have had particular input include Nigel Smart and my internal PhD progress reviewers, Dave Cliff and Bogdan Warinschi. Insightful comments from each of them have substantially influenced the direction of my research, and their general concern for my personal development (especially in taking on new challenges such as speaking and presenting) has been very much appreciated.

I cannot think of a lovelier bunch of people to work alongside than the Cryptography group here at Bristol—a delightfully diverse selection of brilliant minds and warm characters, who have borne with me with great grace and been generous enough to teach, challenge and affirm me as appropriate.

And of course, side-by-side with me in this endeavour has been the amazingly patient and generous Mr. Whitnall. Doing a PhD has been great but it’s just a small part of all that we are thankful for in our lives together—wonderful families, friendship, fellowship, and much else besides. May we learn to hold such blessings lightly and gratefully!
Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University’s Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate’s own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ........................................................................................................................................

DATE: ..........................................................................................................................................
# Contents

List of Figures xi

List of Tables xiv

1 Introduction 2

1.1 Physical security 4

1.1.1 Embedded devices 5

1.1.2 Implementation attacks 6

1.1.3 Block ciphers 8

1.1.4 Power consumption of common technologies 12

1.1.5 Measuring power consumption 14

1.2 Differential side-channel analysis 15

1.2.1 ‘Standard DPA’ 15

1.2.2 DPA distinguishers 17

1.2.3 Evaluating and comparing attacks 22

1.2.4 Factors influencing outcomes 23

1.2.5 Countermeasures 29

1.3 This thesis 35

vii
2 Prior knowledge and ‘generic’ DPA

2.1 Introduction ............................................................... 38
2.2 Delineating leakage assumptions ................................. 40
  2.2.1 Assumptions about the power model ......................... 41
  2.2.2 Assumptions about the noise ................................. 43
2.3 ‘Generic’ DPA: making no assumptions ......................... 44
  2.3.1 Conditions for a generic strategy to succeed ................ 45
  2.3.2 ‘Near-generic’ DPA and arbitrary power models .......... 51
2.4 Impact of model type on practical outcomes ................... 54
2.5 Conclusion ............................................................. 57

3 A theoretic evaluation framework for differential side-channel distin-
guishers ................................................................. 59

3.1 Introduction ............................................................. 59
3.2 Existing evaluation methods ........................................ 61
  3.2.1 Number of traces required for key recovery ................. 61
  3.2.2 Success rate and guessing entropy ........................... 63
3.3 Comparison framework ............................................... 64
  3.3.1 Computing theoretic vectors ................................ 64
  3.3.2 Outcome measures .............................................. 66
3.4 Conclusion ............................................................. 68

4 Mutual information and other nonparametric distinguishers .... 70

4.1 Introduction ............................................................. 70
4.2 Background to the distinguishers .................................. 73
## List of Figures

2.1 Types of leakage model and the assumptions required by common distinguishers. ................................................................. 41

2.2 The ‘near generic’ (7LSB) power model compared with actual trace distributions, when the device leaks the Hamming weight. ................................................................. 53

2.3 The ‘near generic’ (7LSB) power model compared with actual trace distributions, for an example case where the device does not leak the Hamming weight. ................................................................. 53

2.4 Example success rates for attacks against the first DES S-Box using power models of differing measurement level. ................................................................. 56

2.5 Example success rates for attacks against the AES S-Box using power models of differing measurement level. ................................................................. 56

2.6 Example success rates for attacks against the DES and AES S-Boxes using power models of differing measurement level, as signal-to-noise ratio varies. ................................................................. 57

3.1 An example theoretic distinguishing vector showing the correct key and the raw and standardised margin to the nearest rival. ................................................................. 68

3.2 Formulae for the proposed outcome measures comprising the theoretic evaluation framework. ................................................................. 69

4.1 Ideal distinguishing vectors using the HW power model to attack the output of the first DES S-Box. ................................................................. 84
4.2 Theoretic relative and absolute distinguishing margins as the (Gaussian) noise varies, for standard univariate attacks against the first DES S-Box.

4.3 Theoretic relative and absolute distinguishing margins as the (Gaussian) noise varies, for generic univariate attacks against the first DES S-Box.

4.4 Theoretic relative and absolute distinguishing margins as the (Gaussian) noise varies, for standard univariate attacks against the AES S-Box.

4.5 Theoretic relative and absolute distinguishing margins as the (Gaussian) noise varies, for 'near-generic' univariate attacks against the AES S-Box.

4.6 Theoretic relative and absolute distinguishing margins as the (Gaussian) noise varies, for standard univariate attacks against the PRESENT S-Box.

4.7 Theoretic relative and absolute distinguishing margins as the (Gaussian) noise varies, for 'near-generic' univariate attacks against the PRESENT S-Box.

5.1 Single-target and two-target MI-based theoretic distinguishing vectors against unprotected DES with noise-free Hamming weight leakage.

5.2 Theoretic relative and absolute distinguishing margins as the (Gaussian) noise varies, for standard two-target attacks against AddRoundKey and the first S-Box in an unprotected implementation of DES.

5.3 Theoretic relative and absolute distinguishing margins as the (Gaussian) noise varies, for generic two-target attacks against AddRoundKey and the first S-Box in an unprotected implementation of DES.

5.4 Theoretic relative and absolute distinguishing margins as the (Gaussian) noise varies, for two-target attacks against the first DES S-Box in a masked implementation.

5.5 Theoretic relative and absolute distinguishing margins as the (Gaussian) noise varies, for two-target attacks against AddRoundKey and the S-Box in an unprotected implementation of PRESENT.
5.6 Theoretic relative and absolute distinguishing margins as the (Gaussian) noise varies, for ‘near-generic’ two-target attacks against AddRoundKey and the S-Box in an unprotected implementation of PRESENT. 117

5.7 Theoretic relative and absolute distinguishing margins as the (Gaussian) noise varies, for two-target attacks against the PRESENT S-Box in a masked implementation. 119

6.1 Relative margins and critical support of theoretic attacks against Hamming distance leakage (with a reference state of $0100_{(2)}$) for varying levels of Gaussian noise. 129

6.2 Average number of traces required for key recovery in simulated practical attacks against Hamming-distance leakage (with a reference state of $0100_{(2)}$), for varying levels of Gaussian noise. 130

6.3 Theoretic success (in the absence of noise) against the first DES S-Box in the presence of data-dependent reference states of length 8 bits, as the number of different states increases. 134

7.1 Coefficients, in the polynomial expression for $L$, on the covariates as predicted under the correct and an incorrect key hypothesis. 143

7.2 Median distinguishing margins of single-target attacks against unprotected AES, PRESENT and DES S-Boxes as the leakage degree increases. 148

7.3 Median distinguishing margins of two-target attacks against masked AES, PRESENT and DES S-Boxes as the leakage degree increases. 149

7.4 First percentile of the distinguishing margins of single-target attacks against AES, PRESENT and DES S-Boxes as the actual degree of the leakage polynomial increases. 149

7.5 First percentile of the distinguishing margins of two-target attacks against masked AES, PRESENT and DES S-Boxes as the actual degree of the leakage polynomial increases. 150
List of Tables

4.1 Theoretic outcomes of single-target attacks against DES, AES and PRESENT in optimistic, realistic and challenging leakage scenarios without noise. . . \hspace{1em} 95

5.1 Theoretic outcomes against DES in optimistic, realistic and challenging leakage scenarios without noise. \hspace{1em} 113

5.2 Theoretic outcomes against PRESENT in optimistic, realistic and chal-
\hspace{1em} lenging leakage scenarios without noise. \hspace{1em} 120

6.1 Theoretical strength of correlation- and MI-based attacks against DES with
\hspace{1em} Hamming distance leakage from a constant reference state. \hspace{1em} 128
Chapter 1

Introduction

Differential side-channel analysis (DSCA) is a class of methods within the broader family of implementation attacks—cryptanalytic techniques to exploit the physical characteristics of cryptographic devices (and the environment in which they operate) in order to leverage knowledge on otherwise secret information.

The general strategy of a DSCA attack is to make predictions about the behaviour (run time [62], power consumption [63], electromagnetic radiation [44, 91], light [43, 104], heat [20] or sound emissions [7, 100]) according to some hypothesis about some secret the adversary wishes to uncover (part of a secret key, for example), and then to apply a statistical method to compare the predictions with actual measurements in order to test the hypothesis.

The more that is known a priori about the relationship between the processed data and the physical behaviour of a particular device, the more easily and efficiently an attacker can exploit measured data to extract secret information. This is the intuition behind ‘profiled’ attacks [30, 50, 92, 97], in which the adversary is assumed to have access to—and complete control over—an identical device from which to build detailed models of the leakage for different secret values. These models—known as ‘templates’—are supposed to capture characteristics of the leakage which vary with respect to the key but are fixed with respect to a class of devices. Once this has been achieved, measurements obtained from the target device can be compared against the templates (for example, using Bayesian classification) to reveal the most likely secret values.
However, such capabilities are a strong requirement on the adversary and seldom realistic. Indeed, the very notion of ‘identical device’ is problematic, particularly in the context of increasingly prevalent nanoscale technology. It has been shown that, as transistors shrink in size, the functional form of the power consumption no longer obeys expected patterns and, moreover, varies substantially between devices so that templates built from one device may not be good models for the leakage of another [94]. Therefore, it remains interesting to explore what can be achieved with minimal or no information about the form of the leakage. Attacks on such a basis have been loosely termed ‘generic’ because of the potential to apply them (equally fruitfully) to any implementation scenario.

The search for ‘generic’ methods has largely drawn from the field of nonparametric statistics and has uncovered some interesting candidates based on (for example) mutual information [49], the Kolmogorov–Smirnov two-sample test statistic [117, 123], the Cramer–von Mises test [117] and copulas [119]. However, as emphasised in [119], all of these suggestions have inherent limitations in that, unless the target function itself is non-injective, they inevitably require some sort of meaningful, non-injective mapping from the target values to a prediction model. The existence of an as-yet undiscovered method circumventing this requirement has remained an open question, and indeed is one of the motivations behind this thesis.

The aim, then, in what follows is to rigorously explore the scope and limitations of non-profiled DSCA. By formalising what it means for an attack strategy to be ‘generic’ we can make concrete statements about when and in what sense such an approach is possible (see Chapter 2). Within this setting we then explore the various nonparametric candidates (Chapters 4 and 5), as used ‘generically’ and otherwise, within a ‘fair’ theoretic evaluation framework which we propose in Chapter 3 to overcome some of the drawbacks of the usual, experimental approach to comparative analysis. Finally, we introduce a particularly interesting (parametric) approach to non-profiled DSCA which uses linear regression to perform a sort of on-the-fly profiling (Chapter 7). Whilst it can be used as a ‘generic’ strategy, it also motivates an alternative application which, by way of introducing some minimal, non-device-specific intuition, is able to succeed in a far broader range of attack scenarios.

We use the rest of this introductory chapter to provide the relevant motivating background
for our study, and describe the progress made in the existing literature. We first introduce, in §1.1, the problem of physical security and the real-life setting in which DSCA poses a particular threat—namely, embedded devices, with a particular focus on those implementing block ciphers such as the Data Encryption Standard (DES) and the Advanced Encryption Standard (AES). We will motivate the notion of data-dependent side-channel leakage with an in-depth look at power consumption in CMOS and FPGA technology. Then, in §1.2, we will describe DSCA attacks, the factors influencing outcomes, and some of the particular issues relating to improving or preventing DSCA. Finally, we outline the contribution of the remainder of this thesis in §1.3.

1.1 Physical security

Modern cryptography is the art of controlling information exchange. Its origins are in the practice of disguising messages so as to keep them confidential from all but the intended recipients, as, indeed, the etymology of the word implies (from the Greek: “secret”+“writing”). But, over the last century in particular, it has become a banner term encompassing a range of related goals, for example the task of communicating messages in such a way that the identity of the sender is authenticated to the recipient and/or the recipient is assured of the integrity of the information received.

Cryptanalysis is the art of flouting cryptography: of getting at information that somebody doesn’t want you to have, or interfering with communication in such a way as to destroy or divert trust between parties or to modify the information exchanged. The two disciplines co-evolve in the manner of an arms-race, catalysed in recent decades by the rapid technological progress of the 20th century.

Today, many cryptographic protocols are known to be robust against classical cryptanalysis, in the sense that they require a prohibitively demanding amount of computational effort to break (at least while efficient algorithms to solve certain supposed ‘hard problems’ in mathematics remain undiscovered). However, these assurances pre-suppose the classical black box view of protocols, in which an adversary can only view the inputs and outputs of the cryptosystem with no knowledge of the intermediate computations. In reality, this view has proved overly optimistic: even in cases where the security cannot
be compromised from inputs and outputs alone, aspects of the *physical implementation* of the algorithm have been found to ‘leak’ information about the intermediate states of the system. Since this information is unaccounted for in the security model which the system otherwise satisfies, it may be found to introduce unexpected vulnerabilities. It is becoming increasingly clear that there is a need to analyse protocols in a *grey box* setting if we are to get a realistic assessment of their working security. Some steps have been made towards appropriately incorporating physical observables into formal security models [38, 78, 107], but these works have primarily served to clarify the difficulty of the problem, so that a practically workable solution has yet to be seen.

With protocol-level protection a distant ideal, it becomes all the more important to study and understand implementation-level attacks in order, so far as is possible, to secure cryptographic *devices* which necessarily operate in potentially hostile environments.

### 1.1.1 Embedded devices

Modern day-to-day living is increasingly dependent on mobile, embedded and networked systems: smart cards, smart phones, RFID tags, PDAs, smart energy meters, access control—many of which rely on inbuilt cryptography in order to securely perform their intended functionality (make authorised payments, authenticate legitimate customers, protect personal information, and so on).

By contrast with (for example) personal computers which are built to perform, in parallel, a wide range of tasks, these novel devices usually house small systems designed to serve particular, targeted purposes. As such, the two-way interface between the device and the environment in which it operates can be far more directly tied to the specific, isolatable tasks as they are performed (largely in *sequence*) inside the device.

Moreover, the adversary from whom the device needs to be secured is very often the *device owner*, or at least holder. We might think of someone in possession of a pay-TV card or pre-pay travel card with the incentive to clone it or override the access restrictions/payment mechanism. Alternatively, an adversary handling bank cards might seek to extract keys in order to hijack accounts.

These two factors in particular—the inherent (and unavoidable) physical accessibility,
coupled with the natural relationship between the external and internal operation of embedded devices—exposes them to substantial threat from implementation attacks.

1.1.2 Implementation attacks

Passive implementation attacks exploit the fact that the internal state of the device impacts on the physical. A passive adversary therefore attempts to learn about the secret values protected by the protocol by observing the physical state during normal device operation. Conversely, active attacks exploit the physical interface from the other direction, by disturbing the physical state so as to influence the internal one ‘abnormally’.

Implementation attacks are also classified according to whether or not they are invasive. A non-invasive attack is one in which the chip and its packaging are left entirely intact and unaltered; a semi-invasive attack is one in which the packaging may be removed but the chip remains undisturbed; an invasive attack is one in which the electronic circuit is directly modified.

At one extreme, then, invasive active attacks include attempts to de-package and disassemble chips in order to reconstruct the layout by inspection (reverse engineering) or spy on intermediate values by microprobing. Such endeavours aspire to a direct reading of the internal state. However, they require expensive equipment and expertise, and moreover can be prevented to some extent by tamper-resistant countermeasures.

Fault attacks (identified in [18] and first implemented in [9]—see [60] for a comprehensive overview) are active and can be either non-invasive or semi-invasive. They work by disturbing the physical environment (e.g. via electromagnetic pulses [34], lasers [105], clock glitches [5], power surges [98], voltage drops [10]) in order to corrupt the protocol output. The hope is that the nature of the corruption will reveal something about the operation of the device which is unique or specific to a particular key (or a particular class of key candidates).

Passive, non-invasive attacks such as side-channel analysis (SCA) are of particular interest because they are generally less costly and less dependent on particular specialised equipment. In SCA the adversary measures some physical characteristic of the device in operation and combines these with public transcripts in order to leverage what may
be learned about the internal state of the system. For example, the run-time of (parts of) an algorithm may be expected to vary depending on some secret key—e.g. if the algorithm implements modular exponentiation with exponent-dependent branching [62]. Or, we might expect the power consumption as intermediate values are carried over a CMOS circuit to be data-dependent [63] (see §1.1.4 for more details). Other side-channel candidates include electromagnetic radiation [44, 91] and sound [7, 100], heat [20] and light emissions (although not all of these can be measured non-invasively—for example, measuring the light emissions requires depackaging the chip at a minimum [43]).

*Simple* SCA [63] entails visual inspection of one or more (averaged)\(^1\) trace observations relating to a fixed input, which can be particularly enlightening when the underlying algorithms involve data-dependent branching. For example, if the “double-and-add” approach is used for scalar multiplications in Elliptic Curve Cryptography (ECC), then—since the ‘addition’ step for each bit in the binary expansion is conditional on whether or not the bit takes value 1—the resulting computation time for that step, and the total power consumed, are greater than if the bit takes value 0, so that the secret key can be read directly from a trace [68].

Simple SCA can be prevented by avoiding secret-dependent branching. In our ECC example, “double-and-add-always” multiplication, and Montgomery multiplication, are two ways of removing the vulnerability [31, 61]. Each iteration of the algorithm as it loops over the bits of the key are performed in constant time, which reveals nothing about the key (although this protection does come at a cost—in the form of an increased overall running time).

*Differential* SCA (DSCA) [63] can pose a substantial threat even in scenarios protected against simple SCA. It relies on variation in the processed data (over a number of protocol runs) producing variation in the measured side-channel. The distribution of the traces is compared, via some statistical procedure, with a number of hypothetical distributions corresponding to guesses about the hidden algorithm values (for example, parts of a secret key). The intuition is that the correct key-part should produce (one of) the best match(es). Since DSCA is to be the primary focus of this thesis we provide a full explanation in §1.2.

\(^1\)Averaging helps to ameliorate against independent noise.
internal collision occurs when two different inputs produce (some of) the same intermediate values, even though the eventual outputs remain distinct. This cannot be observed from the transcript of the algorithm but can be detected by comparing (for example) the power traces associated with those inputs. Once pairs have been found, known properties of the algorithm can be used in an offline cryptanalytic stage to recover (for example) secret keys. Existing methods are often less efficient than DSCA because of the number of traces discarded in the search for collisions [79].

Algebraic side-channel attacks [28, 82, 93] use side-channel leakage as auxiliary information to simplify systems of algebraic equations for the algorithm outputs in function of the keys, which are then solved (for example, using a SAT solver). Original proposals required the side-channel leakage to be noise-free (for example, the perfectly recovered Hamming weights of the intermediate values); later developments [82] extend to the noisy scenario but the computational complexity of the attacks remains high and very sensitive to noise.

1.1.3 Block ciphers

A popular context for the study of DSCA is that of block ciphers: symmetric-key encryption/decryption for (deterministically) transforming fixed-length blocks of bits.

Formally, a block cipher is a pair of algorithms, $Enc$ to encrypt and $Dec = Enc^{-1}$ to decrypt. The inputs to $Enc$ are an $n_b$-bit plaintext message block $M$ and an $n_k$-bit long key $K$. It returns a $n_b$-bit ciphertext block $C$. $Dec$ takes $C$ and $K$ as inputs and returns the original message block $M$. Encryption can thus be viewed as a key-indexed bijection $Enc_K : \{0, 1\}^{n_b} \rightarrow \{0, 1\}^{n_b}$.

An iterated block cipher (which most are) performs the transformation via a repeated round function $Rnd$. The $i^{th}$ round, then, is of the form $M_i = Rnd_{K_i}(M_{i-1})$, where $M_0 = M$ and (for an $r$-round cipher), $C = M_r$. The round keys $\{K_i\}$ are all derived from the original key in some appropriate way—this part of the algorithm is called the key schedule.

A Feistel cipher is a special construction which proceeds by splitting the message block into two equal parts: the (key-indexed) round function is applied to one half, the output is XORed with the other half, and the halves are then swapped. If, then, $M = (L_0, R_0)$
denotes the original block split equally, and $\text{Rnd}$ the round function, round $i$ returns $(R_i, L_i \oplus \text{Rnd}(R_i, K_i))$. Decryption does not require that $\text{Rnd}$ be invertible: the original message block is recovered from the ciphertext by performing the identical algorithm with the round keys in reverse order.

The widely-used and much studied Data Encryption Standard (DES) [116] is one example of a Feistel network; others include Blowfish, Camellia, CAST-128, FEAL, ICE, KASUMI, LOKI97, Lucifer, MARS, MAGENTA, MISTY1, RC5, TEA, Triple DES, Twofish, XTEA.

Another important (and common) construction to consider is the substitution-permutation network, motivated by Claude Shannon’s recognised goals of confusion and diffusion [101]. Informally, confusion is the property that the message and the ciphertext have a highly complex and involved relationship; diffusion is the property that each input bit affects as many output bits as possible (for example, the strict avalanche criterion [120] requires that $\forall i, j \in \{0, \ldots, n_b - 1\}$, a change to the $i^{th}$ input bit affects the $j^{th}$ output bit with probability one half).

A substitution-permutation network in some sense delegates the particular tasks of achieving these two properties to two distinct stages in each round: a substitution stage (a highly non-linear transformation to provide confusion) followed by a (linear) permutation stage (to create diffusion). The substitution layer does not operate on the entire block, rather it maps small portions (bytes, for example) of input to portions of output via specially designed highly nonlinear functions called ‘substitution boxes’ (S-Boxes—which are also commonly used to provide nonlinearity in other constructions such as the Feistel-based DES). It then relies on the permutation layer to ensure that the output bits of a particular S-Box in one round get dissipated as inputs to as many different S-Boxes as possible in the next round. The decryption algorithm simply reverses the process, so unlike Feistel networks all the layers (subsequently, the S-Boxes) need to be invertible.

The Advanced Encryption Standard (AES) [33] is a substitution-permutation network; other examples include 3-Way, SAFER, SHARK, PRESENT, and Square.

Block ciphers—specifically, Triple DES and AES—are specified in the EMV smart card standard [39] for secure communication between cards and their issuers. This fact alone is enough for them to be considered of great practical relevance; the estimated number
of EMV-compliant cards in circulation by the end of 2011 exceeded 1.5 billion (45% of all payment cards) and the number of EMV-compliant terminals exceeded 21.9 million (76% of all payment terminals) [40]. Another application for block ciphers in embedded cryptography is smart metering; for example, the Universal Metering Interface [23] is an open specification which uses AES-128 for secure communication between customer meters and utility providers. Smart metering technology is expected to become increasingly prevalent over the next few years; the UK government, for example, announced in 2009 a national roll-out for gas and electricity, due to commence in 2014 and reach completion by 2019 [35]. Other (growing) areas of application include e-Passports [59] and GSM [2]. Hence there is plenty of motivation to pay careful attention to the physical security of block ciphers. Indeed, this is the context we consider for much of the analysis to follow.

**Boolean vectorial functions**

Block ciphers can, in general, be decomposed into combinations of Boolean vectorial functions, the algebraic properties of which are known to be particularly important to the cryptanalytic robustness of such systems. We here introduce some basic concepts, as will prove relevant to our later analyses; for a good but fairly brief introduction see [56] or, for a more comprehensive explanation, [26, 27].

A Boolean function accepts multiple binary inputs and maps to a univariate binary output: $f : \mathbb{F}_2^n \rightarrow \mathbb{F}_2$. All such functions have a unique representation in algebraic normal form (ANF): $f(x) = \bigoplus_{a \in \mathbb{F}_2^n} a_u x^u$, $a_u \in \{0, 1\}$, where $x^u = \prod_{i=1}^{n} x[i]^{|u[i]|}$. An $(n-m)$ Boolean vectorial function maps to a multivariate output $F : \mathbb{F}_2^n \rightarrow \mathbb{F}_2^m$; the coordinate functions of $F$ are simply the Boolean functions $f_j : \mathbb{F}_2^n \rightarrow \mathbb{F}_2$, $j = 1, \ldots, m$ such that $F(x) = (f_1(x), \ldots, f_m(x)) \forall x \in \mathbb{F}_2^n$.

The plaintext input $X \in \mathcal{X}$ to a block cipher is combined with secret key material $k \in \mathcal{K}$ so that the component functions can be expressed as key-indexed functions of $X$, that is $F_k(X) = F(k \ast X)$ where $F : \mathbb{F}_2^n \rightarrow \mathbb{F}_2^n$ is an $(n-m)$ Boolean vectorial function and $\ast$ denotes the key combining operator (for example, XOR). Those components which are especially designed to introduce confusion into the system are known as S-Boxes (where ‘S’ stands for ‘substitution’).
$F$ is affine if it can be expressed as a linear map followed by a translation—that is, if there exists a matrix $M \in \mathbb{F}_2^{n \times m}$ 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 (i.e. in terms of Hamming distance) from the set of all Boolean affine functions, in order to defend against linear cryptanalysis [74]. Thus the nonlinearity is defined:

$$N_F = \min_{u \in \mathbb{F}_2^n,v \in \mathbb{F}_2^m \setminus \{0\}} \sum_{x \in \mathbb{F}_2^n} u \cdot x \oplus v \cdot F(x),$$

where “$\cdot$” denotes the scalar product defined for $a,b \in \mathbb{F}_2^d$ as $a \cdot b = \bigoplus_{i=1}^d a[i]b[i]$.

An alternative notion of nonlinearity can be found in earlier S-Box design literature (e.g. [86]), namely, the sum of the coordinate-wise Hamming distances to the corresponding closest approximations:

$$N'_F = \sum_{i=1}^m N_{F_i}.$$

The preimage of $Y \subset \mathbb{F}_2^m$ is the set $F^{-1}[Y] = \{x \in \mathbb{F}_2^n | F(x) \in Y\}$. The preimage of a single element $y \in \mathbb{F}_2^m$ is the set $F^{-1}[\{y\}] = \{x \in \mathbb{F}_2^n | F(x) = y\}$. A Boolean vectorial function is balanced if all singleton preimages 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 [124] where any bias on the unobserved inputs is extremely undesirable.

Global, ‘worst case’ properties such as these are usually considered the most pertinent to cryptographic robustness. It may be, though, that related average or local properties are more important to SCA resistance. A good starting point is to take into account the whole of the linear approximation table, described (for example) in [56]: $\text{LAT}_F(u,v) = \# \{x \in \mathbb{F}_2^n | u \cdot x = v \cdot F(x)\} - 2^{n-1}$. The linear bias associated with a given approximation represented by the pair $(u,v) \in \mathbb{F}_2^n \times \mathbb{F}_2^m$ is $2^{-n}\text{LAT}_F(u,v)$.

The key property providing resistance to differential cryptanalysis as introduced by Biham and Shamir [15] 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 [41]) can be exploited by a cryptanalyst. The space $\{a \in \mathbb{F}_2^n | \#D_a F = 1\}$ is the linear space of $F$ and the larger it is, the more susceptible to cryptanalysis.
1.1.4 Power consumption of common technologies

For DSCA to be a reasonable concern it must be that something measurable in the physical environment in which the protocol runs is related to the intermediate values which are presumed ‘invisible’ on the basis that they are not part of the black box output of the protocol. As a usefully illustrative example we consider the power consumption of a CMOS circuit.

The static power consumption of a CMOS logic cell is typically very low, so that the overall power consumption is dominated by dynamic power consumed whenever an internal or output signal of the cell switches. (Although, in future technologies, the balance is likely to shift: static power consumption increasingly dominates as devices are scaled down to nanometer ranges). It is moreover the case that the power consumed by an output signal switching is much higher than that consumed by an internal signal switching, so that we can disregard the latter. Of course, signals are switched according to the data being processed by the circuit, so that there is a strong dependency between the power consumption and the intermediate values of the algorithm.

For example, the power consumed as a value is transferred over a data bus is known to be approximately proportional to the Hamming distance between that value and the one transferred directly previous, \( \text{HD}(v_0, v_1) = \text{HW}(v_0 \oplus v_1) \), since this gives the number of bits which are flipped on the bus wires. (Note that for this approximation to hold requires that 0 → 1 and 1 → 0 transitions consume roughly the same amount of power, and also that each wire contributes equally and independently).

Therefore, an adversary able to measure the power consumption (e.g. using a digital sampling oscilloscope) as a particular intermediate value is carried over a data bus in a CMOS circuit can guess the intermediate value and the preceding value on the bus and investigate whether the Hamming distance between the two is proportional to the measurements, thereby confirming or confuting the guess. Noise in the measurements—that is, measurement error or random fluctuations in the power consumption independent of the data processed—introduces uncertainty into the comparison, which can be mitigated for by acquiring more data.

Of course, an adversary may not know the preceding value in order to compute the Ham-
ming distance—although Brier et al. [19] explain that, in software implementations where the previous value on the bus is always the address of an instruction, the reference state will at least be constant (which turns out to make it much easier to exploit the leakage). Or, the simplifying assumptions may not be reasonable—for example (as investigated in [3]), 0 → 1 and 1 → 0 transitions may contribute differently; there may be imbalances or interactions between the wires; the data-dependent contribution may not be purely deterministic (i.e. it may have a random component so that the noise cannot be assumed independent of the processed value). To the extent that the attacker is aware of such behaviour he can use it to inform the attack, as we shall explore in our discussion of power models once we have formally introduced differential power analysis (DPA). Otherwise, as we shall also explore, it turns out that even poor approximations of the device behaviour can be adequate for success in many scenarios.

Different leakage characteristics can be observed for hardware implementations. For example, [109] analysed FPGA power consumption, and found that it could be well-predicted by the number of bit transitions inside the device registers. Rather than the Hamming distance from a constant reference state, the device leaks something resembling the Hamming distance from a non-constant but predictable intermediate value (that is, the one preceding the target value).

We see from these examples that the functional form of the data-dependency emerges from the particular way in which the data influence the device operation. Other sources of side-channel leakage produce different functional forms, and good mechanical knowledge of the technology can help in approximating a meaningful leakage model. But the intuition behind power analysis in particular motivates the expectation of a variety of side-channels—not least because varying power consumption itself is known to impact on other observable characteristics such as electromagnetic radiation and thermal dissipation (so that the data-dependency may be thought of as occurring via power consumption-dependency).
1.1.5 Measuring power consumption

An attacker is assumed to have physical access to the device in order to acquire the requisite data to perform an attack. The acquisition stage entails measuring the power consumption as the algorithm operates on different (known) inputs with a fixed (unknown) key.

A typical measurement set-up involves several interacting component parts. A PC is used to communicate with the cryptographic device, triggering the execution of the algorithm and receiving the output. The device generally requires its own (stable) power supply (provided, for example, by a smart card reader in the case that it is a smart card), and an external clock signal. The power consumption signal is usually recorded by a digital sampling oscilloscope, which also receives instruction from (and sends data to) the PC. An oscilloscope is only able to measure voltage, so an attacker must find a way to produce a voltage signal which is proportional to the property (the current, or power) he is interested in measuring. This can be achieved via a power measurement circuit: a small resistor is inserted into the ground line or the positive supply line between the power supply and the device, producing a voltage drop which is proportional to the current (as long as the supply voltage is constant). Alternatively, an electromagnetic probe may be used to indirectly measure the current via the magnetic field around a wire.

Minimising noise in the measurement stage can be crucial to the success of an attack. In particular, the smaller the amplitude of the data-dependent signal, the more sensitive the measurement equipment will need to be in order to detect this against the background (independent) power consumption. Specialist measurement circuits have been designed explicitly to enhance DPA, e.g. by means of a transimpedance amplifier [21] (a device which converts current to voltage). It is also important to have an oscilloscope with sufficient bandwith to capture the frequency of the power consumption signal, with adequate resolution (in reality, each measurement is quantised to take one of a finite number of values—normally $2^8$), and to sample at a high enough frequency to capture the important features of the signal. Most modern oscilloscopes meet these requirements for most typical scenarios.
1.2 Differential side-channel analysis

We have established, then, the basic principle behind DSCA: the observable behaviour of a physical device depends in some way on the data being operated on inside the device. Information about intermediate values which are neither inputs nor outputs to the algorithm (and therefore assumed to be unobserved) may therefore be present in all sorts of measured data—execution time [62], power consumption [63], electromagnetic radiation [44, 91], even sound [7, 100] and light [43, 104] emissions and thermal imaging [20].

We have not yet discussed how exactly this extra information may be exploited. It is not enough simply that unintended leakage occurs, the adversary needs an appropriate strategy to extract some secret value (such as an encryption/decryption key). It is well-established in the literature that detailed a priori knowledge about the physical characteristics of a device (obtained, for example, by ‘profiling’ an identical device under full control of the attacker) facilitates effective, efficient extraction of targeted information [30, 97]. It has also been shown, though, that quite a lot can often be achieved even when such knowledge is lacking or imperfect [12, 49, 48, 89, 117, 119], provided sufficient data are available (consistent with the well-known trade-off between flexibility and efficiency of statistical methods [45]). Methods in this category have come to be known informally as ‘generic side-channel distinguishers’, but it is not always clear precisely what is meant by this. The question as to whether truly ‘generic’ strategies exist, able to remain effective in any leakage scenario without requiring any knowledge on that scenario, is one which we seek to clarify in later chapters (see, in particular, Chapter 2).

Because of the particular emphasis on power analysis (ever since the ground-breaking work of Kocher et al. in 1999 [63]) and the intuitive relationship between processed data and power consumption of common technologies (as per §1.1.4, above) we present the following explanation and discussion of DSCA in the context of DPA, noting that similar principles apply to many types of data-dependent observations.

1.2.1 ‘Standard DPA’

For the purpose of unifying studies across the literature, the notion of a ‘standard DPA attack’ was defined in [71]. Suppose that the power consumption $P$ of the target crypto-
graphic device depends on some internal value (or state) $F_{k^*}(X)$. The state is a function of some part of the plaintext which is a random variable $X \overset{R}{\in} \mathcal{X}$, as well as some part of the secret key $k^* \in \mathcal{K}$. Consequently, we have that $P = L \circ F_{k^*}(X) + \varepsilon$, where $L$ is some function which describes the data-dependent component and $\varepsilon$ comprises the remaining power consumption which can be modelled as independent random noise. 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 $M : F(\mathcal{X}) \mapsto \mathcal{M}$.

A distinguisher $D$ is some quantity which compares two random variables, as applied to the leakage $P = L \circ F_{k^*}(X) + \varepsilon$ and the model prediction $M_k = M \circ F_k(X)$) arising from some key hypothesis $k \in \mathcal{K}$. The theoretic distinguishing vector is $D = \{D(k)\}_{k \in \mathcal{K}} = \{D(L \circ F_{k^*}(X) + \varepsilon, M \circ F_k(X))\}_{k \in \mathcal{K}}$, where the plaintext input $X$ takes values in $\mathcal{X}$ according to some known distribution (usually uniform).

Sometimes it is useful to abstract away from the impact of noise—for example, if we want to focus on the role of the target function or that of the power model quality or type. For convenience, we define the ideal distinguishing vector: $D_{\text{Ideal}} = \{D_{\text{Ideal}}(k)\}_{k \in \mathcal{K}} = \{D(L \circ F_{k^*}(X), M \circ F_k(X))\}_{k \in \mathcal{K}}$.

In practice an adversary only has access to the distribution of $P$ via a sample drawn from that distribution, so that $D$ must be estimated rather than computed directly. Suppose we have observations corresponding to the vector of inputs $x = \{x_i\}_{i=1}^N$, and write $e = \{e_i\}_{i=1}^N$ to be the observed noise (i.e. drawn from the distribution of $\varepsilon$). Then the size $\# \mathcal{K}$ estimated distinguishing vector is $\hat{D}_N = \{\hat{D}_N(k)\}_{k \in \mathcal{K}} = \{\hat{D}_N(L \circ F_{k^*}(x) + e, M \circ F_k(x))\}_{k \in \mathcal{K}}$.

We concentrate on the notion of key-recovery success as formalised by Standaert et al. in [108]. We say the attack is theoretically successful (respectively, ideally successful) if $D(k^*) > D(k) \forall k \neq k^*$ (respectively, if $D_{\text{Ideal}}(k^*) > D_{\text{Ideal}}(k) \forall k \neq k^*$). We say it is $o$-th order theoretically successful (respectively, $o$-th order ideally successful) if $\# \{k \in \mathcal{K} : D(k^*) \leq D(k)\} \leq o$ (respectively, if $\# \{k \in \mathcal{K} : D_{\text{Ideal}}(k^*) \leq D_{\text{Ideal}}(k)\} \leq o$).
From a practical perspective it follows that the attack is successful if \( \hat{D}_N(k^*) > \hat{D}_N(k) \forall k \neq k^* \) and \( o\text{-th order successful} \) if \( \# \{ k \in \mathcal{K} : \hat{D}_N(k^*) \leq \hat{D}_N(k) \} \leq o \).

Put simply, a DPA attack succeeds if the model predictions corresponding to the correct key hypothesis bear more resemblance to the true power traces than the model predictions corresponding to the incorrect alternatives, provided the attacker is able to estimate an appropriately chosen comparison statistic with sufficient precision to detect the hypothesis-dependent difference in model accuracy. In \( \S \)1.2.4 we examine in a bit more detail the way the various components of the attack influence outcomes, and the interesting research questions which arise, but first we discuss what it means for one DPA attack to be ‘better’ than another (or, conversely, for one implementation to be ‘more resistant’ than another).

### 1.2.2 DPA distinguishers

In this section we (very) briefly introduce all the distinguishers which will feature in our analysis and discussion. Details and background information can be found in the relevant chapters, in particular Chapter 4.

**Bayesian classification**

In Bayesian classification DPA [30] the key-dependent ‘model’ possessed by the adversary is a probability density function for the leakage under each key-hypothesis (estimated, for example, in a profiling stage): \( M_k = \hat{g}_{P|K=k}(\cdot) \). The idea is that the actual trace distribution should have a density closely resembling one of these hypothetical densities.

The mechanism to distinguish between the hypotheses uses Bayes’ theorem: The \emph{a posteriori} probability for a discrete random variable \( \theta \) given a (sample of a) continuous random variable \( A \) is defined as \( \mathbb{P}(\theta = t | A) = \frac{\mathbb{P}(\theta = t) \mathcal{L}(\theta | A)}{\mathbb{P}(A)} \), where \( \mathcal{L}(\theta | A) = p_{\theta}(A) \) is the likelihood—the probability density of \( A \) viewed as a function of \( \theta \).

The \emph{a posteriori} probability of a particular hypothesis \( K = k \) given a set of trace measurements \( t \) (that is, independent realisations of the random variable \( P \)) is therefore proportional to the product of the prior probability \( \mathbb{P}(K = k) \) (assumed to be known,
and often uniform) and the likelihood $\mathcal{L}(t|K = k)$ which we compute according to the models $M_k$. Since we have defined the notion of theoretic distinguishing vector in relation to the full input support space $\mathcal{X}$, we consider, for each hypothesis, the expectation (with respect to the actual trace distribution) of the \textit{a posteriori} probability associated with the model densities evaluated at $L \circ F_k^*(x) + \epsilon$ for all $x \in \mathcal{X}$. As is the usual approach with so-called \textit{naive} Bayesian classification, the known-to-be-unlikely simplifying assumption that the conditional distributions $P|K = k, k \in \mathcal{K}$ are independent is made, and the joint density thus computed as the product.

$$D_{BC}(k) = \mathbb{P}(K = k) \mathbb{E}_{\mathcal{X}} \left[ \prod_{x \in \mathcal{X}} M_k(L \circ F_k^*(x) + \epsilon) \right]. \tag{1.1}$$

In a practical attack, the adversary has $N$ observations $t_1, t_2, \ldots, t_N$ associated with input values $x_1, x_2, \ldots, x_N$ and it is this realised sample for which the likelihoods are computed—or rather, in practice, the log-likelihoods:

$$\ln \mathcal{L} \left( \{t_i\}_{i=1}^N | K = k \right) \approx \sum_{i=1}^N \ln M_k(L \circ F_k^*(x_i) + e_i).$$

**Correlation**

Pearson’s correlation coefficient, defined for two random variables $A$ and $B$ as $\rho(A, B) = \frac{\text{cov}(A, B)}{\sqrt{\text{var}(A)\text{var}(B)}}$, can be directly applied as a distinguisher to compare model predictions with the actual trace values [19]:

$$D_{\rho}(k) = \rho(P, M_k) = \frac{\text{cov}(P, M_k)}{\sqrt{\text{var}(P)\text{var}(M_k)}}. \tag{1.2}$$

It is generally estimated via the sample correlation coefficient (which simply entails by plugging the sample variances and covariances into the above formula). For a potentially more robust method (with respect to the model quality), sometimes Spearman’s rank correlation is used instead [11] (in which case the values of the variables are replaced with their rankings).
**Difference-of-means**

The earliest published DPA attacks [63] operated by partitioning the measured traces (or a subset of them) into two sets according to some hypothesis-dependent criteria and comparing the average power consumption in each set, the idea being that only the correct hypothesis would produce a meaningful partition—and therefore a detectable difference—with all other hypotheses producing (approximately) random partitions with (approximately) the same mean.

The model $M_k$ relates to the partition criterion. This could be as simple as the value of a single bit in the predicted target, or could be more complex, involving the values of several bits.

$$D_{\text{DoM}}(k) = \mathbb{E}(P|M_k = m_1) - \mathbb{E}(P|M_k = m_2)$$ (1.3)

It is estimated via the sample means in each of the two groups.

**Non-parametric goodness-of-fit statistics**

Goodness-of-fit statistics make for interesting DPA tools because they extend the ‘partitioning’ mechanism of the difference-of-means distinguisher to allow for more than two partitions\(^3\) (see, in particular, [106]). In general, they operate by comparing the global trace distribution with the resulting conditional distributions in order to detect the key hypothesis which produces the most ‘meaningful’ partition.\(^4\)

The mutual information (MI) based distinguisher of [49] was the first to be widely discussed and falls into this category because it can be framed as the expectation (with respect to the conditioning value) of the Kullback-Leibler divergence between the global

---

\(^3\)In [77] were presented several variations on the partitioning function, incorporating multiple bits but ultimately using only two partitions in the DPA comparison.

\(^4\)The crucial feature of partitioning-based distinguishers, which we will discuss at more length in Chapter 2, is that they are invariant to a re-labelling of the partitions (noting that in a difference-of-means strategy the attacker is looking for the largest *absolute* value, so that swapping the 0/1 labelling is without effect). This is by contrast with (for example) correlation DPA, which—although it does, in some sense, operate by partitioning the global distribution—uses the partition labels themselves (the power model values) in the subsequent computations, so that it is *not* invariant to re-labelling.
and partitioned traces:

\[
D_{MI}(k) = I(P; M_k) = H(P) - H(P|M_k) = H(P) - \mathbb{E}_{m \in \mathcal{M}} [H(P|M_k = m)] \quad (1.4)
\]

\[
= \mathbb{E}_{m \in \mathcal{M}} [\text{Div}_{KL}(P|M_k = m || P)]. \quad (1.5)
\]

Estimation is problematic and generally achieved by first estimating the densities and then substituting these into plug-in estimators for the entropies. We discuss this in considerable depth in Chapter 4.

The above motivated similar distinguishers based on the Kolmogorov–Smirnov test statistic and the Cramér–von Mises criterion, proposed in [117]. These each compare global and conditional cumulative density functions \( F_A(a) = \mathbb{P}(A \leq a) \), which makes them slightly more straightforward (though not necessarily more efficient) to estimate.

The Kolmogorov–Smirnov based distinguisher finds the supremum of the absolute difference between the two distributions, averaged over the conditioning values:

\[
D_{KS}(k) = \mathbb{E}[K(P||P|M_k)] = \mathbb{E}_{m \in \mathcal{M}} \left[ \sup_y |F_P(y) - F_{P|M_k = m}(y)| \right]. \quad (1.6)
\]

The Cramér–von Mises based method is similar but integrates the squared differences across the entire support, rather than just at the point of greatest divergence:

\[
D_{CvM}(k) = \mathbb{E}[C(P||P|M_k)] = \mathbb{E}_{m \in \mathcal{M}} \left[ \int_{-\infty}^{\infty} (F_P(y) - F_{P|M_k = m}(y))^2 dy \right]. \quad (1.7)
\]

**Variance ratio**

A variance ratio-based alternative was suggested in [106], motivated by the fact that, under the commonly-held assumption of Gaussian noise, the entropy of a distribution depends only on its variance. Therefore, rather than the data-intensive process of estimating densities or cumulative densities as required for the non-parametric goodness-of-fit comparisons above, the hypothesis-dependent partitions could be ranked according to the
proportion of the overall variance which is accounted for by the partitioning variable:

\[
D_{VR}(k) = \frac{\text{var}(P)}{\mathbb{E}_{m \in M} \text{var}(P|M_k = m)}.
\] (1.8)

**Copula-based distinguisher**

A novel twist on the usual approach, presented at Crypto 2011 [119], operates by first applying a copula transform to the trace distribution in order to work with \( Z = F_P(P) \) — where \( F_P \) is the cumulative distribution, defined (as above) as \( F_P(y) = \mathbb{P}(P \leq y) \). By definition of the probability integral transform, \( Z \) is known to have a uniform distribution.

The attack mechanism relies on the idea that meaningful partitions on \( Z \) will not be uniform, whilst random partitions will. This is essentially the same intuition as that behind the difference-of-means and the goodness-of-fit statistics discussed earlier. The authors of [119] propose to test each key hypothesis \( k \in K \) by looking at the Manhattan distances between pairs sampled from \( Z|M_k = m, \forall m \in M \). These can be pooled over all \( m \in M \) and the distribution compared with the (known) distribution corresponding to the uniform case (via the integrated square distance). Only in the case of a meaningful partition (i.e. the correct key hypothesis) would we expect this divergence to be large.

Writing down a formula for the distinguishing vector becomes rather involved, and is probably less informative than the preceding description, but we include it for completeness. Suppose \( P_1, P_2 \) are identically distributed according to the leakage distribution; then \( Z_i = F_P(P_i) \sim U(0, 1) \). Denote by \( d_{\text{man}} \) the Manhattan distance (which, in this univariate case is simply \( d_{\text{man}}(a, b) = |a - b| \)) and let \( f_{d_{\text{man}}(Z_1, Z_2)} \) and \( f_{d_{\text{man}}(Z_1|M_k = m, Z_2|M_k = m)} \) denote the probability densities of the global and conditioned Manhattan distances viewed as random variables. Then, noting that pooling the conditioned distances produces a mixture density, the distinguishing vector can be written as:

\[
D_{CP}(k) = \int_0^1 \left( f_{d_{\text{man}}(Z_1, Z_2)}(y) - \sum_{m \in M} \mathbb{P}(M_k = m)f_{d_{\text{man}}(Z_1|M_k = m, Z_2|M_k = m)}(y) \right)^2 dy.
\] (1.9)

The method is shown to perform well by comparison with MI-based distinguishers in a standard univariate scenario, and is particularly suited to attacks of higher dimension.
$d \geq 2$, where the copula transformation can be applied independently for each dimension, resulting in $d$ univariate cumulative functions (whereas MI-based DPA requires estimating multivariate densities). The distributions of the conditioned and global Manhattan distances (summed, of course, over all $d$ dimensions) are compared just as in the univariate case.

A full analysis of this distinguisher within the framework presented in Chapter 3 would be a valuable contribution—particularly with respect to its uniquely simple multivariate operation—but is outside the scope of this thesis. We have rather focused on two broad classes of distinguishers: MI and similarly-operating goodness-of-fit tests (see Chapters 4 and 5), about which several misconceptions persist in spite of (or because of) a fairly voluminous literature, and linear regression-based methods (see Chapter 7), which are emerging as potentially more flexible than previous methods in exploiting minimal prior information.

### 1.2.3 Evaluating and comparing attacks

To avoid over-stating the physical security of a device it is important to take into account the most powerful methods available to an attacker with access to side-channel measurements. Attempts to compare different distinguishers in the search for the ‘most effective’ have thus received considerable attention in the literature (see [106] for a particularly thorough empirical evaluation).

The theoretic performance of a distinguisher can only be ascertained when the device leakage can be fully characterised, which is obviously not possible in practical, real life scenarios, where one deals always with samples from unknown distributions. Nonetheless, it is informative to explore the theoretic feasibility and strength of certain DPA strategies against (well-defined) hypothetical leakage. If the theoretic distinguisher values rank the correct key within the top $o$ hypotheses, under certain ‘reasonable’ stated assumptions, then we would expect a practical instantiation under equivalent conditions to be ($o$-th order) successful ‘eventually’—that is, once provided with enough data to estimate the quantities to sufficient precision. If the correct key is not within the top $o$ hypotheses as ranked by the theoretic quantities then we would not expect a practical attack to be ($o$-th
order) successful except by chance (that is, due to sampling variance in the estimates).

Given, then, a distinguisher which is theoretically capable of recovering the key, a natural question arises: how large does the sample size $N$ need to be for a successful attack? It was recognised, from the very earliest DPA literature [63], that such a figure, if it could be provided, would represent a relevant and highly informative indicator for the physical security of a device. Related metrics are the ($o$-th order) success rate and guessing entropy of [108], defined (respectively) as the probability of ($o$-th order) success and the expected number of candidates remaining to test, as the sample size varies.

The techniques of statistical power analysis$^5$ [66] can be used to compute required sample sizes (conversely, the statistical power for a given sample size, which corresponds to the success rate) provided the sampling distributions of the estimator can be approximated under the various key hypotheses. This, as we shall explore more fully in Chapter 3, is unrealistic in most cases—though with some extra simplifying assumptions it is achievable in the case of correlation DPA [70].

Most of the time, then, sample sizes, success rates and guessing entropies are obtained experimentally, by performing practical attacks against simulated or measured traces. One drawback is that all such evaluations are necessarily estimator specific and may not be good indicators of the inherent theoretic capabilities of the evaluated distinguishers. We tackle this shortfall in Chapter 3, by introducing an evaluation framework for making meaningful comparisons at the theoretic level, with direct implications for practical success.

1.2.4 Factors influencing outcomes

Many conflicting and compounding factors contribute to DPA outcomes. The various DPA strategies available are differently suited to different scenarios, so that general statements about the merits of one approach relative to another are hard to come by. In reality, outcomes are highly scenario-specific so that the ‘best’ (that is, most effective, or efficient) approach to the analysis will be different in each case.

$^5$The name conflict is an unfortunate consequence of the separation between the fields of statistics and cryptography—the ‘power’ in this case refers to the probability that a statistical test will reject the null hypothesis when the null hypothesis is false, and has nothing to do electric current.
In what follows we explore some of these factors and their role in influencing the feasibility or effectiveness of a DPA attack, in order, later, to establish which strategies are best suited to which scenarios.

**Target function**

To begin with, the target intermediate function $F$ is known to play an important role in determining DPA outcomes; some operations—most notably those which are designed to be cryptographically secure (S-Boxes used in block ciphers, for example)—are particularly vulnerable [55, 88]. This is because small changes in the input produce big changes in the output, so that any wrong key hypothesis is far more likely to produce predictions which differ substantially from the true consumption. On the other hand, cryptographically weak operations such as the AddRoundKey functions in block ciphers (where plaintext bytes are XOR-ed with key bytes) are far more resilient to DPA, as similar key bytes produce similar predictions and it becomes harder (and more data-costly) to clearly spot the correct one.

In the quest for ‘generic’ DPA strategies [49, 106, 119, 123], non-injectivity has also been implicated as a relevant target characteristic, determining whether or not the ‘identity’ function can be used effectively in place of an informed power model. Indeed, as summarised in [119], the need for some meaningful, non-injective power model mapping whenever the target function is injective appears inescapable. One of the goals of Chapter 2 of this thesis is to formalise and clarify this observation, and to explore the characteristics of $F$ which make it vulnerable or otherwise to ‘generic’ DPA.

**Finding ‘interesting points’**

In practice, rather than observing a univariate $P$ an attacker usually begins with a time-indexed vector of power consumption values $\{P_t\}_{t=1}^T$, of which a limited (initially unknown) subset indexed by $\tau \subset \{1, \ldots, T\}$ are of the desired data-dependent form. In order to recover the secret key the attacker must therefore first (or simultaneously) recover a non-empty subset $\tau' \subset \tau$.  

24
The usual approach is to estimate the distinguishing vector for every point in the power trace $\hat{D}_N^{(t)} = \{\hat{D}_N(P_t, M \circ F_k(x))\}_{k \in K}$ and then identify the pair $(t, k)$ which maximises $\hat{D}_N^{(t)}[k]$. The attack is $o$-th order successful if $\#\{k \in K : \max_{t \in \tau} \hat{D}_N^{(t)}[k^*] \leq \max_{t \in \{1, \ldots, T\}} \hat{D}_N^{(t)}[k]\} \leq o$ [70].

Of course, this greatly adds to the computational complexity of the attack (by a factor of $T$) so that it is highly desirable to reduce the number of candidate points prior to computing the distinguishing vectors. Visual inspection of the averaged traces can be adequate to detect patterns likely to indicate, for example, the location of S-Box computations. Investigating the data-dependent variance can also reveal time windows which are dominated by such processes: [30] proposes choosing those points which maximise the pair-wise differences between the (data-dependent) conditional means; [92] takes a similar approach and additionally advocates a fast Fourier transform pre-processing to remove high frequency noise.

Another, more sophisticated proposal is to transform the data using principal component analysis [6, 13] and then perform the DPA attack on only the top-ranked few components. A problem here arises with the fact that there are many processes behind the variance in the power consumption (particularly if the initial window is wide) and the targeted function will not necessarily feature early on.

**Power model**

The characteristics of the device leakage—the functional form of the data-dependent component $L$ and the relative size and shape of the independent noise $\varepsilon$—will substantially dictate how easily and effectively the side-channel can be exploited. In particular, studies such as [37] have clearly demonstrated the central role of the power model $M$ in determining distinguisher performance. In the case that an attacker has full control over an identical device, profiling (as, for example, in [30, 97]) can produce a very good approximation of the leakage function. However, in this work we are usually interested in a weaker adversary, with access only to an unverifiable guess based on what is known or assumed about the underlying technology of the device. Therefore the extent to which the device leakage is ‘typical’ or predictable will have a significant bearing on the attack.
outcome.

As we explained in §1.1.4, some devices, such as those built with CMOS technology, are well-known to consume power approximately proportional to the Hamming weight of/Hamming distance between processed values. This has resulted in the widespread popularity of Hamming weight/distance power models in side-channel research, in particular when little is known about the true form of the leakage and the device cannot be profiled. However, not all technology conforms neatly to predictable behaviour. For example, the authors of [94] found evidence that power consumption in emerging nanoscale technologies may not be consistent even between ‘identical’ devices, so that even an attacker with profiling capabilities can have difficulty predicting the leakage of a target device with confidence. Non-standard leakage can also be observed in typical hardware implementations of S-Boxes [72]. This has contributed to the increasing interest in finding so-called ‘generic’ DPA strategies, which are able to succeed with little or no prior information on the leakage. In Chapter 2 we explore more thoroughly the role of the power model, delineating the types of model possible and formalising the notion of a ‘generic power model’ (correspondingly, ‘generic DPA’).

**Choice of distinguisher**

The choice of distinguisher $D$ must be appropriate to the available model (as we explain in more detail in Chapter 2). Statistics which may be efficiently estimated are expected to produce successful outcomes with fewer data, provided a good choice of estimator is made. Choosing a good estimator, though, is not necessarily trivial, particularly for non-parametric statistics where general results about estimator performance are hard to come by.

For example, an adversary who is confident that their model $M$ is a good approximation for the true leakage $L$ up to proportionality would be well-advised to use Pearson’s correlation coefficient [19]—an extremely popular and much documented choice for $D$ on account of its efficiency (being moment-based) and the prevalence, among those scenarios studied in the literature, of Hamming weight-/Hamming distance-style leakage assumptions (which may be readily modelled).
The correlation coefficient quantifies the total linear dependency between two random variables $A$ and $B$, and is defined as $\rho(A, B) = \frac{\text{cov}(A, B)}{\sqrt{\text{var}(A)} \sqrt{\text{var}(B)}}$. It takes values from -1 to 1 and is zero whenever $A$ and $B$ are independent. (However, the converse is not true; namely, $A$ and $B$ may be (non-linearly) dependent with a (linear) correlation of 0). It is estimated from samples $\{a_i\}_{i=1}^n$, $\{b_i\}_{i=1}^n$ via the sample correlation coefficient:

$$r(A, B) = \frac{\sum_{i=1}^n (a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^n (a_i - \bar{a})^2} \sqrt{\sum_{i=1}^n (b_i - \bar{b})^2}}.$$ 

This is a consistent estimator for $\rho(A, B)$ and, moreover, is asymptotically unbiased and efficient if $A$ and $B$ have a joint Gaussian distribution.

The correlation-based DPA distinguisher has been consistently found to outperform other proposed (non-profiled) distinguishers in cases where the ‘proportional model’ assumptions hold, and even (often) to maintain an advantage as the model quality degrades [89, 108, 117]. In practice, its simplicity as a moment-based quantity makes it a strong candidate whenever some linear correspondence exists between the model and the true leakages. Nonparametric alternatives (e.g. mutual information) are far more flexible to different levels of model approximation but carry a hefty overhead in terms of estimation costs which make them far less efficient [103]. They are only really useful in scenarios where the prior information on the leakage is inadequate for correlation DPA.

With the reasonable expectation that correlation DPA is the ‘best’ approach whenever the requirements on the power model are met, we are particularly interested in the use and usefulness of nonparametric alternatives whenever these requirements are not met. See in particular Chapters 4, 5 and 6.

**Noise**

There may be several sources of noise in the measured traces: rounding/measurement error, electronic noise produced by the transistors, algorithmic noise from other, non-targeted, computations, or model-fit error comprising data-dependent variation not accounted for by the power model. Non-profiling adversaries in the standard (univariate) DPA scenario defined above generally assume that the noise is independent of the data and
often make the additional assumption that it is Gaussian (although this latter assumption
is usually only needed for formal statistical inference, such as influences sample size com-
putations). Profiling adversaries can relax the standard assumption of data-independence
and seek to characterise data-dependent noise distributions—generally by retaining the
Gaussian assumption and estimating variances and (in the case of multi-target attacks)
covariances. Such an approach is computationally complex but has the advantage of
exploiting random as well as deterministic data-dependent leakage which may result in
faster (i.e. less data-complex) key recovery.

Supposing, then, a standard non-profiling scenario in which the assumption of independent
noise is reasonable; the size of the noise is naturally highly relevant to the practical effi-
ciency of an attack. The weaker the signal-to-noise ratio (SNR, defined as \( \frac{\text{var}(L_0F_k(X))}{\text{var}(\varepsilon)} \)),
the more data will be required to estimate the distinguishing vector with sufficient preci-
sion to detect the true key (see Chapter 4 of [70]).

In fact, the sample size \( N \) required for correlation DPA can be explicitly approximated in
function of the SNR, under the assumption of Gaussian noise. The correlation between
the traces and the model prediction relating to key hypothesis \( k \in \mathcal{K} \) is simply scaled
in the presence of noise: \( \rho(P, M_k) = \frac{\rho(L_0F_k(X), M_k)}{\sqrt{1 + \frac{1}{\text{SNR}}}} \). The techniques of ‘statistical power
analysis’ (as described in §1.2.3 above) yield the following equation (see Equation (4.43)
in [70]) for the sample size needed to distinguish the correct key \( k^* \) from the nearest rival
hypothesis \( k^{NR} \):

\[
N = 3 + 8 \cdot \frac{z_{1-\alpha}^2}{\left( \ln \frac{1+\rho(P, M_{k^*})}{1-\rho(P, M_{k^*})} - \ln \frac{1+\rho(P, M_{k^{NR}})}{1-\rho(P, M_{k^{NR}})} \right)^2}
\]

where \( z_{1-\alpha} \) is the \( (1-\alpha)^{th} \) quantile of the standard normal distribution.

Surprisingly, though, we uncovered in the course of our research (see Chapter 4 and [121])
that not all distinguishers respond so straightforwardly to noise. In fact, for nonparametric
statistics such as mutual information and the Kolmogorov–Smirnov test, noise can actually
affect the shape of the theoretic distinguishing vector—that is, it doesn’t necessarily scale
the quantities equivalently so as to preserve the relative values. In some scenarios it
actually plays a role in determining whether or not the correct key is identified, as well as
by how great a margin. One of the implications of this discovery is that it is not usually
sufficient to evaluate distinguisher performance in a pure-signal setting and then reason
that noise affects only the \emph{efficiency} of a practical attack (although, of course, this does still hold for correlation DPA).

\section*{Distribution of inputs}

The distribution of the plaintexts (or, conversely, ciphertexts) $X$ may or may not be under the control of the adversary but is generally assumed to be known (most attention has been given to the uniform case). Some work has been done on (adaptively) chosen message side-channel attacks \cite{64, 118}, motivated by the enhancement that adaptive message selection adds to classical cryptanalysis. Indeed these efforts find that required sample sizes are reduced by the ability to adaptively select, so that worst-case security is lower than that implied by known message attacks.

\subsection*{1.2.5 Countermeasures}

Until (indeed, unless) practically realisable protocols can be found which are provably secure according to appropriate leakage-resilient security models, it is left up to designers of cryptographic devices to protect their products against DSCA (and other implementation) attacks. Such attempts are necessarily ad-hoc and reactive (that is, in response to discovered threats); new side-channels are being discovered and exploited all the time and no one technique will prevent all possible attacks, or even, in general, all known attacks.

In some cases, such as the EMV standard \cite{39} for smart cards used in banking, it is possible to enforce regular key updates (“session keys”) which limits the number of repeat observations an attacker can gather for a fixed key and provides very strong protection against DSCA (a threshold sample size must be obtained before patterns and variations begin to manifest). However, even just a single trace measurement can be adequate for simple SCA (if the protocol is not implemented carefully), or for a sophisticated multivariate template attack if the adversary has full access to an identical profiling device. Moreover, it is often not practical or possible to frequently update secret keys in this way, in which case the designer usually chooses a strategy (or combination of strategies) designed to increase the \emph{workload} of an attack, thereby reducing the opportunity of
an typical adversary for whom it becomes increasingly difficult to gather the necessary resources or volume of data.

The basic goal of any such measure is to break the link between the targeted intermediate values and the side-channel leakage, either through *hiding* (altering the physical characteristics of the device in an attempt to make the leakage *random* or *constant* with respect to the processed data) or *masking* (manipulating the intermediate values so that the data-dependent leakage is random with respect to the target).

A comprehensive overview of common countermeasures (and techniques to circumvent them) can be found in Chapters 7 to 10 of [70]. Since many have emerged from industry practice rather than the research community, not all have scientific publications associated with them. Therefore, in our summary, we will endeavour to cite the literature where appropriate and otherwise refer the reader to [70] as being an excellent description of practice.

**Hiding countermeasures**

Hiding countermeasures only necessarily apply to the side-channel they are intended to alter. Since we are concentrating most particularly on DPA (as the most commonly explored variant of DSCA) throughout this thesis, we consider proposals to make the power consumption in each clock cycle random or constant, noting that such measures may or may not impact on other leakage sources.

Power consumption can be randomised in the time dimension by inserting dummy operations (or, in hardware, dummy clock cycles) or shuffling the order in which (necessarily independent) operations (e.g. the 16 S-Box lookups in an AES round) are performed. Both will produce traces which are misaligned in repeated observations. The former adds to the execution time of the algorithm so that there is a trade-off between the amount of randomness introduced and the resulting computational efficiency; the latter has fewer implications for efficiency but is limited by the algorithm-specific scope for swapping operations around whilst producing the same output.

\[\text{For example, the description of electromagnetic radiation in terms of current (via the Biot–Savart law) would indicate that efforts to control the power consumption may in fact impact usefully on electromagnetic leakage also.}\]
To completely randomise the power consumption in the amplitude dimension implies manipulating the signal-to-noise ratio to be 0—either by reducing the signal to zero (that is, making the data-dependent component constant) or by introducing infinite-variance noise. The potential to achieve this in practice is naturally very limited, but to whatever extent the signal can be diminished and/or the noise magnified, the work of the adversary will increase, as precise estimation of the distinguishing statistics requires larger and larger sample sizes the more the noise dominates [66]. Thus, a designer may be content with adding ‘infeasible’ data-complexity to an attack, for some realistic, scenario-specific adversary.

The signal can be reduced by inserting a filter between the power supply pins and the circuit that is computing the cryptographic algorithm, using, for example, switched capacitors or constant current sources. Alternatively, cell-level techniques have been suggested, such as dual-rail precharge (DRP) logic (e.g. [113, 114]), which encodes all logic signals on complementary wires with a constant precharge value of 0 or 1. Thus there are two outputs to each cell: one corresponding to the processed data and one to its complement, so that (supposing the precharge value is 0) always the transition $0 \rightarrow 1 \rightarrow 0$ occurs on one of the outputs while the other stays at 0. This will remove data-dependency from the power consumption only to the extent that the capacitances in the wires are successfully balanced. In practice, the balancing is unlikely to be perfect; we explore the consequences of imperfectly-realised DRP logic in Chapter 6.

Noise can be increased at the hardware level by introducing a noise engine to the device, in the form of a random number generator connected to a network of large capacitors. Another way of increasing it (at the hardware or software level) is to parallelise as much of the algorithm as possible, so that the power consumption during the clock cycle in which the targeted value is processed depends on other (unrelated) intermediate values.

**Circumventing hiding**

Hiding countermeasures, then, increase the data complexity of successful DPA but do not render it asymptotically impossible. In the case of time-domain randomisation, techniques exist for re-aligning traces prior to analysis—pattern matching, for example (which can
also help if the traces are badly aligned due to imperfections in the measurement set-up).
Failing that, integrating traces within fixed-width intervals may help to strengthen the correct key comparison if several of the randomised instances are combined within one window. This is far less effective though—whereas perfectly re-aligned traces will yield the same success rates as unprotected leakage, attacks against pre-processed misaligned traces will require larger samples.

The impact of amplitude-domain measures depends directly on the resulting SNR: the smaller this becomes, the more trace measurements an adversary will need to acquire to perform DPA successfully (as per the discussion under ‘noise’ in §1.2.4).

Masking countermeasures

Masking countermeasures [29, 53] adopt a different approach. Rather than investing effort into changing the leakage characteristics, they accept the data-dependent nature of the leakage and work instead to disguise (by randomising) the intermediate values so that they become unpredictable to an attacker. That is, if \( F_k(X) \) is never processed directly, and moreover is associated with different random values on repeated observations, then an attacker whose key-dependent predictions are based on \( F_k(X) \) will not find a strong match at any point in the measured trace.

In a \( d^\text{th} \)-order masking scheme, an intermediate value \( v \) is concealed via \( d \) secret values \( m_1, \ldots, m_d \) so that the algorithm operates on \( v_m = v * m_1 * \ldots * m_d \) where * denotes the masking operation, typically chosen according to the operations that are used in the cryptographic algorithm (Boolean addition, or arithmetic addition/multiplication modulo \( n \) for some appropriate \( n \)). Masking an algorithm which uses both Boolean and arithmetic operations requires switching between both types [52].

A masked implementation thus processes \( v_m \) instead of \( v \), and simultaneously tracks the appropriate changes to the masks after each operation so that the output may be unmasked at the end. It is straightforward to do this for linear operations as they change the masks in an easily computable way, for example \( g(v \oplus m_1 \oplus \ldots \oplus m_d) = g(v) \oplus g(m_1) \oplus \ldots \oplus g(m_d) \). However nonlinear operations such as S-Boxes (important, of course, for the cryptanalytic resilience of a cipher) are not so easy to mask, and carry a computational overhead, for
example, requiring that each element of a look-up table be individually masked at the start of a protocol run.

Because this thesis most frequently refers to the AES and DES block ciphers, we concentrate on Boolean masking. Provided the Boolean masks are independent and uniformly random, the processed value $v_m$ will be independent of the target value $v$. With the extra condition that the masks and the masked values are not processed at the same time (or consecutively, in a device leaking the Hamming distance) then the leakage as $v_m$ is processed does not reveal anything about $v$.

If an operation is known to leak the Hamming distance, the option to implement a masked scheme may be rejected in favour of implicit masking by way of random precharging [112]. This entails loading or storing a random value immediately prior to the intermediate value $v$, so that the leakage is masked (that is, the device leaks $\text{HD}(v, m)$ where the attacker has no way of predicting $m$).

Random precharging can also be implemented at the hardware level, by sending random values through the circuit in order to randomly precharge all logic cells [22]. Other hardware options include building masked multipliers and adders to explicitly compute operations on masked inputs, returning appropriately masked outputs [16].

**Circumventing masking**

Masking is not always sufficient to prevent standard univariate DPA (where by univariate we mean with respect to the number of trace points exploited in one go). As hinted towards in the above section, if the masks are processed in the same clock cycle as the masked value, or in consecutive clock cycles (in the case of Hamming distance leakage), the power consumption at those points will not be independent of the target intermediate value. The work of [72] shows how the presence of glitches in hardware implementations of masked schemes can also lead to vulnerabilities.

However, even a masking scheme which successfully protects against univariate DPA may yet be vulnerable to multivariate strategies in which multiple trace points are jointly

---

7Other challenges are attendant with arithmetic masking, such as the fact that the intermediate value 0 cannot be concealed with modular multiplication—which introduces a vulnerability to zero-value DPA [4, 51].
exploited. This is motivated by observing the fact that, though $L(v_m)$ might be independent of $v$, the joint distribution of $(L(v_m), L(m_1), \ldots, L(m_d))$ is not independent of $v$. Therefore, if leakage measurements can be made for the points in time at which all the masks and the masked value are processed, there may be opportunity to learn something about the target intermediate value.

Techniques exist to exploit this. In profiled DPA, multivariate templates (models) may be built to optimally account for the information in the combined measurements [83]. In non-profiled DPA, a typical approach is to pre-process multiple points in a trace into a univariate value by way of a ‘combining function’, and then to perform ‘standard DPA’ on that value [76, 110].

Two-target correlation DPA is the classical and most explored example. A popular choice of combining function is the normalised product: $C(L(v_m), L(m)) = (L(v_m) - \hat{E}(L(v_m))) \cdot (L(m) - \hat{E}(L(m)))$ (where $\hat{E}$ denotes the estimated expectation, i.e. the sample mean). It was shown in [90] that in the case of Hamming weight leakage and Boolean masking, the output of this combination function is also proportional to the Hamming weight of $v$, which makes for a very natural attack strategy. However, the problem of choosing an ‘optimal’ combination function is not at all straightforward and depends on the precise scenario; the appropriate prediction function likewise needs to be re-considered in each case. By consequence, whilst variants on (univariate) ‘standard DPA’ have been shown to be essentially equivalent (asymptotically) so that only one approach need be considered for a good security evaluation [71], the various multi-target DPA strategies cannot be similarly unified, which problem was emphasised and explored in [110].

(Non-profiled) linear regression analysis [37, 97], which can be viewed as a generalisation of correlation DPA to the case where the power model is not known a priori and must be recovered ‘on-the-fly’ (simultaneously with the correct key), can also be extended to multi-target attacks by way of a pre-processing step [36]. We explore this further in Chapter 7.

Because relevant information is lost when multiple values are combined in this way, other recent proposals have sought to avoid the pre-processing stage and operate on the raw

---

8These multiple points are generally assumed to be associated with different target values—for example, a mask and a masked value, or an S-Box input and output. We avoid the popular description ‘higher-order DPA (HODPA)’ as it has assumed several subtly different meanings.

34
multivariate data. For example, [12, 48] explore the use of multivariate mutual information, [73, 123] apply a similar adaptation to the Kolmogorov–Smirnov-based distinguisher, and the copula-based method of [119] extends to higher-orders without pre-processing. Often, though, the increased data-complexity of estimating these multivariate statistics make them inefficient relative to correlation DPA with pre-processed traces, in all but the most unusual of leakage scenarios. We discuss multivariate extensions of nonparametric distinguishers in Chapter 5.

1.3 This thesis

A key motivation for this thesis is the prevailing uncertainty over the role of prior knowledge in determining non-profiled DPA outcomes. We seek firstly to establish unambiguously what is possible to an attacker with little or no information about the true relationship between the device leakage and the data. Attacks on such a basis have been the focus of much recent literature and have been loosely termed ‘generic’. However, there has been a lack in the way of general statements about the potential for such strategies, the focus rather being on specific candidate instances and the extent to which they comply with an informal (and changeable) notion of ‘genericity’.

In Chapter 2 we seek to rectify this lack, by introducing an appropriate formal definition and reasoning directly from this definition in order to establish precisely when and in what sense generic success is feasible. A fairly strong requirement emerges—that the target function be non-injective and sufficiently nonlinear (in some appropriate sense)—so that the range of scenarios vulnerable to generic DPA is confirmed to be very limited. Also in this chapter we provide a useful delineation of leakage assumptions which, in extension of ideas hinted towards in [46], recognises the different types of prior knowledge an attacker might have, and the importance of using a statistical method which is appropriate to type.

Having established some fundamentals about prior knowledge in general and generic strategies in particular, the latter, larger part of the thesis is concerned with specific methods for exploiting minimal prior knowledge. However, making meaningful like-for-like comparisons is not straightforward, because of the estimator-dependent nature of the
outcomes and the absence of scenario-transferable best-case estimators for most of the
statistics considered. We discuss this problem in Chapter 3 and, by way of partial resolu-
tion, introduce a framework for evaluating the theoretic properties of distinguishers apart
from the practical performance of the chosen estimation procedures.

This framework provides us with a basis on which to evaluate and compare existing pro-
posals for exploiting minimal prior knowledge, such as those using nonparametric statistics
like mutual information, the Kolmogorov–Smirnov test and the Cramér–von Mises crite-
ron. In Chapter 4 we discuss the theory behind these methods and explore how they
can applied as ‘generic’ distinguishers. We evaluate them so used but also look at their
performance in combination with prior knowledge of varying quality, according to sugges-
tions made in the existing literature. In particular, we test the supposed ‘near-generic’
power model which attempts to group intermediate values according to a truncation of
their binary expression. We show that this does not constitute a suitable analogue to
‘generic’ DPA.

Chapter 5 deals with multivariate extensions of correlation, mutual information and
Kolmogorov–Smirnov based DPA for use against multiple target values, for example,
in attacks against masked implementations. Again, we consider both generic and non-
generic applications.

In Chapter 6 we consider, as a particular case-study, Hamming distance leakage from
unknown reference states, examining in a bit more depth the advantages of mutual
information-based attacks over correlation. We also discuss the relevance of such analysis
to the special, DPA-resistant circuit technology known as dual-rail precharge logic, and
show how imperfections in the implemented logic style may make it vulnerable to such
techniques as mutual information-based analysis.

In Chapter 7 we explore (non-profiled) linear regression-based methods which, we show,
also qualify as ‘generic compatible’ according to our formalisation in Chapter 2 (and
subject to the usual limitations). However, we go on to identify interesting scope for
adaptation, by which non-device-specific intuition about the leakage may be incorporated
in such a way as to substantially increase the range of scenarios which may be successfully
attacked. We describe the resulting strategy as ‘generic-emulating’.

36
Chapter 8 summarises our findings and highlights particular open questions presenting avenues for further research. It also gives details of publications drawn from the content of this thesis.
Chapter 2

Prior knowledge and ‘generic’ DPA

2.1 Introduction

The many and various strategies proposed for differential side-channel analysis have differing requirements on the quantity, quality or nature of the prior knowledge available to the attacker. At one end of the spectrum are fully-profiled attacks, which exploit access to an identical device in order to estimate detailed models for the data-dependent component of the leakage (the deterministic part and sometimes the random part too). Methods to accomplish this vary (for example, the Bayesian classification-style templates of [30] or the stochastic (linear regression) models of [97]), but all culminate in the task of comparing hopefully very accurate power consumption predictions with known distributions in order to reveal the correct hypothesis on the subkey. A maximum likelihood approach takes full account of the (potentially multivariate) estimated distributions but simpler methods like correlation could be used if the data-dependent component of the leakage is thought to be primarily deterministic.

Adversaries with such capabilities can often narrow down the potential key candidates with only very few trace measurements (possibly even just a single trace if the attacker builds a template model targeting more than one intermediate value simultaneously [92]). Attacker capabilities degrade as the quality of the prior knowledge diminishes: the more
vague or unspecific the assumptions, generally the less efficient the resulting attack; or, if the assumptions are sufficiently inaccurate, then the attack may fail altogether. So, for example, if the data-dependent leakage is correctly supposed to be approximately proportional to the Hamming weight of the targeted intermediate value, a correlation attack using this power model can be very effective [19]. If the only assumption made is that bit $i$ has a substantial effect on the leakage, a difference-of-means attack using the $i^{th}$ bit to partition the leakage measurements will require more data but should eventually succeed as long as the assumption is correct. If the leakage is supposed to be proportional to the Hamming weight but is in fact proportional to the Hamming distance from some unknown reference state then a straightforward correlation attack is not likely to succeed [19] (‘wrong’ information is not the same as ‘less’ information).

An interesting research question is the existence of DPA distinguishers which are able to succeed right at the lower end of the ‘information quantity’ spectrum—that is, given little or no information about the true data-dependent leakage. In practice this reflects a more realistic attack scenario; it is a strong requirement that an attacker have access to a device for profiling, besides which such methods can be rendered unreliable by potentially substantial intra-device variation (observed, for example, in nanoscale technologies [94]). Strategies for use in such a setting have been termed ‘generic’ and are widely sought. Proposed examples include mutual information (MI) based DPA using an identity power model [49], distinguishers based on the Kolmogorov–Smirnov (KS) two-sample test statistic [117, 73], linear regression (LR) based methods which can be seen as a sort of on-the-fly profiling [37], and an innovative approach using copulas [119].

However, until now the notion of ‘generic strategy’ has remained informal, and ambiguities persist throughout the above literature. In the case that the target function is injective, existing distinguishers, without exception, require some piece of knowledge about the leakage: for example, a meaningful partition (in the case of MI- and KS-based DPA)\(^1\) or a restriction on the appropriate set of covariates (in the case of LR-based DPA) [12, 89, 119]. This reliance on knowledge, however vaguely expressed, rather contravenes even our informal expectations from a generic strategy.

\(^1\)The 7LSB model has been proposed as a circumvention of this requirement [12, 49], but as we shall show later, attacks exploiting it are only able to succeed if and when noise in the measurements distorts the true leakage distribution towards the model. In other words, the model is meaningful, albeit unintentionally.
In this chapter, we offer valuable clarification on the various forms that prior knowledge can take, with recourse to the widely accepted ‘levels of measurement’ laid out in [111]. Where most previous studies have talked about ‘good’ power models, in an arbitrary sense, we recognise (as was hinted towards in [11, 46]) the material distinction between different levels of model. A given distinguisher will implicitly interpret a model (for the data-dependent, deterministic part of the leakage) according to one of these types, and the resulting outcomes will depend on how accurate that model is according to the appropriate notion of accuracy for that type. It is therefore crucial that we are able to delineate between model types in order to apply strategies which are compatible with our assumptions and prior knowledge.

These clarifications tend towards an appropriate formalisation for the notion of ‘generic strategy’, which we shall define as the pairing of a generic-compatible distinguisher with the generic power model (also to be defined). By reasoning directly from these definitions we can abstract away from the analysis of specific distinguishers and begin to make statements which are more conclusive and generally-applicable. For the first time we are able to demonstrate that non-injectivity is a prerequisite for any generic strategy to succeed; the hunt for universally-applicable distinguishers requiring no information about the leakage will only uncover variations on a theme (though with some scope for improved efficiency).

We further show that non-injectivity alone is not sufficient to distinguish between hypotheses and investigate additional required properties of the target function. It is already known that there is an inverse relationship between performance against certain S-Box criteria and susceptibility to DPA [88]; we demonstrate a sufficient condition for generic success which is promoted (though not inevitably produced) by the particular goal of differential uniformity [15].

### 2.2 Delineating leakage assumptions

There are two aspects to the assumptions made by an attacker about the side-channel leakage. First there are the assumptions made about the deterministic, data-dependent component, as captured by the power model. Then there are the assumptions about
the distribution of the noise—which in most cases are not directly exploited but rather play an implicit role in determining how accurately or efficiently certain statistics may be estimated. Figure 2.1 visualises this two-dimensional continuum, and indicates the suitability of some popular distinguishers as assumptions vary.

![Figure 2.1: Types of leakage model and the assumptions required by common distinguishers.](image)

* The justification for the claim that linear regression-based distinguishers exploit nominal power models is developed in §7.2.1 of Chapter 7.

2.2.1 Assumptions about the power model

Stevens [111] identified four distinct types of scale arising from the various rules by which numerals can be assigned as measurements: nominal, ordinal, interval and ratio. The level of measurement restricts the types of operation which can be meaningfully performed on the data—and, by consequence, the types of statistic which are admissible. The power model approximation exploited by a given DPA strategy can be classified accordingly, depending on the assumptions implied by the distinguisher with which it is paired. The performance of a strategy will depend on the accuracy of the power model, for a notion of accuracy which is specific to the approximation type.

Profiled attacks (e.g. Bayesian classification-style templates [30] and stochastic profiling [97]) depend on a direct approximation of the deterministic data-dependent power consumption, corresponding to the ratio scale of [111]. This requirement is the most demanding possible, expressed as $M \approx L$. We say a distinguisher interprets a power model directly...
if it is not invariant to any transformation on the power model. 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: \( SS_{err}(M) = \sum_{z \in \mathcal{Z}} (M(z) - L(z))^2 \).

Less demanding is the requirement that \( M \) is a proportional approximation for \( L \): \( M \approx \alpha L \) (c.f. the ‘interval scale’ of [111]). We say a distinguisher interprets a power model proportionally if it is invariant to a linear transformation of the model values. Pearson’s correlation coefficient [19] is a popular choice for use in such a scenario: as a simple, moment-based statistic it can usually be estimated very efficiently with respect to the number of trace measurements required. (Formal efficiency guarantees for the sample estimator assume Gaussian noise, but it remains consistent and meaningful in most circumstances by the law of large numbers). A natural way of quantifying the accuracy of such a model is the correlation itself:

\[
Corr(M, L) = \frac{\sum_{z \in \mathcal{Z}} (M(z) - E[M(z)])(L(z) - E[L(z)])}{\sqrt{\sum_{z \in \mathcal{Z}} (M(z) - E[M(z)])^2 \sum_{z \in \mathcal{Z}} (L(z) - E[L(z)])^2}}.
\]

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 [111]). We say a distinguisher interprets a power model ordinally if it is invariant to any monotonically increasing transformation of the model values; for example, a variant of correlation DPA using Spearman’s rank correlation coefficient, as proposed in [11]. As above, the accuracy of the model can be quantified via the rank correlation itself, which is computed by replacing the values with the rankings of the values in the above formula.

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 [111]). We say a distinguisher interprets a power model nominally if it is invariant to an arbitrary re-labelling of the model values. In fact, these correspond to the ‘partition-based’ distinguishers of [106]. Typical examples include statistics which are used to compare arbitrary distributions, such as mutual information [49] the Cramér–von Mises criterion [117] and the Kolmogorov–Smirnov test statistic [117, 123].
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., [65]).

\[
\begin{align*}
\text{Precision}(M) &= \mathbb{P}(L(z) = L(z')|M(z) = M(z')), \\
\text{Recall}(M) &= \mathbb{P}(M(z) = M(z')|L(z) = L(z')).
\end{align*}
\]

### 2.2.2 Assumptions about the noise

In Bayesian classification-style template attacks the distribution of the noise is fully characterised so that any data-dependency in the *variation* may be optimally exploited. At the other end of the scale, where the attacker is not prepared to make *any* assumptions about the noise distribution, nonparametric statistics such as mutual information and the Kolmogorov–Smirnov test may come in handy. Most cases fall somewhere in the middle: it is often reasonable to assume that the noise is data-independent and Gaussian (at least approximately), in which case a broad range of (semi-)parametric options become available. These are to be preferred where possible because they are inherently less costly to estimate [45].

However, in practice it is not the *true* shape of the noise which impacts on the estimation but the distribution of the residuals after the assumed power model is taken into account. This is rather problematic: it means that even if the *actual* noise is ‘well-behaved’ as above, the unexplained variance in the leakage may be data-dependent if the power model is not accurate. This would inevitably bias the estimates. Of course, under an incorrect hypothesis the power model is assumed to be inaccurate and it is precisely this feature which motivates DPA; in such cases we *expect*—even *rely on*—meaningless estimates.

This is a sharp reminder of the somewhat (unavoidably) heuristic nature of DPA. Many rigorous results in statistical theory cannot be applied to the quantities we are attempting to estimate—particularly those under an incorrect hypothesis. It is important to note

\[^2\text{The classification theory literature more frequently states these definitions in terms of ratios of counts—practically convenient but less directly translatable across contexts. See [54] for a more explicit probabilistic interpretation; though in our case we are, of course, averaging over multiple classes.}\]
that the statistical artefacts introduced into the estimated distinguishing vector may affect
DPA outcomes in either direction: the goal of DPA is not precise estimates but a separation between the estimate produced under the correct hypothesis and those produced under the alternatives. Imprecision in the estimates could exaggerate the underlying theoretical distinguishing margin, or it could understate it. Unfortunately, this behaviour depends not only on the statistic chosen as distinguisher but also (and interactively) on the precise leakage scenario.

2.3 ‘Generic’ DPA: making no assumptions

A generic strategy has been informally understood as one which does not depend on prior knowledge about the data-dependent leakage. In practice, to make no assumptions about the data-dependent leakage implies assigning 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 $F_k$: $[x]_k = F_k^{-1}[F_k(x)] \forall x \in X$.

Definition 1. The generic power model associated with key hypothesis $k \in K$ 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 §2.2.1 above) which interpret hypothesis-dependent predictions as an approximation up to nominality of the data-dependent leakage.

Definition 2. A distinguisher is generic-compatible if it interprets the generic power model nominally (that is, as an approximation up to nominality of the data-dependent leakage).

This provides valuable clarification on previous work such as [12], 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 1 and 2 together give rise to a natural notion of a ‘generic strategy’:

**Definition 3.** A generic strategy pairs the (hypothesis-dependent) generic power models with a generic-compatible distinguisher in order to perform DPA.

However, as previous work on ‘partition-based’ distinguishers (separately, e.g. [49, 119, 123], and collectively [106]) has consistently noted, not all (indeed, not many) scenarios are suited to a generic strategy.

### 2.3.1 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 more accurate 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—indisputably 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 §2.2.1:
\begin{align*}
\text{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]
\end{align*}

\begin{align*}
\text{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{align*}

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 can succeed against an injective target function.

**Proof.** If the $F_k$ are injective then $\forall x \in \mathcal{X}$, $F_k^{-1}[F_k(x)] = F_k^{-1}[F_k^*(x)] = \{x\}$. Nominal accuracy therefore does not vary by key hypothesis—no generic-compatible distinguisher can be expected to separate any two key candidates. \hfill \Box

Indeed, all of the known generic-compatible distinguishers, from the seminal CHES ’08 paper on the use of MI [49] to the recent copula-based method presented at Crypto ’11 [119], 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 [106]. 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.
Non-injectivity 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 XOR key addition. It then becomes fairly straightforward to draw out such function characteristics as will obstruct a generic strategy.

Propositions 2 to 4 are increasingly general: the order in which they are presented provide an intuitive progression of insight into the operation of generic DPA but the proofs follow naturally in reverse order and so are presented after all three have been stated.

**Proposition 2.** If \( F \) is affine then no generic strategy is able to distinguish the correct key \( k^* \) from any \( k \in \mathcal{K} \setminus \{k^*\} \).

The distinguishing vector produced by such an attack would be flat and maximal across all hypotheses, just as in the case of an injective target.

**Proposition 3.** If \( a \in \mathbb{F}_2^n \) is a linear structure of \( F \) then no generic strategy is able to distinguish between \( k^* \) and \( k^* \oplus a \).

That is, \( k^* \oplus a \) 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 \( k^* \oplus a \) [19]; [88], notes a corresponding phenomenon for correlation DPA.

**Proposition 4.** 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 \( k^* \oplus a \).

In other words, \( k^* \oplus a \) 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 [108]. Our observation is consistent with, and further illuminates, past works such as

\[47\]
which recognised the unusual operation of DPA distinguishers confronted with this particular S-Box/offset combination.

**Proof.** (Of Proposition 4). 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 $\text{Precision}(F_k) = \text{Precision}(F_{k^*}) = 1$ and $\text{Recall}(F_k) = \text{Recall}(F_{k^*})$ 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 $c : \mathbb{F}_2^n \rightarrow \mathbb{F}_2^n.$

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

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

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

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

Proposition 3 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 $c \in \mathbb{F}_2^n$ constant over all $x.$

Proposition 2 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 $a \in \mathbb{F}_2^n \setminus \{0\}$ (and thus for all $k \in \mathcal{K} \setminus \{k^*\} \subseteq \mathbb{F}_2^n$) by the same argument.

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

**Proposition 5.** Suppose $F$ is a balanced, non-injective $n$-$m$ 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.

**Proof.** (Of Proposition 5). That the condition is necessary follows directly from Proposition 4. Now suppose that, additionally, $L$ is injective.

Choose $x' \in \mathbb{F}_2^n$ such that $\#D_k \oplus k F(F^{-1}[F(x')] \neq 1$—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:

$$D_k \oplus k F(x' + k) \neq D_k \oplus k F(x'' + k)$$

$$\Rightarrow F(x' + k \oplus k^* + k) \oplus F(x' + k) \neq F(x'' + k \oplus k^* + k) \oplus F(x'' + k)$$

$$\Rightarrow F(x' + k^*) \oplus F(x' + k) \neq F(x'' + k^*) \oplus F(x'' + k)$$

$$\Rightarrow F_{k^*}(x') \oplus F_k(x') \neq F_{k^*}(x'') \oplus F_k(x'')$$

$$\Rightarrow F_{k^*}(x') \neq F_{k^*}(x'') \quad \text{(since $x'' \in F_k^{-1}(F_k(x'))$)}$$

$$\Rightarrow x'' \notin F_k^{-1}[F_{k^*}(x')]$$

$$\Rightarrow F_k^{-1}[F_{k^*}(x')] \neq F_k^{-1}[F_k(x')]$$

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 $\text{Precision}(F_{k^*}) = \text{Recall}(F_{k^*}) = 1$ whilst $\text{Precision}(F_k) = \text{Recall}(F_k) < 1$, so that a sufficiently sensitive generic-compatible distinguisher will be able to reject the hypothesis $k$. □

Proposition 5 can be 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 following toy example demonstrates that we can no longer claim sufficiency if \( L \) is non-injective:

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

\[
F(x) = \begin{cases}
  0, & x \in \{0, 3\} \\
  1, & x \in \{1, 2\} \\
  2, & x \in \{4, 5\} \\
  3, & x \in \{6, 7\},
\end{cases}
\]

\[
L(z) = \begin{cases}
  1, & z \in \{0, 1\} \\
  2, & z \in \{2, 3\}.
\end{cases}
\]

So \( F_0(x) = F(x \oplus 0) = F(x) \)

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) \( F_0^{-1}[F_0(0)] = \{0, 3\} \neq \{0, 1\} = F_4^{-1}[F_4(0)] \), whilst \( F_0^{-1}[L^{-1}[L \circ F_0(0)]] = \{0, 1, 2, 3\} = F_4^{-1}[L^{-1}[L \circ F_4(0)]] \supset F_4^{-1}[F_4(0)] \); 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 \( \text{Precision}(M_4) = \text{Precision}(M_0) = 1 \) and \( \text{Recall}(M_4) = \text{Recall}(M_0) \), implying that key candidates 0 and 4 cannot be distinguished from one another.

**Implications for practice**  It is an explicit design goal that S-Boxes should have high differential entropy [15]; 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 \( D_a F \) partitions \( \mathbb{F}_2^n \), the fewer the possible refinements into \( 2^m \) (balanced) parts. Therefore, among the (already restricted)
class of non-injective 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.

Our observations about the feasibility of generic DPA share similarities with previous work such as the efforts by Prouff [88] to relate the efficiency of correlation DPA (using a Hamming weight power model) to the cryptographic properties of the target function in the case that the device leaks the Hamming distance from some unknown, constant, reference state. The ‘transparency order’ of an S-Box increases as the coordinate-wise derivatives approach uniformity, with the implication that Hamming distance leakage of a cryptographically strong S-Box is inherently vulnerable to correlation DPA (see also [25]).

2.3.2 ‘Near-generic’ DPA and arbitrary power models

Previous work [12, 49, 80, 106] proposed using the 7 least significant bits (7LSB) of the AES S-Box output as an ‘arbitrary’ power model to achieve ‘generic’ MI-based DPA against injective target functions. Such an approach has been found to produce successful attacks sometimes, but not always. A surprising discovery of [122] is that attacks using the 7LSB model against Hamming weight leakage of the AES S-Box only become theoretically feasible when the noise magnitude reaches a certain threshold! The model is demonstrably not accurate for $L$ in this case: the classes it produces are pairs of values which always differ in Hamming weight by 1, so that:

\[
\text{Precision}(7\text{LSB}_{k^*}) = \mathbb{E}_{x \in \mathbb{F}_2^8} \left[ \frac{\#F_{k^*}^{-1}[L^{-1}[L \circ F_{k^*}(x)] \cap 7\text{LSB}_{k}^{-1}[7\text{LSB}_{k}(x)]]}{\#7\text{LSB}_{k}^{-1}[7\text{LSB}_{k}(x)]} \right] = \frac{1}{2},
\]

and \[
\text{Recall}(7\text{LSB}_{k^*}) = \mathbb{E}_{x \in \mathbb{F}_2^8} \left[ \frac{\#F_{k^*}^{-1}[L^{-1}[L \circ F_{k^*}(x)] \cap 7\text{LSB}_{k}^{-1}[7\text{LSB}_{k}(x)]]}{\#F_{k^*}^{-1}[L^{-1}[L \circ F_{k^*}(x)]]} \right] = \sum_{i=0}^{8} \frac{\binom{8}{i}}{2^8} \frac{1}{\binom{8}{i}} = \frac{9}{256} = 0.035.
\]

However, when we measure side-channel leakage, we don’t observe $L$ directly, rather a noisy version $L + \epsilon$. In very low levels of noise the preimages of $L$ produce discernibly different trace measurements (see, for example, the first panel of Figure 2.2); but, because
the 7LSB model pairs values with adjacent Hamming weights, increasing amounts of noise distort the trace distribution towards the model (as can be seen in the second and third panels of Figure 2.2). That is, given the observed evidence, it becomes less clear that the values paired by the model do not produce identical leakage signals.

Of course, DPA outcomes depend not just on the accuracy of the correct key model but on the relative inaccuracy of the models under alternative hypotheses. With this in mind, Figure 2.2 can help to even more fully illuminate the noise-dependence of 7LSB strategies. The solid red line, and the dotted black and red lines, depict examples of conditioned trace distributions for a correct-key model pairing and two incorrect-key model pairings respectively, when the device leaks the Hamming weight.

Correct-key pairings in this common but highly specific example scenario always take adjacent deterministic leakage function values, so when the signal is very strong (as in the first panel), the inaccuracy of the model is evident from the distribution. By contrast, the cryptographic properties of the AES S-Box have the effect that values paired by an alternative-key model are not always adjacent in the leakage function (as per the dotted blue lines) and may actually coincide (as per the dotted red lines). Non-adjacency does not affect apparent accuracy since adjacent components are distinguishable anyway, whilst agreement actually increases overall precision! Hence the catastrophic failure of the 7LSB model in such scenarios—the incorrect hypotheses can produce seemingly more accurate models than the correct hypothesis. (Indeed, [122] showed that MI distinguishing vectors are minimised—by a substantial margin—under the correct key until the noise reaches a certain threshold: see Chapter 4 for more).

As the signal degrades, adjacent components become increasingly mixed so that the perceived accuracy of the correct-key model increases. As for the alternative-key models, conditional distributions comprising non-adjacent components will require higher levels of noise before mixing occurs, whilst noise in the single component conditional distributions will diminish their contribution to precision. From these concurrent developments arises the potential for 7LSB-based strategies to eventually distinguish the correct key, supposing the perceived correct-key accuracy overtakes the perceived alternative-key accuracies (which will depend at least in part on the properties of the target function).

By contrast, when the device does not leak the Hamming weight, values which are paired
under the correct key no longer necessarily correspond to adjacent (close) values in the true leakage function, as for example in Figure 2.3. In such cases, it is far less clear what will happen to the distinguishing vector as signal strength varies; more noise is required for the components of the conditional distributions to mix, keeping the conditional information high (and the MI low). Key distinguishability will be highly sensitive to the combinations of conditional distributions produced under the correct and incorrect hypothesis.

From the above examples it becomes clear that a 7LSB strategy can only be successful when it amounts to coincidental use of meaningful prior knowledge—that is, the supposedly arbitrary choice of model in fact turns out to correspond to something apparent in the leakage measurements. This is the case for noisy Hamming weight leakage, but is by no

---

*For illustrative purposes, we here imagine a highly nonlinear leakage function—the Hamming weight of the AES S-Box transform.*
means true in general. For this reason, it cannot be considered ‘generic’. (See Chapter 4 (and [122]) for further investigation of the 7LSB model in different scenarios, by which we clearly see how leakage function-dependent it is). We would expect any power model based on an arbitrary compression of the intermediate value to exhibit similar behaviour; that is, to produce successful outcomes if and only if it happens to sufficiently coincide with reality after all. We conclude that there is no injective-target counterpart to the genuinely generic strategies against non-injective targets.

More generally, such an incidence highlights the difference between absent knowledge (which is what the generic power model is supposed to encapsulate) and incorrect knowledge (which the 7LSB model could be described as). That a certain statistic remains effective when information is taken away does not justify the expectation that it will remain effective when it is provided with wrong information. This subtle distinction has, however, been repeatedly overlooked in discussions about candidate ‘generic’ distinguishers.

### 2.4 Impact of model type on practical outcomes

Setting aside questions of model accuracy it is natural to ask what impact model type itself has on DPA outcomes. It is well established that there is a trade-off between flexibility and efficiency when it comes to statistical estimation [45], so that as fewer and fewer assumptions are made, the amount of data required to obtain equivalently precise estimates increases. This implies that ascending measurement levels will facilitate increasingly efficient DPA attacks, as long as appropriate distinguishing strategies are chosen.

As a simple illustrative experiment, we consider the success rates of attacks against arbitrary leakage of the (first) DES and the AES S-Boxes, using known power models but of descending measurement level. The leakages are simulated according to the following rules:

- The deterministic part is randomly chosen to take values between 0 and \((m+1)\times 4\) such that the cardinality of the leakage function image \((L(\mathbb{F}_2^m))\) is the same as that
of the Hamming weight function, i.e. 5 in the case of DES \((m = 4)\) and 9 in the case of AES \((m = 8)\). This produces leakages which are comparable to a known typical behaviour in the extent to which they compress the target function output space but less constrained in that leaked values are not adjacent or evenly spaced.

- We allow the random part of the leakage \(\varepsilon\) to be data-dependent also by associating to each value in the range of the target a Gaussian fluctuation of a particular standard deviation. These are assigned randomly but such that the average signal-to-noise ratio (SNR) is the same in each experiment.

We experiment with two different (average) SNRs (a ‘low’ (0.0125) and a ‘high’ (4)) by simulating 500 different such leakages and incrementally launching the various attacks to determine the ‘success’ thresholds in each instance. We choose, for each model type, the most straightforward appropriate strategy: Bayesian classification for ratio-level models (with fully characterised noise), Pearson’s correlation coefficient for interval-level models, Spearman’s rank correlation for ordinal-level models, and the 2-sample Kolmogorov–Smirnov test statistic for the nominal-level models. This last one is not, perhaps, so widely considered in the literature as, say, mutual information, but it has the advantage that it is unambiguous in its estimation methodology, so that we do not need to take into account user-defined choices such as bin-width.\(^4\) (The need to select particular strategies cannot be avoided for the purposes of experimentation, and our results are intended to be indicative rather than conclusive: better outcomes may be achievable in any of the categories). For the nominal level attacks, we consider both the known nominal model and the generic model according to Definition 1. We also consider the ‘near-generic’ 7LSB model which we describe in §2.3.2 above.

Figures 2.4 and 2.5 show the results of these experiments. As we expect, the Bayesian attacks are most efficient, closely followed by Pearson’s correlation and then Spearman’s rank correlation. In these instances, the size of the differences are not likely to make much practical difference. That is, an attacker with only an accurate ranking of the true leakages does not appear very disadvantaged next to an attacker with templates. The gap between the ordinal-model results and the nominal-model results is larger, (though

\(^4\)See chapter 4 for a full explanation and evaluation of the Kolmogorov–Smirnov and the mutual information based distinguishers.
of course could be narrowed by an improved strategy), but still the disadvantage is not prohibitive. At the higher noise level the advantage to the Bayesian strategy appears more marked and the gap between interval- and ordinal-based correlation strategies narrows.

There is a further substantial loss of efficiency between the nominal strategy with the known model and the nominal strategy with the generic model, as we would fully expect because the latter is not as accurate as the former. However, generic attacks on the DES S-Box remain consistently successful against these arbitrarily constructed leakages, given sufficient data. By comparison, the so-called ‘near-generic’ attacks using the 7LSB do not succeed against the (injective) AES S-Box: as we showed in §2.3.2, their effectiveness is derived from the coincidental meaningfulness of the 7LSB in some scenarios (i.e. Hamming weight leakage with adequate noise, and similar). We would, as such, not expect them to be useful in attacking arbitrary leakage functions; this we have now confirmed experimentally.

![Figure 2.4: Example success rates for attacks against the first DES S-Box using power models of differing measurement level.](image1)

![Figure 2.5: Example success rates for attacks against the AES S-Box using power models of differing measurement level.](image2)
2.5 Conclusion

In this chapter we have differentiated between the types of prior knowledge about the device leakage which might be available to an attacker. In particular, the power model for the deterministic data-dependent leakage can operate on one of four widely-recognised measurement scales, namely (in descending order) ratio, interval, ordinal and nominal. Each permits different operations and so must be combined with appropriate statistics as distinguishers. Model type has a bearing on DPA outcomes independently of model accuracy, because statistics requiring fewer assumptions are more data-costly to estimate.

This delineation has enabled us to clarify precisely what qualifies a power model as ‘generic’, and to explore the distinguishers and scenarios which are compatible with such a definition. In particular, we have been able to conclusively demonstrate that non-injectivity 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 (additional) 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.

We have further considered one frequently suggested work-around for achieving (informally) ‘near-generic’ key recovery against injective functions, which is to arbitrarily truncate the target output (usually by removing the most significant bit). We have concluded that this only works if and when the arbitrary choice happens to correspond, in some meaningful way, to the trace measurements after all—and therefore that example suc-
cesses in the literature do not qualify as ‘generic’.

By way of illustrating the impact of model type (as opposed to accuracy) we have conducted experiments against simulated arbitrary leakage of the first DES and the AES S-Boxes. We have confirmed the intuition that attack efficiency diminishes as the measurement level downgrades, due to the trade-off between flexibility and efficiency in estimation theory. We have moreover confirmed the consistent efficacy of generic strategies against non-injective targets and the general inefficacy of so-called ‘near-generic’ strategies for use against injective targets.
Chapter 3

A theoretic evaluation framework for differential side-channel distinguishers

3.1 Introduction

The ability to make meaningful comparisons between side-channel distinguishers is important both to attackers seeking an optimal strategy and to designers wishing to secure a device against the strongest possible threat. Projects such as the NIST Non-Invasive Attack Testing Workshop [81] have been working towards the development of industry-wide, recognisable test methods and metrics suitable for incorporation into security standards and for providing realistic reassurances to potential end-users.

Perhaps the most natural indicator for the strength of an attack (or, from another perspective, the vulnerability of a device) is the number of trace measurements required for an attack to be successful. Attempts to approximate sample sizes have played a prominent part in evaluations from the very earliest DPA literature [63]. Such information equips designers to make risk assessments about the capability of an attacker to acquire or process the required amount of data. Depending on the nature of the application, it could even be used to determine an appropriate upper limit on the number of encryptions/protocol
runs permitted under any given secret key. Or, it could be used to tune countermeasures to an acceptable cost/security trade-off.

Motivated by this notion are the \((o\)-th order) success rate and guessing entropy of [108], defined (respectively) as the probability of \((o\)-th order) success and the expected number of candidates remaining to test, as the sample size varies. They are well-established as a useful basis on which to compare the practical efficiency of different DPA strategies against different leakage scenarios. However, they are ill-suited to measuring the inherent capabilities of a particular distinguisher because they are estimator specific: that is, they relate to the practical instantiation of a distinguisher via a given estimation procedure. This does not really matter for simple, moment-based statistics such as Pearson’s correlation coefficient, for which the sample correlation is a ‘well-behaved’ estimator under fairly minimal assumptions. But for other, more complex quantities—particularly nonparametric statistics such as mutual information—there is usually no obvious ‘best’ choice of estimator and performance is known to vary widely [84]. As such, estimator-based evaluations do not allow conclusive statements about the underlying capabilities of the distinguishers, because they offer no guarantees that results would not be improved by changes to the estimation procedure.

Another drawback of metrics relating sample size to practical outcomes is that direct computation—via the techniques of statistical power analysis [66]—requires being able to at least approximate the sampling distributions of the estimator in the case that the correct key is selected and in the case that one of the alternatives is selected. Some extra simplifying assumptions make this possible for correlation DPA [70], but it is rarely achievable in general. Most of the time sample sizes, success rates and guessing entropies must be obtained experimentally, on a case-by-case basis, by performing practical attacks against simulated or measured traces.

In recognition of these limitations we propose an alternative (or complementary) evaluation framework which is based directly on the theoretic quantities taken by a distinguisher under given leakage conditions, bypassing altogether the confounding problem of estimation. Prouff et al. in [89] recognise the crucial distinction between the theoretic and practical factors influencing attack success, and devote some discussion to conditions under which an attack is theoretically plausible. We extend their notion of plausibility to
develop various measures of *distinguishing strength* by which to assess and compare a variety of attack methodologies and scenarios. This facilitates evaluations which are like-for-like in comparing the underlying capabilities of different strategies. But it does have its own drawbacks: different statistics are differently problematic and costly to estimate, so that theoretic comparisons will not necessarily translate into the practical realm. However, the outcome measures we select are motivated by statistical power analysis and can be demonstrated to be highly relevant to practical performance.

In what follows we review existing methods for evaluating and comparing DPA attacks (§3.2) before explaining the advantages of a theoretical perspective and presenting our proposed framework (§3.3).

### 3.2 Existing evaluation methods

The notion of *key-recovery success* was formalised by Standaert et al. in [108]. The theoretic attack distinguisher is \( D = \{ D(k) \}_{k \in K} = \{ D(L \circ F_k^*(X) + \varepsilon, M \circ F_k(X)) \}_{k \in K} \), where the plaintext input \( X \) takes values in \( \mathcal{X} \) according to some known distribution (usually uniform). We say the attack is *theoretically successful* if \( D(k^*) > D(k) \) \( \forall k \neq k^* \). We say it is *o-th order theoretically successful* if \#\{\( k \in K : D(k^*) \leq D(k) \}\} < o.

However, in practice \( D \) must be estimated. Suppose we have observations corresponding to the vector of inputs \( x = \{x_i\}_{i=1}^N \), and write \( e = \{e_i\}_{i=1}^N \) to be the observed noise (i.e. drawn from the distribution of \( \varepsilon \)). Then the size \#\( K \) estimated vector is \( \hat{D}_N = \{ \hat{D}_N(k) \}_{k \in K} = \{ \hat{D}_N(L \circ F_k^*(x) + e, M \circ F_k(x)) \}_{k \in K} \). We then say the attack is *successful* if \( \hat{D}_N(k^*) > \hat{D}_N(k) \) \( \forall k \neq k^* \) and *o-th order successful* if \#\{\( k \in K : \hat{D}_N(k^*) \leq \hat{D}_N(k) \}\} < o.

### 3.2.1 Number of traces required for key recovery

*Statistical power analysis*\(^1\) [66] studies the relationship between sample size, population quantities and the *power* of statistical tests in frequentist inference. It is designed to answer questions about any one of these three aspects given information on the other

\(^1\)The ‘power’ in this context relates to the power of a statistical hypothesis test to (loosely speaking) detect a genuine difference between two quantities, and has nothing to do with a device’s *power consumption* viewed as a side-channel.
two, for tests of a given confidence level $\alpha$. For example, if you want to be able to detect, with probability $1 - \beta$ (the ‘power’), whether there’s a difference of at least $\nu$ standard deviations (the ‘effect size’) between two population quantities, power analysis can be used to determine the number of measurements $N$ you will need to collect (the ‘sample size’) in order to produce a conclusive $\alpha$-level test. Conversely, it can be used to determine the power of a $\alpha$-level test with a given effect size and sample size, or the effect size detectable by an $\alpha$-level test with a given power and sample size. However, these computations all require at least being able to approximate the sampling distributions of the test statistic (in both the presence and absence of a real effect).

In the case of DPA we would like to know how many traces we need to measure before the value associated with the correct key dominates in the estimated distinguishing vector. That is, we want to compute $N^*$ s.t. $N > N^* \Rightarrow \hat{D}_N(k^*) > \hat{D}_N(k) \forall k \neq k^*$. Each pairwise comparison can be thought of as an implicit, informal hypothesis test $H_0$: $\hat{D}_N(k_1) - \hat{D}_N(k_2) = 0$ vs. $H_1$: $\hat{D}_N(k_1) - \hat{D}_N(k_2) > 0$. Having some idea of how large $D(k^*)$ and $D(k)$, $k \neq k^*$ are will help in choosing an appropriate effect size. A common choice for the power is 80%. The remaining ingredients for a statistical power analysis are the sampling distributions for $\hat{D}_N(k^*)$ and $\hat{D}_N(k)$, $k \neq k^*$.

Since the Fisher transformation of the sample correlation coefficient $z = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$ is approximately normally distributed (with mean $\frac{1}{2} \ln \frac{1+\rho}{1-\rho}$ and standard deviation $1/\sqrt{N-3}$), the number of traces needed to distinguish an estimate of the ‘correct key’ correlation $\rho_{ck}$ from an estimate of the correlation under some rival key hypothesis $\rho_{rk}$ (at a confidence level of $\alpha$) can be computed as follows [70]:

$$N^* = 3 + 8 \frac{z_{1-\alpha}^2}{\left( \ln \frac{1+\rho_{ck}}{1-\rho_{ck}} - \ln \frac{1+\rho_{rk}}{1-\rho_{rk}} \right)^2}$$

In general, however, the required distributions are not known: for example, mutual information, as a functional of probability distributions, is notoriously problematic to estimate and the sampling distribution is highly sensitive to the choice of estimation procedure and to the true underlying distribution [84]. Even the two-sample Kolmogorov-Smirnov test statistic, which has a known distribution under the ‘null hypothesis’ ($H_0$) that two distributions are independent, does not yield a sample size approximation because we do not
know the distribution under the ‘alternative hypothesis’ ($H_1$) of dependence (which relates to the correct key distinguisher value). The type of neat result obtained for the simple, moment-based correlation coefficient is out of reach for these more complex methods.

### 3.2.2 Success rate and guessing entropy

Standaert et al. [108] propose to evaluate key-recovery attacks via the (o-th order) success rate and the guessing entropy: respectively, the probability of (o-th order) success and the expected number of key candidates remaining to test after a practical attack with a given number of traces.

\[
SR(\hat{D}_N, o) = \mathbb{P}(\#\{k \in \mathcal{K} : \hat{D}_N(k^*) \leq \hat{D}_N(k)\} < o)
\]

\[
GE(\hat{D}_N) = \mathbb{E}[\#\{k \in \mathcal{K} : \hat{D}_N(k^*) \leq \hat{D}_N(k)\}].
\]

Note that these metrics are usually estimated from practical experiments against simulated or measured traces, and as such are necessarily estimator specific. In the case of moment-based distinguishers such as the simple difference-of-means and the correlation coefficient, the drawbacks are not immediately apparent; natural, ‘good’ estimators are usually readily available (for example, the sample mean and the sample correlation, which are consistent and meaningful in most circumstances by the law of large numbers, even if they are not always the most efficient estimator available).

However, nonparametric statistics such as mutual information can be estimated in a variety of different ways and performance is known to be extremely sensitive to the chosen approach. The best estimator for a given scenario will depend on the (unknown) underlying distributions; there are no universal rates of convergence [84], so that whatever estimator we pick, we can always find a distribution for which the error vanishes arbitrarily slowly.

In short, there is no such thing as a universally ‘best’ estimator for any given distinguisher, by which to fairly measure its best-case capabilities in a given leakage scenario. This rather undermines comparisons based on practical experiments with simulated or
measured traces: perceived advantages/disadvantages are inconclusive as we do not know if they truly indicate inherent strengths/weaknesses of the distinguishers or merely arise from the choice of estimation procedure.

A common theme of the research into MI-based DPA, for example, has been the attempt to establish whether, by improved estimation techniques, it can achieve the much-sought-after advantage over correlation DPA [8, 117]. Prior to our evaluation framework (first published in [121]) we are unaware of any consistent attempt to establish whether (or when) this advantage is theoretically possible given the inherent attributes of the two distinguishers.

By evaluating theoretic attributes we are able to make comparisons which are more conclusive than those based on particular estimators. The differential burden of estimation incurred by different statistics limits the extent to which such comparisons can be supposed to directly translate into the practical realm; we are able to mitigate this to some degree by choosing outcome measures which are demonstrably indicative of practical performance, as we will explain in the next section.

### 3.3 Comparison framework

Our proposed approach towards like-for-like, estimator-independent comparisons between methodologies is to focus on the theoretic distinguishing vectors rather than estimated vectors from practical attacks. In this section, we will first clarify what we mean by ‘theoretic distinguishing vectors’ and describe how these can be computed. We will subsequently introduce a range of outcome measures based on those characteristics of the theoretic vectors which have the greatest bearing on the trace efficiency of a practical attack.

#### 3.3.1 Computing theoretic vectors

Statistical methods are widely used in many areas of study; however, outside the field of statistical theory the distinction between a statistic (which is a function of observed data) and a statistical parameter (an underlying quantity which is a feature of the population)
is not always properly understood. When we talk about an estimator, we are referring to a statistic which is used to estimate a particular population quantity—such as the mean of a distribution, or the correlation coefficient or mutual information between two distributions.

Hence our goal of evaluating the theoretic characteristics of DPA distinguishers, independently of estimation procedure, entails computing the population quantities for given distributions as opposed to estimating them from measured or simulated traces. A common approach in the literature has been to simulate traces according to given distributions and then plug these into formulae for estimators; the quality (loosely speaking, the ‘closeness to reality’) of the consequent estimates depends on the number of samples generated and the appropriateness of the chosen estimation procedure. But in simulated experiments the conjectured leakage can be fully characterised so that we can, in fact, directly obtain the population quantities, thus arriving at a conclusive picture of the asymptotic capabilities of a certain distinguisher in a certain leakage scenario.

Of course, in real-life attack scenarios the measurements arise from unknown distributions which we cannot control or observe directly. We cannot compute the underlying ‘theoretic’ distinguishing vector, but remembering that it is there helps us to understand the process of estimation (and any prior knowledge we do have may help us improve our estimates). It is, in a sense, ‘what we are aiming for’. Hence, by computing and evaluating theoretic distinguishers for a variety of hypothesised leakage scenarios (particularly those inspired by suspected or even profiled device behaviour) we are much better informed about what to expect in real life and about which DPA strategies may be most worth pursuing given our suspicions in a particular instance.

Arriving at the theoretical vectors is straightforward: For each possible input $x \in \mathcal{X}$ to the cryptographic function we obtain a vector evaluating the Gaussian density centered at the corresponding data-dependent leakage value $L \circ F_k^*(x)$ and having variance $\text{Var}(\varepsilon)$. The average of these vectors, weighted by the input probabilities $\mathbb{P}(X = x)$, then gives the probability density of the power consumption evaluated over the full range of possible leakage values. Conditional densities, corresponding to each possible prediction value $m \in \mathcal{M}$ under each key hypothesis $k \in \mathcal{K}$, are constructed similarly. From these probability densities we are able to directly compute (via numerical integration) the various moments,
entropies and cumulative probabilities which make up the formulae by which the various distinguishing statistics are defined.\textsuperscript{2}

\subsection*{3.3.2 Outcome measures}

Having established why it is useful to consider theoretic distinguishing vectors, and explained how to go about computing them, we now seek to provide a means of summarising relevant characteristics in such a way as to facilitate comparisons between different strategies in different attack scenarios. Recall, from §3.2, that what really matters in practice is how many trace measurements are required for an attack to be successful. As such, the outcome measures we propose are based on characteristics of the vectors having a direct impact on expected data complexity. Figure 3.2 displays the precise formulae; descriptions and rationale are provided below.

The first needs little explanation:

1. \textit{Correct key ranking}: The position of the correct key when ranked by distinguisher value. If the correct key is ranked joint first the \textit{ranking order} is the number of keys sharing position 1, so that an attack with a ranking order of $o$ is $o^{th}$-order theoretically successful as defined in §3.2. The relationship with practical efficiency is obvious: attacks which are not first-order successful will not be able to uniquely extract the correct key given \textit{any} number of trace measurements (except by random chance).

The theory behind \textit{statistical power analysis} tells us that, when estimating population quantities, the sample size required to detect a statistically significant difference increases as the actual magnitude of the true difference decreases. Therefore, of crucial relevance to practical attack outcomes are the \textit{theoretical} margins by which the true key is isolated from the remaining keys. This idea is illustrated in Figure 3.1 and is the motivation for the next three measures:

\footnote{This does involve a certain amount of numerical approximation where symbolic integration is not possible; care should be taken, therefore, to use sufficiently many evaluation points to achieve the required level of precision. In our analyses, we experimented to make sure that our reported results were robust to increases in the number of evaluation points; we were confident that the precision of our approximations far exceeded that required for the level at which it was informative to report them.}
2. **Relative distinguishing margin**: The distance between the correct key distinguisher value and the value for the highest ranked alternative, normalised by the standard deviation of the distinguishing vector so that scale-free comparisons can be made between different distinguishers in different leakage scenarios. (Note that it is zero for attacks with success orders greater than 1, and negative for failed attacks, where it gives further indication of the extent of the failure).

3. **Absolute distinguishing margin**: The relative margin allows us to summarise the shape of a distinguishing vector and how this responds to noise or scenario degradation. However, it disguises changes in the actual magnitude of the margin and the fact that this is more sensitive for some methodologies than for others. We need some way to take into account raw margin size as well as size relative to the vector as a whole, which is still scale-independent so that we can make like-for-like comparisons between distinguishers. We therefore report the ratio between the nearest-rival margin and that of the corresponding ‘optimal’ vector: the univariate equivalent in an optimistic (i.e. known Hamming weight power model) noise-free setting. This will allow us to comment meaningfully on the impact of model degradation and noise on the real size of the margins to be estimated.

4. **Standard score**: This is the same as the “DPA signal-to-noise ratio” described by [55]: the number of standard deviations above (or below) the mean, for the correct key distinguisher value. It provides a more general measure of the sensitivity of an attack in isolating the correct key. A theoretically ‘unsuccessful’ attack may still be able to return a small candidate subset containing the correct key if the standard score is high.

By computing the above measures for uniformly drawn plaintexts $X \leftarrow X^{unif}$, we are able to compare theoretic behaviour of attacks when provided with full information. We propose to explore the sensitivity of attacks to incomplete information by inspecting theoretic attack vectors as restricted on reduced subsets of the plaintext space: $D|_{X'}$ where $X' \subseteq X$. These vectors depend not only on the size but also on the composition of the input set $X'$; we cannot perform the computations exhaustively over the entire space of possible subsets (it is too large), but by repeated random draws of increasing size we can estimate the support size needed for theoretic success. We argue that this provides
insight into the relative data complexity of distinguishers and their particular limitations in small samples. We thus add the following measures (defined for theoretically successful distinguishers only):

5. **Average critical support**: On average, the required support size of the input distribution for the attack to achieve $o^{th}$-order success (where $o$ is the ranking order).

6. **Critical support for $100 \times p\%$ success rate**: The support size for which the rate of success (of the appropriate order) is at least $100 \times p$ per cent.

Our criteria are best viewed in conjunction with one another rather than in isolation, and trade-offs between them will interplay differently with practical considerations. For instance, a methodology which achieves only $o^{th}$-order success (where $o > 1$) might be preferable to one achieving $1^{st}$-order success if the distinguishing vector can be estimated more precisely and/or efficiently. Likewise, nearest-rival distinguishability may be more important than average critical support in the presence of high noise.

### 3.4 Conclusion

In this chapter we have identified a shortcoming of existing approaches to DPA evaluation, which only managed to capture the comparative performance of particular estimator-
True key ranking (and associated order):

$$\text{Rank}(D) = 1 + \sum_{k \in K} I_{\{D(k) > D(k^*)\}}$$

$$\text{RankOrder}(D) = \sum_{k \in K} I_{\{D(k) = D(k^*)\}}$$

Relative distinguishing margin:

$$\text{RelMarg}(D) = \frac{D(k^*) - \max_{k \neq k^*} \{D(k)\}}{\sqrt{\text{Var}\{D(k)\} | k \in K}}$$

Absolute distinguishing margin:

$$\text{AbsMarg}(D) = \frac{D(k^*) - \max_{k \neq k^*} \{D(k)\}}{D(L \circ F_k(X), L \circ F_k(X)) - \max_{k \neq k^*} \{D(L \circ F_k(X), L \circ F_k(X))\}}$$

Standard score:

$$\text{StdScore}(D) = \frac{D(k^*) - \mathbb{E}\{D(k) | k \in K\}}{\sqrt{\text{Var}\{D(k)\} | k \in K}}$$

Critical support:

$$\text{AveSupp}(D) = \mathbb{E}[\min\{\#\mathcal{X}' | \mathcal{X}' \subseteq \mathcal{X} \land \text{Rank}(D|_{\mathcal{X}'}) = 1\}]$$

$$\text{PctSupp}(D, p) = \min\{\#\mathcal{X}' | \mathcal{X}' \subseteq \mathcal{X} \land \mathbb{E}[I_{\{\text{Rank}(D|_{\mathcal{X}'}) = 1\}}] = p\}$$

**Figure 3.2:** Formulae for the outcome measures.

Specific practices. To counter this, we propose an evaluation framework which makes a nuanced account of the theoretic attributes of a distinguisher, independent of any particular estimation procedure.

Our methodology entails summarising the features of a distinguishing vector which most contribute to the precision required to produce a successful attack in the practical estimation stage—indicative of the number of trace measurements a real-world attacker would need to sample. This approach has not been considered previously in the side-channel literature, where the difference between estimating attack outcomes from simulated traces and numerically approximating the theoretical quantities from the (known) density functions has tended to be overlooked.

All of the following chapters make use of aspects of this framework; it is particularly useful when evaluating nonparametric methodologies as general results about sampling distributions are hard to come by.
Chapter 4

Mutual information and other nonparametric distinguishers

4.1 Introduction

One response to the ‘generic DPA’ challenge has been to seek out nonparametric (‘distribution-free’) statistics, such as mutual information (MI) [49], the Kolmogorov–Smirnov (KS) test statistic [117, 73] and the Cramér–von Mises (CvM) criterion [117] for use as distinguishers. The reasoning behind this is that nonparametric statistics require fewer assumptions on the underlying distribution than do parametric statistics (such as the moment-based Pearson’s correlation coefficient). Thus the approach is well-motivated up to a point but, in the absence of a formal understanding of what it means for a strategy to be generic, the role of the power model has become somewhat overlooked. In fact, these methods are often presented as generic in the sense that they should be robust to incorrect or naive power models rather than in the sense that they are able to operate with the generic power model. We have already argued in Chapter 2 that such reasoning is not justifiable as, in fact, any successful distinguishing strategy relies on a power model which captures something meaningful about the true leakage. One priority of this chapter is therefore to explain and re-frame MI-, KS-, and CvM-based DPA in the light of our formal reasoning about the role and nature of the power model and our proposed understanding of what it means for an attack to be generic.
Moreover, the trade-off between flexibility and efficiency in statistical methods [45] has at times been overlooked, or inadequately understood, so that the preferability of parametric methods when appropriate assumptions do hold has not been properly anticipated. In particular, part of the initial appeal of using the MI was that, as a population quantity, it represents an information-theoretically optimal counterpart to linear correlation. This prompted the conjecture that it might be a more efficient means of exploiting the available data, and so require fewer measurements. In reality, the greater data-complexity of actually estimating the MI makes this extremely unlikely in any situation where correlation DPA is also feasible, even in scenarios where the knowledge on the leakage was not perfect, as reflected in experimental studies [12, 49, 89, 117]. Several studies [8, 89, 117] have highlighted the sensitivity of MI-based DPA to the estimation methods used and have attempted to improve outcomes by finding better estimators. However, they have not first verified whether or not a theoretic advantage exists; that is, if the underlying quantities do not themselves distinguish the key more strongly than the population correlation coefficients, then no estimation method will achieve the desired practical advantage. This open question points towards the usefulness of our evaluation framework of Chapter 3; by its application to the theoretic MI vectors in different scenarios we are able to re-assess the a priori reasoning of information-theoretic optimality and conclusively confirm that the persistent under-performance reported by empirical studies arises from the complexity of estimation and not any underlying weakness of the estimated quantity itself.

Motivated by the importance of not introducing more flexibility than is needed, we recall another distinguisher, interesting because it is generic-compatible with respect to the power model but relies on assumptions with respect to the noise: the variance-ratio (VR) proposed in [106]. In the terminology of [106], it belongs to the class of ‘partition-based’ distinguishers—which, as discussed in Chapter 2, is equivalent to saying that it interprets the power model as a nominal approximation for the deterministic data-dependent leakage. However, it is not nonparametric, as it assumes the model errors are independent, identically-distributed Gaussian noise. We explore the theoretic performance of this distinguisher as an efficiently-estimated alternative to nonparametric methods.

Having established, then, that there is an overlap between nonparametric distinguishers and ‘generic DPA’, but that the one does not imply the other (since nonparametric dis-
tinguishers can be used in combination with an informed power model, and parametric methods exist which do not require such knowledge) the latter part of this chapter is concerned with evaluation and comparison of all five different proposed methods, as applied ‘generically’ as well as in combination with prior knowledge. Previous such studies have tended to rely on experimental comparisons alone which, as we discussed in Chapter 3, are necessarily estimator-specific and unable to yield conclusive statements about the inherent theoretical capabilities of the distinguishers (i.e. the estimands) themselves. Our analysis places the distinguishers on a like-for-like footing, independently of estimator choices so that our comparisons are in some sense conclusive, though with the limitation that theoretic advantages may not translate into practical advantages because of the disparity in estimation requirements.

The nonparametric methods we test are found to display a surprising sensitivity to noise, resembling a type of stochastic resonance, which can even play a critical role in determining the theoretic success or failure of the distinguishers in certain scenarios. This is by contrast with correlation DPA which is only affected by noise at the practical level—that is, the sample size required for precise estimation increases with noise but the underlying ratios between the correct key vector value and the rest of the vector (for a given target function) depends only on the relationship between the power model and the deterministic component of the leakage, and thus remains unchanged.

This noise sensitivity proves especially significant in the case of the ‘near-generic’ attacks described by the authors of [49], who propose to use the 7 least significant bits (7LSB) as a power model against injective 8-bit target functions. Although experimentally verified in subsequent investigations such as [80], in Chapter 2 we showed that they were (by stark contrast with ‘generic’ attacks against non-injective targets) highly dependent on the true underlying leakage distribution. In this chapter we provide a detailed theoretical-level evaluation of near-generic attacks in various scenarios, discovering that, even when the leakage function itself is compatible, a certain amount of noise is needed in order to achieve theoretic success: key recovery fails comprehensively in strong-signal settings. It is clear that, on the rare and particular occasions when they do succeed, they do not supply the hoped-for advantages displayed by generic attacks against non-injective targets.

Our analysis emphasises the fact that ‘generic’ strategies, where possible (that is, when
the target function is nonlinear and non-injective), are very robust to different forms of
the model leakage. The various generic-compatible distinguishers differ in effectiveness
by scenario and by magnitude of noise so that the ‘optimal’ strategy is hard to identify
in advance. Strategies making use of a power model require that model to be meaningful
regardless of the statistic used as a distinguisher, and when a ‘good’ power model
is available, and standard assumptions hold, parametric statistics are theoretically just
as capable of successful key recovery and likely to be more efficient in practice—that is,
in the estimation stage. Put simply, statistics making few assumptions are more heavily
dependent on the observed data for information about the underlying population.
Where assumptions can be made, the data have ‘less work’ to do and more immediately
contribute towards refining prior knowledge rather than providing new information. Applying nonparametric (‘robust’) statistics to scenarios known to comply with simplifying
assumptions amounts to ignoring what we already know and tasking the data with informing us from scratch. This is why it may be fruitful to explore half-way strategies
such as the VR-based distinguisher, which potentially progress us towards a ‘no more
flexibility than is needed’ approach.

4.2 Background to the distinguishers

In Chapter 1 we introduced the notion of a ‘standard DPA attack’. We now turn our attention to five particular statistics—Pearson’s correlation coefficient, mutual information,
the Kolmogorov–Smirnov test statistic, the Cramér–von Mises criterion, and the variance
ratio—and explain how they can be used to construct (univariate) DPA distinguishers.
(Higher-dimension variants will be dealt with in Chapter 5). We also discuss the various
issues relating to requisite assumptions and estimation procedures.

4.2.1 Pearson’s correlation coefficient

Pearson’s correlation coefficient measures the total linear dependency between two random
variables $A$ and $B$. It is defined as $\rho(A, B) = \frac{\text{cov}(A, B)}{\sqrt{\text{var}(A)} \sqrt{\text{var}(B)}}$. It takes values from -1 to 1 and is zero whenever $A$ and $B$ are independent. However, the converse is not true; namely, $A$ and $B$ may be (nonlinearly) dependent with a (linear) correlation of 0.
Because we are primarily interested in the magnitude (as opposed to the direction) of the relationship between the true and modelled leakage we base our distinguisher on the absolute value of the correlation, comparing measured traces \( P = L + \varepsilon \) with the hypothesis-dependent predictions \( M_k \) as follows:

\[
D_\rho(k) = |\rho(P, M_k)| = \left| \frac{\text{cov}(P, M_k)}{\sqrt{\text{var}(P)} \sqrt{\text{var}(M_k)}} \right| .
\] (4.1)

Thus applied, Pearson’s correlation coefficient is an example of an *interval-level* distinguisher as described in Chapter 2. Hence, if the model \( M \) adequately approximates the data-dependent leakage \( L \) up to proportionality then we expect (4.1) to be maximised for the correct key hypothesis \( k = k^* \).\(^1\)

**Estimation**

It is estimated from samples \( \{a_i\}_{i=1}^n \), \( \{b_i\}_{i=1}^n \) via the sample correlation coefficient:

\[
r(A, B) = \frac{\sum_{i=1}^n (a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^n (a_i - \bar{a})^2} \sqrt{\sum_{i=1}^n (b_i - \bar{b})^2}}.
\]

This is a consistent estimator for \( \rho(A, B) \) and, moreover, is asymptotically unbiased and efficient if \( A \) and \( B \) have a joint Gaussian distribution.

Under the same assumptions, we can even approximate the sampling distribution via *Fisher’s transformation* which, in the context of DPA, leads to ‘nice’ results such as the number of trace measurements required for attacks to be successful (as discussed in Chapter 3—see also Chapter 6.4 of [70]). Note that both the efficiency of the statistic and the availability of general results derive from its parametric nature.

\(^1\)Accordingly, it has been found (in [12], for example) to successfully recover the key using an ‘identity’ power model when the true leakage is proportional to the Hamming weight. This, as the authors remarked, was possible because the identity is itself sufficiently proportional to the Hamming weight over a small enough range of integers from 0 up. That is, it provided a proportional approximation which, though not perfect, was good enough.
Impact of noise

The impact of noise on the distinguisher is straightforward. In fact, as derived in Chapter 6.3 of [70],
\[ \rho(L + \varepsilon, M_k) = \frac{\rho(L, M_k)}{\sqrt{1 + \frac{\sigma^2}{\text{var}(L)}}}, \]
where \( \sigma \) is the noise standard deviation. Thus, the larger the noise, the more diminished are the correlations. But—crucially—the denominator does not depend on the key hypothesis; the theoretic distinguisher vector is thus scaled in such a way that the rankings and other relative features (such as the standard score and distinguishability measures defined in Chapter 3) are preserved. This does not at all imply that practical correlation-based attacks are immune to noise: As the sample variance of the estimator increases, the number of traces required to reach a sufficient level of precision also increases (see §3.2.1 of Chapter 3). Rivain [95] explores this thoroughly by applying the aforementioned techniques of statistical power analysis to the multivariate distributions of correlation- and template-based distinguishing vectors, thus approximating the sample size needed for correct key recovery in the presence of varying amounts of noise.

Reported performance

The original DPA attacks of [63] looked for a key hypothesis under which a single predicted bit in the target value partitioned the traces into significantly different distributions. When normalised by the standard deviations of the leakage and the predictions, this becomes equivalent to the correlation distinguisher using a 1-bit power model. The idea of combining all the predicted target bits into a prediction for the actual leakage function, and then correlating with the actual traces (first formally proposed in [19] but hinted towards in previous work such as [32, 75]) is a natural progression towards better exploiting the information in the measured data. Of course, it relies on knowing something about the form of the device leakage, but in many cases (CMOS technology, for example) we do [3]. Moreover, the Gaussian noise assumption is very reasonable for the actual noise in a system (though not necessarily for the residual noise when an imperfect model is fitted). Thus correlation DPA has tended to perform extremely efficiently in typical leakage scenarios.
4.2.2 Mutual information

The appeal of MI for use in DPA is that, rather than measuring linear dependencies only, it quantifies the total information (in bits) shared between two random variables. It is defined as

$$I(A; B) = \sum_{a \in A} \sum_{b \in B} p_{A,B}(a, b) \log_2 \left( \frac{p_{A,B}(a, b)}{p_A(a) p_B(b)} \right),$$

where $p_{A,B}$ is the joint probability density of $A$ and $B$ and $p_A$, $p_B$ are the marginal densities.

However, it is more intuitive (and often more convenient) to express it in terms of entropies via Shannon’s formula:

$$I(A; B) = H(A) - H(A|B)$$
$$= H(B) - H(B|A)$$
$$= H(A) + H(B) - H(A, B)$$
$$= H(A, B) - H(A|B) - H(B|A),$$

where $H(A) = -\sum_{i=1}^{n} p_A(a) \log_2 p_A(a)$ is called the Shannon entropy of $A$, and $H(A|B) = \mathbb{E}_{b \in B}[H(A|B = b)]$ is the conditional entropy.

These equivalent formulations emphasise the fact that $I$ is symmetric and that it communicates the reduction in entropy (uncertainty) of $A$ that results from knowing $B$ (or vice-versa). If $A$ and $B$ are independent then observing $B$ tells you nothing about $A$ so that $I(A, B) = 0$—the lower bound. The maximum value that $I(A, B)$ can take is $H(A)$ [= $H(B)$], occurring in the event that the two variables are completely determined by one another.

It can also be thought of as the Kullback–Leibler divergence\(^2\) between the joint distribution of $A$ and $B$ and the product of the marginal distributions. This, again, aids intuition as $A$ and $B$ are independent if and only if their joint density is the product of their marginal densities (i.e. $p_{A,B} = p_A p_B$)—in which event the divergence (i.e. the MI) is of course zero.

---

\(^2\)The Kullback–Leibler divergence is a non-symmetric measure of the distributional distance between two random variables $P$ and $Q$, defined as $\text{KL}(P||Q) = \int_{-\infty}^{\infty} p_P(x) \ln \frac{p_P(x)}{p_Q(x)} \, dx$. 

76
Its application as an attack distinguisher is as follows:

\[ D_{\text{MI}}(k) = I(P; M_k) = H(P) - H(P|M_k) = H(P) - \mathbb{E}_{m \in M} [H(P|M_k = m)], \quad (4.2) \]

and because the ‘unexplained’ entropy (the second term) is smallest when the predictions are good, we expect (4.2) to be maximised for the correct key hypothesis \( k = k^* \)— provided, of course, that the correct-key power model \( M \) is a good approximation for \( L \) up to nominality (MI is an example of a nominal-level distinguisher as described in Chapter 2).

**Estimation**

As a functional of probability distributions, MI is notoriously problematic to estimate [17, 58, 84, 102, 115]. All estimators are biased, and further no universally ‘best’ estimator exists—that is to say, different estimators perform differently depending on the underlying structure of the data. The usual approach is to first estimate the underlying marginal and conditional densities and then to substitute these into Shannon’s formula via a ‘plug-in’ estimator for discrete entropy. There are many different ways to estimate densities and the quality of the associated MI estimator is very sensitive to the methods and parameters chosen.

If we have a good understanding of the underlying distributions we can fit a parametric model such as a Gaussian mixture (see Veyrat-Charvillon et al. [117]). However, since MI-based DPA has been proposed for use in scenarios where our usual assumptions do not hold we are generally more interested in nonparametric methods, which are somewhat sensitive to user approach and known to incur an overhead in terms of estimation costs. In practice, due to the large sample space and small datasets we usually estimate the densities via an \( m \)-bin regularisation of the space. By an important data processing inequality\(^3\) this means we are always estimating a lower bound on the MI — as the binning or mesh becomes finer the estimate approaches the true MI monotonically from below [84].

\(^3\) \( I(S(A); T(B)) \leq I(A; B) \) for any random variables \( A \) and \( B \) and any functions \( S \) and \( T \) on the range of \( A \) and \( B \)
The maximum likelihood estimator for the entropy is 
\[- \sum_{i=1}^{m} \hat{p}_{N,i} \log_2(\hat{p}_{N,i}),\] 
where \(N\) is the sample size and \(\hat{p}_{N,i}\) is the discretised empirical probability. The bias is large and the variance small for small \(N\), especially if \(m\) (the number of bins) is large. Much study has been devoted to reducing the bias, but the goal of DPA is key distinguishability and so precision is far more important than accuracy. From the information theoretic literature ([17, 84, 115]) we learn that the leading order term in the expansion of the bias does not depend on the underlying distribution, which goes some way to argue that bias correction is irrelevant to key distinguishability (any adjustment would be approximately constant for all key hypotheses and may even incur a loss of precision, which would have a detrimental impact on attack effectiveness).\(^4\)

Unfortunately, estimators for MI do not behave so ‘nicely’ as the sample correlation coefficient; in fact, there are no universal rates of convergence [84], so that whatever estimator we pick, we can always find a distribution for which the error vanishes arbitrarily slowly. In the absence of general results about the sampling distribution of the estimators, we cannot compute (for example) the number of traces needed for an attack to be successful, except under the strongest of assumptions\(^5\).

**Impact of noise**

Unlike correlation DPA the impact of noise on the MI-based distinguishing vector is complex (see, for example, [69]). In particular, whilst \(I(L + \varepsilon; M_k) \leq I(L; M_k)\) \((L, \varepsilon\) independent), nonetheless \(I(L; M_k)/I(L + \varepsilon; M_k) \neq I(L; M_{k'})/I(L + \varepsilon; M_{k'}).\) Hence, the vector elements are differentially affected so that theoretic outcomes in a pure-signal setting do not directly generalise to theoretic outcomes in the presence of noise. To our knowledge, we were the first to explore the implications of this fact for DPA, which will become clearer in the analysis of §4.4.

\(^4\)A note of warning: the higher order terms do depend on the distributions and are not necessarily insignificant in the expansion [115]. Therefore we cannot be certain that bias does not affect key hypotheses differentially.

\(^5\)Under strong simplifying assumptions, estimating an MI-based DPA parametrically can be shown to be equivalent to conducting a correlation attack [71].
Reported performance

When MI was first proposed in [49], it was presented as an enhancement of correlation DPA which was expected to display certain advantages over current practice, loosely summarised by three (informal) conjectures:

1. Optimality: By comprehensively exploiting all of the information contained within trace measurements it could have an efficiency advantage.

2. Genericity: By capturing total dependency (linear + nonlinear) between the true device leakage and the model it could prove effective in the absence of an accurate model. (Note that this is a vaguer notion of ‘genericity’ than the one we formalise in Chapter 2).

3. Multivariate applicability: Because it has natural multivariate extensions it might be adapted to higher-dimension attacks without the need for (information-lossy) pre-processing.

In practice it has largely disappointed with respect to all but the third of these expectations [12, 49, 89, 117], which actuality we seek to explore in this thesis. We reserve our analysis of multivariate extensions for Chapter 5; this Chapter focuses on the goal of genericity, with a brief discussion of efficiency in §4.3.

4.2.3 Kolmogorov–Smirnov

The KS distance between the distributions of random variables $A$ and $B$ is defined as $K(A\|B) = \sup_{x \in A \cup B} |F_A(x) - F_B(x)|$ where $F_A$, $F_B$ are the cumulative distribution functions (CDFs) of $A$ and $B$, i.e. $F_A(x) = \mathbb{P}(A \leq x)$. In a two-sample KS test designed to test the null hypothesis that $A$ and $B$ share the same distribution, the empirical CDFs are estimated from samples $\{a_i\}_{i=1}^n$, $\{b_i\}_{i=1}^n$, e.g. $\hat{F}_A(x) = \frac{1}{n} \sum_{i=1}^n I_{a_i \leq x}$ ($I_{a_i \leq x}$ is the indicator function, taking the value 1 if $a_i \leq x$ and 0 otherwise). The fact that the KS test statistic does not require explicit density estimation is what makes it appealing as an alternative to MI.
Just as MI-based DPA can be understood to operate by comparing the global traces $P$ with the hypothesis-dependent conditional traces $P|M_k$—via the expected change in entropy—a KS-inspired distinguisher measures the maximum distance between the global and the conditional trace distributions, as averaged over the prediction space:

$$D_{KS}(k) = \mathbb{E}[K(P||P|M_k)] = \mathbb{E}_{m\in M} \left[ \sup_y |F_P(y) - F_{P|M_k=m}(y)| \right].$$

(4.3)

In case of the correct key hypothesis we expect the test statistic to return a large difference. This again depends on $M$ approximating $L$ up to nominality.

**Estimation**

One advantage of KS over MI for use in DPA is that the former does not require explicit estimation of the probability density functions, and the difficult user choices attendant with such problems (kernel vs. histogram vs. other, bin width, etc.). Empirical CDFs converge ‘almost surely’ to their associated population distributions, unlike the corresponding sample densities which do not converge to the underlying PDFs. However, estimation can be computationally more costly (for example, compared with histogram estimations of MI) precisely because no ‘binning’ of the traces takes place.

**Reported performance**

The KS-based distinguisher was introduced and empirically compared with MI in [117]; both were applied against the DPA contest dataset [1]—known to be favourable to the Hamming weight model, which was used in both cases—and the KS was found less efficient in terms of the number of traces needed for success. To our knowledge, this is the only place in the literature where KS-based DPA was implemented prior to our work (some of which has been published in [123]); since it deals with just one (correlation-favouring) scenario, only presents experimental results, and does not explore the ‘generic’ potential of the distinguisher, we consider that there is room for further investigation.
4.2.4 Cramér–von Mises

The CvM criterion is similar to the KS test statistic in that it also compares two distributions via their CDFs. Instead of the maximum absolute difference over the support space, CvM integrates the squared difference:

\[ C(A||B) = \int_{-\infty}^{\infty} (F_A(x) - F_B(x))^2 dx. \]

As above, this is used to test the null hypothesis that \( A \) and \( B \) share the same distribution, and is estimated via the empirical CDFs.

We construct a CvM-based DPA distinguisher similarly to the MI- and KS-based ones—that is, we average the statistic (as applied to compare the global with the hypothesis-dependent conditional traces) over the entire prediction space:

\[
D_{\text{CvM}}(k) = \mathbb{E}[C(P||P|M_k)] = \mathbb{E}_{m \in \mathcal{M}} \left[ \int_{-\infty}^{\infty} (F_P(y) - F_{P|M_k=m}(y))^2 dy \right].
\]

By the same intuition as the preceding distinguishers, as long as \( M \) approximates \( L \) up to nominality we expect the test statistic to return a large difference under the correct key hypothesis.

Reported performance

The only place, to our knowledge, that this has been tested in the literature is again in [117]; unlike the KS-based distinguisher, CvM performed almost comparably with the MI-based methods, when tested (using a Hamming weight power model) against the DPA contest dataset.

4.2.5 Variance ratio

The entropy of a Gaussian distribution \( N(\mu, \sigma^2) \) depends only on its variance:

\[ H(N(\mu, \sigma^2)) = \frac{1}{2} \ln ((2\pi e)\sigma^2). \]

If the non-deterministic part of the leakage can be assumed to be Gaussian, then the conditional traces produced under the correct key hypothesis using a nominal power model which is perfectly precise (though need not have perfect recall)—so that this includes the generic power model—will follow a Gaussian distribution. Thus motivated, the authors of [106] propose to avoid density estimation altogether and
simply compute variances. As such, they propose a distinguisher based on the ratio of
the global trace variance to the (average) conditional trace variance:

\[ D_{VR}(k) = \frac{\mathbb{E} \var(P)}{\var(P|M_k = m)}. \] (4.5)

By combining a nominally approximating model with a Gaussian noise assumption in
this way, we mitigate against introducing more flexibility than is needed to our DPA
methodology, remembering that there is always a trade-off between flexibility and data-
complexity. We therefore expect that, in scenarios where these assumptions reasonably
hold, the VR-based distinguisher will outperform other distinguishers suitable for use with
a nominal model.

**Reported performance**

In [106] where the VR is introduced, it is experimentally tested, relative to MI- and
correlation-based distinguishers, against implementations of AES-128 on two 8-bit RISC-
based microcontrollers (a PIC 16F877 and an Atmel ATmega163 in a smart card body).
They found that it nearly always outperforms MI-based distinguishers when both are pro-
vided with the same power model (‘partition’), but is, in turn, nearly always outperformed
by correlation DPA. The study as a whole, though, emphasises the scenario-dependency
of outcomes and indicates the need for more understanding of how the different distin-
guishers behave in different situations.

**4.3 Mutual information as ‘theoretically optimal’**

As previously mentioned, MI was initially hoped to produce more efficient DPA attacks by
virtue of its information theoretic optimality. However, it was quickly found to disappoint
on this front.

Empirical evaluations of MI-based DPA have indicated that, in scenarios favouring corre-
lation DPA (such as those where the data-dependent leakage is known to be well approxi-
mated by the Hamming weight) it is highly unlikely to offer any advantage over the latter
(see, for example, the analysis of [117] which tested them both (using Hamming weight
power models) against the DPA contest dataset [1] and also showed, using simulated traces, that the correlation-based distinguisher continues to outperform the MI-based up to quite a high degree of divergence between the model and the true leakages). This is hardly surprising, in view of the problematic and costly estimation procedures which we described above in §4.2.2.

However, such analyses are necessarily estimator-specific and do not permit conclusions about the underlying theoretic capabilities of the MI-based distinguisher. Indeed, much work has been devoted to finding improved estimators which do produce the looked-for advantages [8, 89, 117]. Meanwhile, little attempt has been made to identify whether or not these advantages are theoretically attainable even given perfect estimation procedures.

This is not immediately obvious. In particular, it is hard to say a priori what will happen to the “ghost peaks” [19] in a distinguishing vector should we succeed in completely quantifying the total dependencies rather than just the linear. These originate from subtle dependencies between the leakages and the incorrect key predictions (induced by the target function), producing non-zero (potentially quite large) distinguisher values under incorrect key predictions. Attack success does not simply depend on the value produced by the correct key hypothesis but on the relative size as compared with all the alternatives. Using MI increases the capture of information for the ghost peaks concurrently with the true peak; this increase need not be constant—indeed we might intuitively expect it to be greater for the incorrect hypotheses where the underlying relationships are more complex (§4 of [80] expounds on this idea). As such, enhanced information capture may in fact be to the detriment of distinguishability, even in the theoretical realm. (This is before considering the distribution-dependent performance of MI estimators (see, for example, [84]), which makes it all the more likely that the impact of choosing MI over correlation will differ by key hypothesis).

In order to settle the question we simply examine the theoretic vectors in a correlation-favouring scenario of (known) Hamming weight leakage of the first DES S-Box with zero noise. These are displayed in Figure 4.1. (Note that these outcomes are key-independent [71]: because the target function has the Equal Images under different Subkeys (EIS) property [97] and the plaintexts are assumed uniformly distributed, the correct hypothesis yields the same distinguisher value under any key, and only the arrangement of the
remaining vector entries changes).

It is evident that both attacks are first-order successful by a clear margin, but that MI has a substantial advantage, with a relative margin of 5.61 compared with just 3.56 for correlation. (The corresponding standard scores are 6.59 and 5.14). Thus far the a priori intuition is confirmed: MI does produce a more theoretically optimal attack and it must instead be the relative efficiency of estimating the correlation coefficient which enables correlation to consistently outperform MI in practical attacks with a good power model.

![Figure 4.1](image_url)

**Figure 4.1:** Ideal distinguishing vectors using the HW power model to attack the output of the first DES S-Box.

As a partial insight into the quantity of data needed we next look at the critical support size required for the distinguishers to approach their full ideal potential, as also introduced in Chapter 3. The space of possible plaintext combinations is too large to explore exhaustively, so we look at the average behaviour of the attacks in repeated random draws from the plaintext space. We find that correlation DPA is able to identify the correct key from a far smaller support than MI-based DPA, with an average critical support of 6 and a threshold of 16 for a 100% success rate, compared with an average of 8 and threshold of 19 for MI. Note as well that even once a high ideal success rate is achieved, it may be that a broader support is required before the MI-based distinguisher regains the advantage it displays with respect to the full distribution.

### 4.4 Comparative evaluation

We have thus clarified that the disadvantage of the MI-based distinguisher in typical leakage scenarios arises from the data-complexity of estimation rather than any theoretically inherent weakness relative to correlation DPA. We next seek to establish scenarios
where the advantages afforded by its flexibility are sufficient to outweigh this efficiency disadvantage. That is, we seek to explore its application as a generic distinguisher. We simultaneously investigate the CvM and the KS variants as these are conceptually similar, and the VR which presents a potentially more efficient alternative by reliance on assumptions about the noise.

4.4.1 Evaluation scenarios

We consider a non-profiling adversary who can use either a ‘standard’ power model, namely the Hamming weight, or, when appropriate to the distinguisher, a ‘generic’ (for non-injective target functions) or ‘near-generic’ (for injective target functions) power model, namely the identity or the \((m - 1)\) least significant bits \(((m - 1)\text{-LSB})\) of the hypothesised processed value. These latter two were proposed in [49] as suitable for use with an MI-based distinguisher, and have been consistently promoted in subsequent studies such as [12, 80, 106]. Some papers [89, 117, 119] have explicitly recognised that whatever the chosen model, it must at least maximise the MI under the correct hypothesis, and that this condition will not be guaranteed for an arbitrary choice such as the \((m - 1)\text{-LSB}\). However, it is often used regardless, in the absence of any meaningful compressing information. Hence, the theoretic capabilities of ‘near-generic’ strategies relative to ‘generic’ strategies is one point of particular interest, as well as the theoretic capabilities of nonparametric-based distinguishers relative to one another and relative to correlation DPA.

Our evaluation will be with respect to three different leakage scenarios: The first, and simplest, is an optimistic scenario in which the data-dependent leakage really is proportional to the Hamming weight as per the adversary’s standard model. The second we term realistic: the attacker’s power model is meaningful but degraded relative to the true leakage. As motivated by [3], we suppose that the true leakage is actually an unevenly-weighted sum of the bits, deliberately picking a weighting which was found in [117] to produce sufficient distortion away from the Hamming weight power model for MI-based DPA to have an advantage over correlation DPA. 6 Lastly, as an example of a challenging

\(^6\)Specifically, we allow the least significant bit (LSB) to dominate with a relative weight of 10, as per the experiments of [117]. We stress that this is not intended to be realistic in the sense of representing known device leakage—rather, in capturing a realistic degree of model degradation, which can easily be
(but still realistic) scenario, we consider the case where the attacker’s power model is essentially meaningless relative to the true leakage. We explore this by supposing that the true leakage is a highly nonlinear function of the intermediate data. This is chosen to represent an informal ‘worst case’ scenario; it is not motivated by any specific known device behaviour, though there is plenty of evidence for device leakage capable of being totally unrelated to standard power models (examples include typical hardware implementations of substitution boxes [72] and emerging nanoscale technologies [94]).

We test our distinguishers against three target algorithms running in these leakage environments: the DES, AES and PRESENT block ciphers. In DES, 6-bit chunks of plaintext are XOR-ed with 6 key bits before being input into one of eight S-Boxes, each with 4-bit outputs (so that they are non-injective). The algorithm is a popular context for analysis because of its long-established and widespread use, its general susceptibility to DPA (arising from the high degree of nonlinearity built into the S-Boxes), and its particular amenability to generic attacks (arising from the non-injectivity of the S-Boxes). By contrast, the AES S-Box accepts 8-bit inputs and is injective, and the PRESENT S-Box accepts 4-bit inputs and is also injective, which make them interesting comparators for investigating near-generic strategies.

Within each leakage/target function scenario we will first consider the behaviour of our distinguishers in a noise-free setting—to demonstrate the role of data-dependent leakage in determining attack outcomes—and then go on to show how independent Gaussian noise of varying size impacts on the distinguisher outcomes. We focus on an SNR range chosen to illustrate the interesting noise-dependent behaviour of the theoretical vectors, first uncovered by our investigations. However, the smallest SNR we consider is 0.125; in interpreting our analysis it should be remembered that many real-life leakage scenarios fall below this bound.

---

7Specifically, we map the target value to the Hamming weight of the AES S-Box output. There is no significance to this choice other than that it is well-known and specially fitted with the nonlinearity properties useful to produce our hypothetical degraded model scenario.
4.4.2 Targeting the first DES S-Box

The noise-free setting  The first two blocks of Table 4.1 report the outcomes of standard and generic univariate attacks on an unprotected DES S-Box with noise-free data-dependent leakage. As we have already seen in §4.3, in the optimistic scenario, the MI-based distinguishers exhibit substantially larger relative margins than standard correlation DPA, but also require a larger support to be successful. It is this initial ‘information overhead’, combined with the relative efficiency of estimating the correlation coefficient, which accounts for the consistently reported advantage held by correlation in practical attacks with a good power model. The other generic-compatible distinguishers display similar characteristics; CvM actually outperforms MI and KS in terms of relative margin (as per the experimental findings of [117]). (Note that the theoretic correct key VR (under a perfectly precise partition) tends towards infinity as the noise tends to zero, so that the associated distinguishing margins are undefined in the noise-free setting).

When the standard Hamming weight power model is a good fit to the true leakage, generic MI-based DPA offers no advantage, exhibiting a substantially reduced margin in absolute terms and requiring a larger input support to succeed. This is the expected consequence of exchanging the power model for one which is less accurate (in ‘recall’ terms, at least), and accords with the theory laid out in Chapter 2. The same applies for the CvM-based distinguisher (KS does have slightly increased margins but it also has a larger critical support).

As the true leakage diverges from the standard power model, the advantage to MI increases. In the challenging scenario, correlation DPA actually fails whilst the MI-based distinguisher with the degraded HW power model continues to identify the correct key. CvM and KS variants also continue to succeed, though with less decisive margins and greater critical support. The VR-based distinguisher (with power model) eventually fails in the challenging scenario.

The flexibility of the generic variants (where applicable) is evident—the distinguishing margins and the critical support size are remarkably robust to the deterioration of the leakage. Again, this is what we would expect, since the nominal accuracy of the generic power model is in some sense independent of the true leakage: whilst it is always preferable
to leverage prior knowledge when a meaningful model is available, a generic model will work just as well however typical or unusual an unknown leakage function really is.

The impact of noise on distinguishing margins  Figure 4.2 shows the impact of noise on the distinguishing margins of standard attacks. As we know already (from §4.2.1), the correlation vector is merely scaled by a constant as the SNR varies, so that the relative distinguishing margin is unchanged. By contrast, the relative margins for MI, KS, CvM and VR are affected by noise, and in such a way that the relationships are not necessarily monotonic. This is particularly marked for MI and KS: in each leakage scenario there seems to be an optimal SNR at which the margin reaches a maximum, subsequently diminishing to that of the noise-free setting. Such a phenomenon is a type of stochastic resonance [14], which can (in principle) occur in any nonlinear measurement system.

In the optimistic scenario, the standard VR distinguisher displays the greatest relative margins across the tested range. The ‘noise-free’ advantage exhibited by the CvM distinguisher gives way to the MI distinguisher in noisy scenarios, whilst the KS margins diminish to below those of correlation. In the realistic scenario, the standard MI and KS distinguishers appear very sensitive to noise, whilst standard CvM and VR remain fairly robust. In the challenging scenario, the success of the standard MI and KS distinguishers is noise-dependent (in this case, it only manifests with a strong enough signal), whilst the other distinguishers remain ineffective across the tested range.

The lower part of the figure shows absolute margins as the SNR varies. These are most robust for correlation DPA, in such cases that it is theoretically successful (i.e. the optimistic and realistic scenarios). Since the actual size of the margins to be estimated has a bearing on the amount of data needed for estimation (in addition to the size relative to the variation in the vector), this is likely only to enhance its proven advantage in practical attacks in the presence of noise. It is interesting to note that KS absolute margins are more robust to noise than those of the MI distinguisher, so that the former method may actually prove the preferable of the two in (noisy) practical settings.

Figure 4.3 displays the counterpart generic distinguishers. There is little to comment on here; the stochastic resonance-type behaviour is still evident, particularly for MI; all four generic-compatible methods remain successful across the tested range, even in the
The impact of noise on critical support size  
Within each scenario, we tested the strongest MI, KS, CvM and VR variants (standard in the optimistic scenario, generic in the realistic and challenging scenarios) to see whether or not noise had any detrimental
effect on the support size required for key recovery. We found that it did not (i.e. the outcome measures relating to support size remained constant across the tested SNR range) with one minor exception: the generic KS distinguisher against realistic leakage required on average one more trace in the highest noise scenario than it did in the pure-signal scenario. (For correlation we do not need to test this because of the noise-invariance of the shape of the distinguishing vector). Thus any advantages of MI, KS, CvM and VR in terms of distinguishing margin size and (in the generic case) scenario and noise robustness are not in general undermined by increased support size costs as noise varies.

4.4.3 Targeting the AES S-Box

The noise-free setting We next consider attacks against the AES algorithm. Block 3 of Table 4.1 reports outcomes of standard attacks against the AES S-Box with noise-free data-dependent leakage. As was the case for DES, the MI- and CvM- based attacks with the Hamming weight power model have substantially larger relative margins than correlation, with KS falling only slightly behind. However, all standard attacks are less robust to severe model degradation and fail in the challenging scenario.

Block 4 summarises the theoretic outcomes of ‘near-generic’ attacks in the absence of noise; it is immediately apparent that these are not comparable to the generic variants as used against DES. In fact, MI with a 7LSB power model fails dramatically, with the correct key appearing last in the ranked vector (by a substantial margin). The equivalent KS and CvM attacks succeed, but with greatly reduced distinguishing power and requiring almost half of the total input support space. The VR attack succeeds with apparently large margins—the explanation provided by the model is no longer ‘perfect’ as before so that the true-key quantity does not tend to infinity—but also requiring substantial support.

‘Near-generic’ KS, CvM and VR attacks remain successful in the ‘realistic’ leakage scenario—in fact, with better margins and reduced critical support, which is a quirk of the way the power model interacts with the particular leakage function chosen. They all, however, fail in the ‘challenging’ scenario.

The failure of ‘near-generic’ MI is surprising—and even concerning—given the reported
effectiveness of the attack in practical scenarios (for example in the experiments of [80]). For clues to this apparent incongruity we must look beyond the noise-free setting.

**The impact of noise on distinguishing margins** The relative margins of the distinguishers using the standard Hamming weight power model (see Figure 4.4) appear less affected by noise than in the attacks against DES. There are some particular similarities: in the optimistic scenario the VR distinguisher achieves the largest relative margin and the CvM advantage over MI disappears as noise increases. They all succeed across the tested range in both this and the realistic scenario, and they all fail across the tested range in the challenging scenario. In absolute terms (see bottom left panel of Figure 4.4) the KS distinguisher once more appears more robust to noise than the MI, with correlation the most robust of all.

Examining the first panel of Figure 4.5 it becomes clear that it is only in the strong-signal setting (SNR > 10, approx.) that the attack fails to distinguish the correct key. As the data-dependent signal weakens, the attack becomes theoretically distinguishing, though with consistently smaller relative and absolute margins than the standard attacks.

A closer inspection of Shannon’s formula in the noise-free setting (I(P; M_k) = H(P) − H(P|M_k)) reveals the reason for the eventual failure: except for the correct key hypothesis, certain model predictions induce conditional distributions which are supported on a single point and therefore contribute zero entropy to the overall (expected) conditional entropy component H(P|M_k). All other conditional distributions, including all those induced under the correct key hypothesis, are supported on two values and thus contribute entropy of one bit. So the expected conditional entropy is 1 for the correct key and less than 1 everywhere else; since the global entropy H(P) does not change, I(P; M_k) is minimal for the correct key.

In the ‘realistic’ scenario the MI margins achieved in the noisy setting are larger, again reflecting the interaction between the choice of power model and the actual leakage. However, in the challenging scenario all of the attacks fail across the tested range, emphasising the fact that the 7LSB model does not provide a generic equivalent for extension to injective functions.

91
Figure 4.4: Theoretic relative and absolute distinguishing margins as the (Gaussian) noise varies, for standard univariate attacks against the AES S-Box.

Figure 4.5: Theoretic relative and absolute distinguishing margins as the (Gaussian) noise varies, for ‘near-generic’ univariate attacks against the AES S-Box.

4.4.4 Targeting the PRESENT S-Box

The noise-free setting Standard attacks against noise-free leakage of the PRESENT S-Box display smaller margins than those against the AES S-Box, as only half the number of bits are being targeted so that there is less information by which to distinguish between
the correct and incorrect keys.

Otherwise, they compare similarly. In the optimistic scenario, CvM again has the largest margins, whilst all generic-compatible distinguishers require a similar increase over correlation in terms of critical support. In the realistic scenario, MI overtakes CvM and the VR method seems to require a larger increase in support size than the three nonparametric distinguishers. All tested standard attacks fail in the challenging scenario.

The ‘near-generic’ attacks, this time using the 3LSB model against the 4-bit injective target, perform even less well than the 7LSB attacks against the 8-bit AES S-Box. In the optimistic scenario, not only the MI but also the KS and CvM distinguishers fail. The VR succeeds but does less well than simply using correlation with a 3LSB power model, which is counter-intuitive. In the realistic scenario we again see improved outcomes for CvM and KS—so that they become successful—by dint of the convenient interaction between the power model and actual leakage. As with AES, all 3LSB attacks fail in the challenging scenario.

The impact of noise on distinguishing margins  In the optimistic scenario we again observe the standard VR distinguisher to have the largest relative margins across the tested range, and again note that the CvM distinguisher loses its advantage over MI in the presence of enough noise. As before, the correlation and, secondly, the KS attacks have the best noise-robustness in absolute terms.

The realistic scenario is interesting, as the success of both MI and KS is shown to be critically sensitive to noise within the tested range due to the stochastic resonance-type behaviour noted previously. The KS distinguisher fails for SNR below around 64, whilst the MI distinguisher fails between SNR of 2 and around 16 (that is, it succeeds again when the noise is high enough as well as when it is low enough).

The failure of all the standard attacks in the challenging leakage scenario persists across the tested range.

The 3LSB attacks are persistently ineffective in the optimistic and challenging scenarios. Strangely, in the realistic scenario the CvM, MI and KS attacks become first-order successful (by small margins) within small regions of the tested range. But since this is a
scenario in which successful standard attacks exist, this is not to say that the 3LSB model would ever be of practical use.

**Figure 4.6:** Theoretic relative and absolute distinguishing margins as the (Gaussian) noise varies, for standard univariate attacks against the PRESENT S-Box.

**Figure 4.7:** Theoretic relative and absolute distinguishing margins as the (Gaussian) noise varies, for ‘near-generic’ univariate attacks against the PRESENT S-Box.
Table 4.1: Theoretic outcomes in optimistic, realistic and challenging leakage scenarios without noise.

<table>
<thead>
<tr>
<th></th>
<th>Optimistic</th>
<th></th>
<th>Realistic</th>
<th></th>
<th>Challenging</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Corr.</td>
<td>MI</td>
<td>KS</td>
<td>CvM</td>
<td>VR</td>
<td>Corr.</td>
</tr>
<tr>
<td>1. Standard,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| against DES    | Correct key ranking (order) | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1
|                | Standard score | 5.14 | 6.59 | 5.95 | 7.34 | – | 3.21 | 6.38 | 5.49 | 5.09 | 5.64 | 0.74 | 5.23 | 2.66 | 3.07 | 0.37 |
|                | Relative margin | 3.56 | 5.61 | 4.24 | 6.38 | – | 1.22 | 5.12 | 3.61 | 3.17 | 3.57 | –2.38 | 3.22 | 0.40 | 0.16 | -2.78 |
|                | Absolute margin | 1.00 | 1.00 | 1.00 | 1.00 | – | 0.30 | 0.86 | 0.66 | 1.20 | 0.00 | –0.28 | 0.21 | 0.04 | 0.01 | 0.00 |
|                | Average critical support | 6 8 8 8 8 | 17 10 | 12 | 11 | 21 | – | 26 | 39 | 39 | – | 37 | 61 | 57 | – |
|                | Critical support for 90% SR | 16 19 | 19 | 19 | 19 | 19 | 32 | 14 | 20 | 18 | 37 | – | 37 | 61 | 57 | – |
|                | Critical support for 100% SR | 16 | 19 | 19 | 19 | 19 | 49 | 21 | 34 | 30 | 51 | – | 46 | 64 | 63 | – |
| 2. Generic,   |            |             |           |             |             |             |             |             |             |             |             |
| against DES    | Correct key ranking (order) | 1 | 1 | 1 | 1 | 1 | 8 | 1 | 1 | 1 | 1 | 64 | 1 | 1 | 1 | 1 |
|                | Standard score | 5.39 | 6.35 | 6.20 | 6.81 | – | 1.45 | 6.66 | 5.77 | 5.90 | – | 1.29 | 6.48 | 5.94 | 6.84 | – |
|                | Relative margin | 3.61 | 5.08 | 4.60 | 5.66 | – | -0.81 | 5.45 | 4.12 | 4.39 | – | -3.95 | 5.30 | 4.41 | 5.55 | – |
|                | Absolute margin | 0.85 | 0.68 | 0.74 | 0.76 | – | -0.14 | 0.86 | 0.80 | 3.30 | – | -0.55 | 0.77 | 0.80 | 1.08 | – |
|                | Average critical support | 9 | 16 | 16 | 15 | 16 | – | 15 | 15 | 14 | 15 | – | 15 | 15 | 14 | 15 |
|                | Critical support for 90% SR | 14 | 19 | 19 | 18 | 19 | – | 17 | 17 | 16 | 17 | – | 18 | 18 | 17 | 18 |
|                | Critical support for 100% SR | 27 | 24 | 24 | 22 | 24 | – | 21 | 21 | 19 | 21 | – | 25 | 25 | 21 | 25 |
| 3. Standard,  |            |             |           |             |             |             |             |             |             |             |             |
| against AES    | Correct key ranking (order) | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 186 | 193 | 214 | 245 | 251 |
|                | Standard score | 12.24 | 15.60 | 13.91 | 15.74 | – | 9.61 | 15.49 | 12.69 | 13.90 | 14.29 | -0.77 | -0.75 | -1.03 | -1.41 | -1.72 |
|                | Absolute margin | 1.00 | 1.00 | 1.00 | 1.00 | – | 0.49 | 0.94 | 0.58 | 0.92 | 0.00 | -0.26 | -0.04 | -0.13 | -0.06 | 0.00 |
|                | Average critical support | 5 | 9 | 9 | 9 | 9 | 22 | 12 | 18 | 18 | 37 | – | – | – | – |
|                | Critical support for 90% SR | 6 | 11 | 11 | 10 | 11 | 39 | 16 | 30 | 32 | 61 | – | – | – | – |
|                | Critical support for 100% SR | 9 | 15 | 15 | 14 | 15 | 71 | 23 | 47 | 68 | 101 | – | – | – | – |

Continued on next page...
Table 4.1: Theoretic outcomes in optimistic, realistic and challenging leakage scenarios without noise (continued).

<table>
<thead>
<tr>
<th></th>
<th>Optimistic</th>
<th>Realistic</th>
<th>Challenging</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Corr.</td>
<td>MI</td>
<td>KS</td>
</tr>
<tr>
<td>Correct key ranking (order)</td>
<td>1 256 1 1 1</td>
<td>1 256 1 1 1</td>
<td>153 212 164 218 237</td>
</tr>
<tr>
<td>Standard score</td>
<td>11.23 -5.75 6.12 6.46 14.50</td>
<td>3.86 -3.98 10.23 11.17 15.96</td>
<td>-0.44 -1.08 -0.42 -1.03 -1.31</td>
</tr>
<tr>
<td>Relative margin</td>
<td>8.88 -7.74 3.50 3.50 12.71</td>
<td>0.70 -0.79 7.96 9.22 15.95</td>
<td>-3.87 -4.06 -3.72 -4.16 -5.65</td>
</tr>
<tr>
<td>Absolute margin</td>
<td>0.57 -0.11 0.11 0.11 0.00</td>
<td>0.03 -0.08 0.26 1.34 0.00</td>
<td>-0.19 -0.06 -0.11 -0.13 0.00</td>
</tr>
<tr>
<td>Average critical support</td>
<td>21 - 125 124 90</td>
<td>153 - 76 72 64</td>
<td>-</td>
</tr>
<tr>
<td>Critical support for 90% SR</td>
<td>35 - 155 154 104</td>
<td>234 - 89 84 72</td>
<td>-</td>
</tr>
<tr>
<td>Critical support for 100% SR</td>
<td>73 - 186 189 135</td>
<td>255 - 110 107 94</td>
<td>-</td>
</tr>
</tbody>
</table>

5. Standard, against PRESENT

<table>
<thead>
<tr>
<th></th>
<th>Optimistic</th>
<th>Realistic</th>
<th>Challenging</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Corr.</td>
<td>MI</td>
<td>KS</td>
</tr>
<tr>
<td>Correct key ranking (order)</td>
<td>1 1 1 1 1</td>
<td>1 1 1 1 1</td>
<td>9 6 10 12 12</td>
</tr>
<tr>
<td>Standard score</td>
<td>2.52 2.94 2.79 3.16 -</td>
<td>1.44 3.17 2.07 1.74 2.06</td>
<td>-0.30 -0.30 -0.63 -0.61 -0.51</td>
</tr>
<tr>
<td>Relative margin</td>
<td>1.06 2.09 1.94 2.55 -</td>
<td>0.25 2.08 0.62 0.35 1.05</td>
<td>-1.81 -2.69 -3.27 -3.09 -3.15</td>
</tr>
<tr>
<td>Absolute margin</td>
<td>1.00 1.00 1.00 1.00 -</td>
<td>0.20 0.63 0.22 0.20 0.00</td>
<td>-0.51 -0.46 -1.00 -1.07 0.00</td>
</tr>
<tr>
<td>Average critical support</td>
<td>4 7 7 7 7</td>
<td>10 9 9 9 12</td>
<td>-</td>
</tr>
<tr>
<td>Critical support for 90% SR</td>
<td>6 9 9 9 9</td>
<td>15 12 14 13 16</td>
<td>-</td>
</tr>
<tr>
<td>Critical support for 100% SR</td>
<td>10 13 12 13 13</td>
<td>16 14 16 16 16</td>
<td>-</td>
</tr>
</tbody>
</table>

6. Near-generic, against PRESENT

<table>
<thead>
<tr>
<th></th>
<th>Optimistic</th>
<th>Realistic</th>
<th>Challenging</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Corr.</td>
<td>MI</td>
<td>KS</td>
</tr>
<tr>
<td>Correct key ranking (order)</td>
<td>1 14 5 4 1 (4)</td>
<td>1 9 1 (2) 1 (2) 1</td>
<td>1 (2) 12 9 6 3</td>
</tr>
<tr>
<td>Standard score</td>
<td>2.45 -1.41 0.50 0.53 1.64</td>
<td>2.10 -0.90 1.65 1.65 2.83</td>
<td>2.00 -1.13 -0.17 -0.15 0.78</td>
</tr>
<tr>
<td>Relative margin</td>
<td>1.19 -2.83 -1.45 -1.07 0.00</td>
<td>1.00 -2.41 0.00 0.00 1.14</td>
<td>0.00 -3.28 -2.09 -2.33 -2.07</td>
</tr>
<tr>
<td>Absolute margin</td>
<td>0.98 -0.50 -0.35 -0.22 0.00</td>
<td>0.30 -0.33 0.00 0.00 0.00</td>
<td>0.00 -0.67 -0.52 -0.44 0.00</td>
</tr>
<tr>
<td>Average critical support</td>
<td>7 - - - 11</td>
<td>12 - 11 11 12</td>
<td>-</td>
</tr>
<tr>
<td>Critical support for 90% SR</td>
<td>10 - - - 14</td>
<td>16 - 14 14 14</td>
<td>-</td>
</tr>
<tr>
<td>Critical support for 100% SR</td>
<td>14 - - - 15</td>
<td>16 - 16 16 15</td>
<td>-</td>
</tr>
</tbody>
</table>
4.5 Conclusion

This chapter has examined the role of nonparametric statistics in non-profiling DPA, with a particular focus on their contribution towards achieving ‘generic’ DPA. The sense in which they have hitherto been promoted as generic is somewhat loose and does not align with the re-evaluated formalisation we presented in Chapter 2. As per our earlier discussion and investigation, they can only reasonably be understood as generic strategies in the case that they are paired with the generic power model (i.e. in attacks against non-injective targets); on all other occasions, they still, unavoidably, rely on meaningful insight about the true leakage. That is, whilst they only require a nominal approximation (as opposed to, e.g., a proportional approximation), they nonetheless require a certain level of accuracy in that approximation. Completely arbitrary models do not facilitate key recovery, except by chance.

We have also sought to dispel certain other prevailing misunderstandings as to the nature of the advantages potentially offered by a nonparametric approach. It had initially been anticipated that MI-based DPA could even provide efficiency gains over correlation DPA, on the basis that the former quantity is ‘more optimal’ in an information theoretic sense, capturing total dependency between variables rather than simply linear dependencies. However, this is to overlook the substantially increased data complexity associated with the estimation of nonparametric statistics: at the practical level they remain inherently disadvantaged relative to simpler statistics (such as the correlation coefficient) whenever the additional assumptions required by those simpler statistics are reasonably well met. Thus they are only of interest in scenarios where the conditions required for correlation DPA do not hold.

We have next focused on three examples in particular—MI, the two sample KS test statistic, and the CvM criterion—and shown how each can be adapted for use as a DPA distinguisher, discussing the associated assumptions and implementation issues. The three are conceptually similar but the latter two are implementationally ‘simpler’ as they rely less on user-determined parameter tuning. We applied the theoretic evaluation framework introduced in Chapter 3 to compare their theoretic capabilities with those of correlation DPA.
DPA in a variety of leakage scenarios, when used in the manner suggested by the literature. This evaluation revealed that MI, KS and CvM do indeed have certain theoretic advantages even in scenarios which are particularly favourable to correlation DPA (i.e. when the attacker has a good power model), confirming that the under-performance frequently observed in practical experiments can be largely attributed to estimation overheads. When applied as generic strategies (that is, when paired with the generic power model against a non-injective target such as the first DES S-Box) they operate less efficiently than when paired with a good power model, but are remarkably robust and continue to perform well in unusual leakage scenarios. This suggests them to be practically useful alternatives to correlation DPA, which loses considerable distinguishing power as the accuracy of the power model degrades, and often fails altogether.

However, as anticipated by our prior reasoning, the ‘near-generic’ approach using the $(m - 1)$-LSB power model does not supply an equivalent functionality against injective targets (such as the AES and PRESENT S-Boxes)—rather it produces some strikingly unusual results, with success found to depend not only on the true form of the leakage function but also on the amount of noise distorting the data-dependent signal. This, again, is due to the fact that the robustness offered by nonparametric approaches is not by way of an invariance to incorrect prior information (as has been sometimes supposed), rather by way of their ability to operate with less specific prior information.

The noise dependency of theoretic outcomes is itself a very new result: it has always been expected that the presence of noise affects an attack at the practical stage—i.e. the precision with which the distinguishing vector can be estimated—but it has not, to our knowledge, been previously observed that the underlying ability of a distinguisher to recover the key can itself vary, and to a substantial degree. Correlation-based distinguishers do not possess this property, which perhaps accounts for the fact that it has not been previously investigated. KS-based distinguishers, whilst inferior to MI-based in noise-free settings, exhibit a similar adaptability to non-standard leakage and moreover appear to be more robust to increasing noise so that they may prove practically useful alternatives to correlation and MI when the side-channel leakage is both unusual and noisy.
Chapter 5

Multi-target extensions with nonparametric statistics

5.1 Introduction

As mentioned in the previous chapter, mutual information (MI) was promoted as a DPA distinguisher [49] not only because of its flexibility in standard univariate attack scenarios but for its natural multivariate extensions which, it was expected, would enable it to be readily extended to multi-target attacks. Multi-target DPA\(^1\), as we explained in Chapter 1, exploits multiple points in a trace, usually in order to circumvent some countermeasure such as random masking.

Early proposals required a pre-processing step to combine these multivariate measurements into a single ‘trace’ before proceeding with standard DPA methods (using, for example, the correlation coefficient) [76, 110]. Loss of information is inevitable ([29, 90]), and moreover it has been noted that the effectiveness of a given strategy is scenario-specific and relies heavily on an appropriate combination of power model, pre-processing function and distinguisher ([110]).

\(^1\)Elsewhere, what we refer to as ‘multi-target DPA’ has been described as ‘higher-order DPA’ (HODPA), but this label has become ambiguous, with some preferring to apply it to the use of higher-order statistical moments rather than to the exploitation of multivariate targets.
We might expect a method which does not require pre-processing to produce stronger attack outcomes, as it will be exploiting all the information in the trace points. Applications of multivariate mutual information (MMI) have shown some promise: for the bivariate attacks in [89] they proved more efficient than correlation DPA in the presence of high noise, and in the examples in [12] they also compare favourably (depending on the method of estimation). Gierlichs et al. [48] present a once-masked scenario in which two-target MMI outperforms correlation DPA using absolute difference pre-processing, and also a twice-masked scenario in which three-target MMI succeeds but correlation DPA with product pre-processing fails, which they use to argue that the cost of combined-target correlation DPA grows exponentially with target dimension while that of MMI grows subexponentially.

Previous evaluations have been empirical and have covered only a limited range of leakage scenarios. Moreover, with the exception of recent work in [36], the focus has almost exclusively been on standard Hamming weight power model variants. We address these gaps with a theoretic analysis of attack outcomes in a variety of scenarios, using both standard and generic (or near-generic) power models (similar to our analysis of single-target attacks in Chapter 4). In so doing we hope to better understand the inherent capabilities of explicitly multivariate non-profiled distinguishers. In addition to MMI-based DPA we consider a bivariate extension of the Kolmogorov–Smirnov (KS) based distinguisher as proposed by [73].

Another potential application of multivariate statistics is to incorporate multiple leakage points (and the corresponding hypothesis-dependent predictions) from an unprotected implementation. This is somewhat inspired by the well-documented advantages of multivariate profiled attacks over univariate counterparts [30, 92]. The idea of extending MI-based DPA in this manner was hinted towards briefly in [47] (a pre-print version of [48]), although more with a focus on targeting multiple noisy measurements relating to the same target. Either way, this has not since been followed up to our knowledge. We clarify the advantages—or rather, as it turns out, the disadvantages—of such an approach.

---

2\(^2\)We are not aware of a reasonable multivariate extension to the Cramér-von Mises criterion or the variance ratio, so we are unable at this time to extend the analysis to all the distinguishers considered in the univariate analysis of Chapter 4.
In this chapter.

In what follows we will first introduce the multi-target adaptations of correlation-, MI- and KS-based DPA distinguishers (§5.2). We will then, in §5.3, apply our evaluation framework to attacks against DES and PRESENT under a variety of leakage assumptions in both protected and unprotected scenarios. (We are unable to extend this analysis to AES because the complexity of the computations precludes calculating the theoretic distinguishing vectors in the presence of masking).

5.2 Multi-target distinguishers

In Chapter 4 we introduced single-target formulations of distinguishers based on Pearson’s correlation coefficient, mutual information, the Kolmogorov–Smirnov test statistic, the Cramér–von Mises criterion, and the variance ratio. In this section, we introduce the two-target extensions which have been proposed for the first three of these; the latter two have no existing extensions, which we leave as further work.

5.2.1 Pearson’s correlation coefficient

Pearson’s correlation coefficient has no natural multivariate extension, but it has been adapted for use in multi-target DPA attacks against masked implementations by means of a data pre-processing step in which multivariate trace measurements are mapped to a univariate trace which is then compared with the model predictions in the usual way [29, 76]. A two-target CPA distinguisher takes the form:

\[ D_{\text{PP}_r}^{\rho}(k) = |\rho(C(P_1, P_2), M_k)| = \left| \frac{\text{cov}(C(P_1, P_2), M_k)}{\sqrt{\text{var}(C(P_1, P_2)) \cdot \text{var}(M_k)}} \right|, \]  

(5.1)

where \( C : \mathcal{P}_1 \times \mathcal{P}_2 \rightarrow \mathcal{P}_C \) is the pre-processing function. We choose \( C \) to be the normalised product: \( C(P_1, P_2) = (P_1 - \mathbb{E}(P_1)) \times (P_2 - \mathbb{E}(P_2)) \), as representing the best available in the literature (at least in the case of known Hamming weight leakage—see [90]).
5.2.2 Mutual information

MI as defined in Chapter 4 (§4.2.2) is bivariate in that it quantifies the information between the distributions of (specifically) two random variables—but those random variables need not be scalar, they can just as well be vectors (without altering the definition). That is, if \( A = (A_1, \ldots, A_n) \) and \( B = (B_1, \ldots, B_m) \) then

\[
I(A; B) = \sum_{a \in A_1 \times \ldots \times A_n} \sum_{b \in B_1 \times \ldots \times B_m} p_{A,B}(a, b) \log_2 \left( \frac{p_{A,B}(a, b)}{p_A(a) p_B(b)} \right),
\]

where \( p_{A,B} \) is the combined joint probability density of \((A_1, \ldots, A_n, B_1, \ldots, B_m)\) and \( p_A, p_B \) are the joint densities of \((A_1, \ldots, A_n)\) and \((B_1, \ldots, B_m)\).

Therefore, a natural way to extend MI-based DPA to a multi-target strategy is simply to quantify the information shared between tuples of trace points and/or model predictions taken jointly. In the once-masked scenario we attack a pair of leakage measurements (associated with the masked S-Box and the mask) via a univariate prediction based on the unmasked S-Box. To disambiguate this from the single-target strategy we coin the shorthand description ‘joint’ MI (JMI), with the distinguisher taking the form:

\[
D_{\text{JMI}^m}(k) = I((P_1, P_2); M_k) = H(P_1, P_2) - H(P_1, P_2|M_k). \tag{5.2}
\]

When, instead, we wish to target two functions in a standard unprotected scenario, we are able to predict each intermediate value explicitly. We therefore seek to quantify the information shared between the pair of trace values and the corresponding pair of hypothesis-dependent predictions:

\[
D_{\text{JMI}^u}(k) = I((P_1, P_2); (M_1, M_2)_k) = H(P_1, P_2) - H(P_1, P_2|(M_1, M_2)_k). \tag{5.3}
\]

An alternative approach to multi-target MI-based DPA is via multivariate MI (MMI), which quantifies the information shared between \( n \) random variables (scalar or otherwise)
$A_1, A_2, \ldots, A_n$ and is defined as

$$I(A_1; A_2; \ldots; A_n) = I(A_1; A_2; \ldots; A_{n-1}) - I(A_1; A_2; \ldots; A_{n-1}|A_n),$$

where $I(A_1; A_2; \ldots; A_{n-1}|A_n) = \mathbb{E}_{a \in A_n}[I(A_1; A_2; \ldots; A_{n-1}|A_n = a)]$. MMI has the interesting property that it can actually take negative values in the case where a third variable proves to be a common effect for two otherwise independent variables (the symmetry of the definition implies that the relationship is independent of the choice of which of the variables to consider the ‘conditioning’ variable). This is precisely the nature of the masking scenario we consider. Therefore, when attacking a masked S-Box by combining the associated leakage measurement with one arising from the mask itself we look for the key hypothesis that produces the the largest negative MMI value:

$$D_{\text{MMI}}^m(k) = -I(P_1; P_2; M_k) = I(P_1; P_2) - I(P_1; P_2|M_k). \quad (5.4)$$

Again, in the unprotected scenario, we are able to predict each intermediate value explicitly, so the MMI is between four variables, rather than three, and we are now interested in the hypothesis producing the largest positive value.

$$D_{\text{MMI}}^u(k) = I(P_1; P_2; M_{1,k}; M_{2,k}). \quad (5.5)$$

Both of these extensions were explored thoroughly in [12] and demonstrated to be essentially (theoretically, asymptotically) equivalent in the case of a perfectly implemented masking scheme (i.e. ensuring that $I(P_1; M_{k^*}) = I(P_2; M_{k^*}) = 0$), which is precisely the scenario we consider for our two-target attacks. However, since they exhibited some differences in practical attacks we consider both to be of interest for further theoretic evaluation.
5.2.3 Kolmogorov–Smirnov

Multivariate extensions of the KS test are somewhat more difficult to achieve, as we first need to formulate an appropriate notion of a multivariate CDF. In the one-dimensional case there are only two ways of ordering the data, namely \( P(A \geq x) \) and \( P(A \leq x) \). As we have that \( P(A \geq x) = 1 - P(A \leq x) \) the choice turns out to be arbitrary.

In higher dimensions the choice of ordering is no longer inconsequential: there is no direct way to map (e.g.) between \( P(A_1 \leq x, A_2 \leq y) \) and \( P(A_1 \geq x, A_2 \leq y) \). In fact for \( d \) different random variables, there are \( 2^d \) possible orderings we need to consider. Peacock (in [85]) proposes to define the KS distance as the maximum distributional difference taken over all the orderings, so that (for example) in the bivariate case:

\[
K(A_1, A_2||B_1, B_2) = \max \left\{ \sup_{(x, y) \in (A_1 \cup B_1) \times (A_2 \cup B_2)} \left| F_{A_1, A_2}^{(i)}(x, y) - F_{B_1, B_2}^{(i)}(x, y) \right| \right\}_{i=1}^{4},
\]

where \( S = (A_1 \cup B_1) \times (A_2 \cup B_2) \) and \( F^{(1)} \) to \( F^{(4)} \) are the CDFs based on all four possible orderings. He shows that a bivariate KS test statistic according to this approach is close enough to being distribution-free to be useful in practice. Inevitably, this impacts not just the data complexity of estimation but also the computational costs; in spite of various optimisations proposed in [42], extensions to dimensions greater than 2 quickly become infeasible.

For two-target attacks against a masked implementation we adapt the KS-based distinguisher to compare the global joint CDF of the traces with the joint CDFs as partitioned by the model predictions under each key hypothesis (this is conceptually comparable to the JMI-based distinguisher):

\[
D_{\text{JKS}}^{m}(k) = \mathbb{E}[K(P_1, P_2||P_1, P_2|M_k)]
= \mathbb{E}_{m \in \mathcal{M}} \left[ \max_{y_1, y_2} \left\{ \sup_{y_1, y_2} \left| F_{P_1, P_2}^{(i)}(y_1, y_2) - F_{P_1, P_2|m_k=m}(y_1, y_2) \right| \right\}_{i=1}^{4} \right]. \tag{5.6}
\]

Similarly, following the construction of the JMI-based distinguisher for unprotected implementations we consider a KS analogue whereby the global joint CDFs are compared
with the joint CDFs as conditioned on the bivariate predictions:

\[
D_{JKS^u}(k) = \mathbb{E}[K(P_1, P_2||P_1, P_2|(M_1, M_2)_k)]
= \mathbb{E}_{(m_1, m_2) \in M_1 \times M_2} \left[ \max \left\{ \sup_{y_1, y_2} \left| F^{(i)}_{P_1, P_2}(y_1, y_2) - F^{(i)}_{P_1, P_2|(M_1, M_2)_k=(m_1, m_2)}(y_1, y_2) \right| \right\}^{4}_{i=1} \right].
\]

(5.7)

5.3 Comparative evaluation

We consider the same adversary and leakage scenarios as in Chapter 4. To briefly recap, the adversary is non-profiling and either uses a (standard) Hamming weight power model, the (generic) identity power model (when targeting a non-injective function) or the (‘near-generic’) \((m-1)\)-LSB power model (when targeting an injective function).

The evaluation is with respect to three leakage assumptions: an optimistic scenario in which the data-dependent leakage really is proportional to the Hamming weight (so favouring the standard model); a realistic in which the true leakage is an unevenly weighted sum of the bits; and a challenging in which the true leakage is a highly nonlinear function of the leaked value. (See Chapter 4 (§4.4) for full details).

In this instance we are only able to test our distinguishers against the DES and PRESENT block ciphers, as the 8-bit AES S-Box (× 8 mask bits) proves too large for the theoretic vectors to be computed. (Of course, a more space-efficient computation may be possible; for now, we are satisfied that the analysis of attacks targeting PRESENT gives us good indication of how the \((m-1)\)-LSB variants perform, which results we may reasonably expect to extend to AES).

As in Chapter 4 we will consider both the theoretic performance in a noise-free setting and the impact of independent Gaussian noise of varying size.
5.3.1 Targeting DES

In DES, 6-bit chunks of plaintext are XOR-ed with 6 key bits before being input into one of eight S-Boxes, each with 4-bit outputs. As discussed in Chapter 2, the non-injectivity of the S-Boxes makes them amenable to generic DPA strategies, where an appropriately-chosen distinguisher is provided with a nominal identity mapping—i.e. the hypothesis-dependent predicted values themselves. As before (in Chapter 4) we also consider standard Hamming weight power model attacks as this is the mode in which multivariate extensions of MI-based DPA have most frequently been used in the literature to date.

Multi-target attacks against unprotected DES

The noise-free setting To put our multi-target attacks in context we first consider standard Hamming weight attacks against DES AddRoundKey. The outcomes in the noise-free setting (block 2 of Table 5.1) nicely illustrate the reduced effectiveness of DPA against linear target functions. Even in the optimistic scenario the distinguishers are no longer able to uniquely identify the correct key, instead ranking it equally with its bitwise complement \( \bar{k} \). The standard scores are reduced relative to those reported for the S-Box attacks with the standard power model; MI and KS again seem to have an advantage over correlation but are no longer robust to severe model degradation.

Block 3 of Table 5.1 relates to multivariate extensions of MI- and KS-based DPA by which we attempt to enhance outcomes by jointly targeting AddRoundKey and the first S-Box. Although we are exploiting a larger amount of information, this increase applies across the range of key hypotheses so that distinguishing margins are not automatically increased. In fact, even in the optimistic scenario the true key is less strongly distinguished (in relative and even absolute terms) than in the attacks against the S-Box alone (Figure 5.1 illustrates this for the MI-based distinguishers), and, at least in the case of JMI\(^a\) and JKS\(^a\) variants, with a substantial increase in the support size required.

The attacks moreover appear less robust to model degradation, and, as with the AddRoundKey attacks, they actually fail in the challenging scenario. Of the two MI versions
Figure 5.1: Single-target and two-target MI-based theoretic distinguishing vectors against unprotected DES with noise-free Hamming weight leakage.

considered, JMI\textsuperscript{u} exhibits larger margins but appears to require a substantially larger input support—likely to impact on practical efficiency.

Generic single-target attacks against AddRoundKey do not work because the target is injective (the correlation attack using the identity succeeds with order 2 because of the aforementioned correlation between the Hamming weight and the actual values between 0 and 15, but this is not really of practical interest). Interestingly, because the mapping to (AddRoundKey,S-Box) output pairs is also injective, the generic JMI\textsuperscript{u} and JKS\textsuperscript{u} strategies produce flat distinguishing vectors (block 6). Meanwhile, generic MMI\textsuperscript{u} (I(P\textsubscript{1};P\textsubscript{2};M\textsubscript{1,k};M\textsubscript{2,k})) continues to succeed but, as with the standard power model attacks, margins are smaller (and critical support larger) than those of the generic single-target attacks against the S-Box alone.

The impact of noise on distinguishing margins Figures 5.2 and 5.3 confirm that, in each of our three leakage scenarios, the advantage of the single-target attacks over
the two-target persists across the tested noise range (and for both standard and generic variants), thus reinforcing our conclusion that more information does not necessarily imply greater attack effectiveness. (Standard MMI\textsuperscript{u} even fails in sufficiently noisy optimistic and realistic scenarios).

This discovery appears counter-intuitive alongside what is known about other classes of multivariate attack. Template attacks, in particular, are known to be enhanced by the incorporation of multiple data points (see [30, 92]).

However, the two are not as conceptually analogous as they may appear. The use of multiple points in a template attack can be thought of as a refining process whereby candidate subsets from different leakage points are intersected to narrow in on the most likely hypothesis. A more suitable comparator would be two separate MI-based attacks in which each of AddRoundKey and the S-Box were only able to reveal reduced lists of likely keys (for example because of insufficient trace measurements); the intersection between the two lists would then reduce uncertainty about the key. By contrast, our multivariate distinguishers as used here have more the effect (though this is somewhat of a simplification) of implicating the union between two such lists; since AddRoundKey exhibits similar leakage for similar keys, using it as a second target confounds rather than refines knowledge on the key. By this reasoning, any such multi-target attack can at best achieve the theoretic success of the strongest of all the univariate attacks taken separately (and will more likely have diminished outcomes).

Attacks against masked DES

The noise-free setting  We next consider an implementation in which the outputs of the first DES S-Box are XOR-ed with a uniformly distributed random mask of the same length as the S-Box outputs (i.e. 4 bits).\textsuperscript{3} Blocks 7 and 8 of Table 5.1 report the ideal outcomes of two-target attacks in such a scenario, using standard and generic power models.

\textsuperscript{3}This scenario may not be realistic as masks are more likely take up the full word length (e.g. 8 bits in the case of a typical 8-bit microcontroller) but is sufficient for our illustrative purposes and moreover keeps computational complexity at a manageable level.
In the optimistic scenario the relative distinguishing margins for all four two-target attacks do not appear greatly reduced from those of their (unprotected) single-target counterparts; the penalty in absolute terms is more evident. Here, the difference between JMI\textsuperscript{univ} and
$\text{MMI}^m$ really becomes apparent. Whilst both have identical margins\(^4\), the latter has far smaller input support requirements (in fact, no greater than the single-target attack on an unmasked implementation) which may make it the preferable of the two when it comes to practical implementation.

$\text{MMI}^m$ is therefore the most robust to the masking with an absolute margin of 57\% compared with 45\% for correlation and just 14\% for $\text{JKS}^m$, and very minimal average critical support. The critical support sizes for $\text{JMI}^m$ and correlation are comparable and modest compared with the substantially inflated requirements for $\text{JKS}^m$ (note that the full support set comprises all the possible input-mask pairs—$64 \times 16 = 1024$ in total).

Interestingly, correlation with pre-processing continues to succeed even in the challenging scenario, where single-target correlation against an unprotected implementation has been found to fail. Standard $\text{JMI}^m$ and $\text{JKS}^m$ also continue to succeed, but $\text{MMI}^m$ is decidedly the most robust to degradation of the model. It appears that, theoretically at least, $\text{MMI}^m$ could be the most well suited of the tested distinguishers to exploiting masked leakage in unusual leakage scenarios. Its advantage over correlation—which relies on pre-processing—has been anticipated in the literature (in particular, [48]) but it is interesting to note that standard $\text{JKS}^m$ no longer achieves close performance in spite of its conceptual similarity to $\text{JMI}^m$ and consistently similar behaviour in univariate scenarios.

As with single-target attacks (against unprotected implementations), generic variants of $\text{MMI}^m$, $\text{JMI}^m$ and $\text{JKS}^m$-based DPA do not have an advantage over standard power model variants in the optimistic scenario (to which the power model is well suited). In this instance they have comparable margins (the $\text{MMI}^m$/$\text{JMI}^m$ lower margins, in fact) and a larger critical support. However, they are, once again, far more robust to degradation of the leakage. The margins do not decrease and there are not the same penalties in terms of additional support required (in fact, in the ‘realistic’ scenario critical support seems to drop; evidently a quirk of the joint distribution under that particular leakage function. We noted similar patterns for univariate attacks against this leakage function in Chapter 4).

The extremely low critical support required by $\text{MMI}^m$ is interesting to note, but the

\(^4\)This is due to the information theoretic relationship $I(A; (B, C)) = I(A; B) + I(A; C) - I(A; B; C)$ and the fact that the masking is perfectly implemented so that $I(P_1; M_{k*}) = I(P_2; M_{k*}) = 0$. 

110
estimation of multi-dimensional densities will remain a costly enterprise, particularly if the typical DPA literature heuristic of ‘one bin per leakage value’ is followed.

**The impact of noise on distinguishing margins** Figure 5.4 indicates that the impact of noise on two-target attacks against masking can be quite profound. In the optimistic scenario the relative margin achieved by standard $\text{MMI}^m/\text{JMI}^m$ is enhanced by noise, whilst in the realistic scenario noise works to the substantial detriment of standard $\text{MMI}^m/\text{JMI}^m$ until the SNR is at least 8. In the challenging scenario noise appears to enhance the relative margins of both $\text{MMI}^m/\text{JMI}^m$ and $\text{JKS}^m$.

In the optimistic scenario the advantages of the power model over the generic variants persists across the tested range. In the realistic scenario generic $\text{MMI}^m/\text{JMI}^m$ do not suffer the same noise impact as their standard counterparts. The generic advantage in the challenging scenario appears to be compromised by noise.

In absolute terms the $\text{MMI}^m/\text{JMI}^m$ margins are substantially reduced by noise, requiring a strong signal before they begin to approach their pure-signal capabilities. Correlation and $\text{JKS}^m$ appear more robust, so that in spite of their inferiority in noise-free settings they may yet be preferable in certain noisy settings.
Figure 5.4: Theoretic relative and absolute distinguishing margins as the (Gaussian) noise varies, for two-target attacks against the first DES S-Box in a masked implementation.
Table 5.1: Theoretic outcomes against DES in optimistic, realistic and challenging leakage scenarios without noise. (Note that MMI and JMI coincide in the univariate case, hence we have combined these into one column where appropriate).

<table>
<thead>
<tr>
<th></th>
<th>Optimistic</th>
<th></th>
<th>Realistic</th>
<th></th>
<th>Challenging</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Standard S-Box attacks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct key ranking (order)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Standard score</td>
<td>5.14</td>
<td>6.59</td>
<td>5.95</td>
<td></td>
<td>3.21</td>
<td>6.38</td>
</tr>
<tr>
<td>Relative margin</td>
<td>3.56</td>
<td>5.61</td>
<td>4.24</td>
<td></td>
<td>1.22</td>
<td>5.12</td>
</tr>
<tr>
<td>Absolute margin</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
<td>0.30</td>
<td>0.86</td>
</tr>
<tr>
<td>Average critical support</td>
<td>6</td>
<td>8</td>
<td>8</td>
<td></td>
<td>17</td>
<td>10</td>
</tr>
<tr>
<td>Critical support for 90% SR</td>
<td>8</td>
<td>11</td>
<td>11</td>
<td></td>
<td>32</td>
<td>14</td>
</tr>
<tr>
<td>Critical support for 100% SR</td>
<td>16</td>
<td>19</td>
<td>19</td>
<td></td>
<td>49</td>
<td>21</td>
</tr>
<tr>
<td>2. Standard AddRoundKey attacks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct key ranking (order)</td>
<td>1 (2)</td>
<td>1 (2)</td>
<td>1 (2)</td>
<td></td>
<td>1 (2)</td>
<td>1 (4)</td>
</tr>
<tr>
<td>Standard score</td>
<td>2.62</td>
<td>4.48</td>
<td>3.76</td>
<td></td>
<td>2.24</td>
<td>3.44</td>
</tr>
<tr>
<td>Relative margin</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Absolute margin</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td></td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Average critical support</td>
<td>6</td>
<td>9</td>
<td>9</td>
<td></td>
<td>26</td>
<td>10</td>
</tr>
<tr>
<td>Critical support for 90% SR</td>
<td>8</td>
<td>12</td>
<td>12</td>
<td></td>
<td>47</td>
<td>15</td>
</tr>
<tr>
<td>Critical support for 100% SR</td>
<td>14</td>
<td>21</td>
<td>21</td>
<td></td>
<td>58</td>
<td>23</td>
</tr>
<tr>
<td>3. Standard two-target attacks (unprotected)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct key ranking (order)</td>
<td>–</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>–</td>
<td>1</td>
</tr>
<tr>
<td>Standard score</td>
<td>–</td>
<td>3.91</td>
<td>6.04</td>
<td>4.70</td>
<td>–</td>
<td>4.47</td>
</tr>
<tr>
<td>Relative margin</td>
<td>–</td>
<td>2.47</td>
<td>3.66</td>
<td>1.93</td>
<td>–</td>
<td>2.93</td>
</tr>
<tr>
<td>Absolute margin</td>
<td>–</td>
<td>0.28</td>
<td>0.64</td>
<td>0.41</td>
<td>–</td>
<td>0.41</td>
</tr>
<tr>
<td>Average critical support</td>
<td>–</td>
<td>7</td>
<td>14</td>
<td>14</td>
<td>–</td>
<td>10</td>
</tr>
<tr>
<td>Critical support for 90% SR</td>
<td>–</td>
<td>10</td>
<td>16</td>
<td>16</td>
<td>–</td>
<td>15</td>
</tr>
<tr>
<td>Critical support for 100% SR</td>
<td>–</td>
<td>15</td>
<td>25</td>
<td>25</td>
<td>–</td>
<td>33</td>
</tr>
</tbody>
</table>

Continued on next page...
Table 5.1: Theoretic outcomes against DES in optimistic, realistic and challenging leakage scenarios without noise (continued).

<table>
<thead>
<tr>
<th></th>
<th>Optimistic</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Corr.</td>
<td>MMI</td>
<td>JMI</td>
<td>K-S</td>
</tr>
<tr>
<td>Correct key ranking (order)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Standard score</td>
<td>5.39</td>
<td>6.35</td>
<td>6.20</td>
<td></td>
</tr>
<tr>
<td>Relative margin</td>
<td>3.61</td>
<td>5.08</td>
<td>4.60</td>
<td></td>
</tr>
<tr>
<td>Absolute margin</td>
<td>0.85</td>
<td>0.68</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>Average critical support</td>
<td>9</td>
<td>16</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Critical support for 90% SR</td>
<td>14</td>
<td>19</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>Critical support for 100% SR</td>
<td>27</td>
<td>24</td>
<td>24</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Realistic</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Corr.</td>
<td>MMI</td>
<td>JMI</td>
<td>K-S</td>
</tr>
<tr>
<td>Correct key ranking (order)</td>
<td>1 (2)</td>
<td>1</td>
<td>1</td>
<td>(64)</td>
</tr>
<tr>
<td>Standard score</td>
<td>1.68</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Relative margin</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Absolute margin</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Average critical support</td>
<td>8</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Critical support for 90% SR</td>
<td>11</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Critical support for 100% SR</td>
<td>23</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Challenging</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Corr.</td>
<td>MMI</td>
<td>JMI</td>
<td>K-S</td>
</tr>
<tr>
<td>Correct key ranking (order)</td>
<td>–</td>
<td>1</td>
<td>1</td>
<td>(64)</td>
</tr>
<tr>
<td>Standard score</td>
<td>–</td>
<td>5.35</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Relative margin</td>
<td>–</td>
<td>3.87</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Absolute margin</td>
<td>–</td>
<td>0.37</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Average critical support</td>
<td>–</td>
<td>20</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Critical support for 90% SR</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Critical support for 100% SR</td>
<td>–</td>
<td>23</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

Continued on next page...
Table 5.1: Theoretic outcomes against DES in optimistic, realistic and challenging leakage scenarios without noise (continued).

<table>
<thead>
<tr>
<th></th>
<th>Optimistic</th>
<th></th>
<th>Realistic</th>
<th></th>
<th>Challenging</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>7. Standard attacks against masking</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct key ranking (order)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Standard score</td>
<td>5.14</td>
<td>6.16</td>
<td>6.16</td>
<td>5.39</td>
<td>2.46</td>
<td>5.90</td>
</tr>
<tr>
<td>Relative margin</td>
<td>3.56</td>
<td>4.55</td>
<td>4.55</td>
<td>3.72</td>
<td>0.35</td>
<td>4.48</td>
</tr>
<tr>
<td>Absolute margin</td>
<td>0.45</td>
<td>0.57</td>
<td>0.57</td>
<td>0.14</td>
<td>0.07</td>
<td>0.62</td>
</tr>
<tr>
<td>Average critical support</td>
<td>28</td>
<td>8</td>
<td>29</td>
<td>115</td>
<td>75</td>
<td>11</td>
</tr>
<tr>
<td>Critical support for 90% SR</td>
<td>50</td>
<td>10</td>
<td>46</td>
<td>219</td>
<td>189</td>
<td>16</td>
</tr>
<tr>
<td>Critical support for 100% SR</td>
<td>137</td>
<td>16</td>
<td>74</td>
<td>385</td>
<td>894</td>
<td>24</td>
</tr>
<tr>
<td>8. Generic attacks against masking</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct key ranking (order)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>29</td>
<td>1</td>
</tr>
<tr>
<td>Standard score</td>
<td>5.39</td>
<td>5.48</td>
<td>5.48</td>
<td>5.48</td>
<td>-0.01</td>
<td>6.36</td>
</tr>
<tr>
<td>Relative margin</td>
<td>3.61</td>
<td>3.78</td>
<td>3.78</td>
<td>3.99</td>
<td>-2.42</td>
<td>4.92</td>
</tr>
<tr>
<td>Absolute margin</td>
<td>0.38</td>
<td>0.42</td>
<td>0.42</td>
<td>0.10</td>
<td>-0.35</td>
<td>0.73</td>
</tr>
<tr>
<td>Average critical support</td>
<td>39</td>
<td>14</td>
<td>52</td>
<td>175</td>
<td>-</td>
<td>13</td>
</tr>
<tr>
<td>Critical support for 90% SR</td>
<td>76</td>
<td>18</td>
<td>81</td>
<td>344</td>
<td>-</td>
<td>16</td>
</tr>
<tr>
<td>Critical support for 100% SR</td>
<td>204</td>
<td>23</td>
<td>135</td>
<td>531</td>
<td>-</td>
<td>20</td>
</tr>
</tbody>
</table>
5.3.2 Targeting PRESENT

The PRESENT S-Box maps 4-bit inputs (XOR-ed with 4 secret key bits) to 4-bit outputs and is thus injective. It therefore enables us to test the ‘near-generic’ \((m - 1)\)-LSB power models proposed in [49] for applying MI and other generic-compatible distinguishers against injective targets. (The 8-bit injective AES S-Box proved too large for similar analysis).

Multi-target attacks against unprotected PRESENT

The noise-free setting  Block 3 of Table 5.2 reports the outcomes of attacks seeking to exploit the joint leakage of AddRoundKey and the S-Box output. Interestingly, the MMI\textsuperscript{a} attack succeeds in the challenging scenario even though both single-target attacks fail. In all other cases (as witnessed previously in the case of DES) the two-target distinguishers offer no advantages over standard attacks against the S-Box alone, with much reduced margins and increased critical support, where attacks succeed at all (the KS-based distinguisher fails in the ‘realistic’ setting even though the univariate attacks both succeed).

As for the ‘near-generic’ attacks using the 3LSB power model, even the single-target S-Box attacks fail in most cases, as we have already seen in Chapter 4. The joint MI and KS attacks produce flat distinguishing vectors whilst the multivariate MI approach ranks the true key last, again, similarly to the univariate case.

The impact of noise on distinguishing margins  That the two-target disadvantage persists across the tested SNR range is demonstrated in Figure 5.5. The surprising success of MMI\textsuperscript{a} in the challenging leakage scenario turns out to be noise-dependent; indeed, many of the multivariate attacks which succeed in the perfect signal scenario fail as noise increases.

In the realistic scenario, just as the univariate MI distinguisher with the 3LSB power model is successful (with very small margins) within a small SNR range, so is the multivariate
counterpart—but, again, with reduced margins. Two-target 3LSB attacks in the other two leakage scenarios fail across the tested range.

Figure 5.5: Theoretic relative and absolute distinguishing margins as the (Gaussian) noise varies, for two-target attacks against AddRoundKey and the S-Box in an unprotected implementation of PRESENT.

Figure 5.6: Theoretic relative and absolute distinguishing margins as the (Gaussian) noise varies, for ‘near-generic’ two-target attacks against AddRoundKey and the S-Box in an unprotected implementation of PRESENT.
Attacks against masked PRESENT

The noise-free setting  Blocks 7 and 8 of Table 5.2 report the ideal outcomes of (standard and ‘near-generic’) two-target attacks against (once-)masked PRESENT.

Against this smaller target, the standard attacks are less robust to degradation of the leakage, just as are the standard (unprotected) single-target equivalents (see blocks 1 and 2 of the table, and Chapter 4 (§4.4.4) for more details). The KS-based distinguisher even fails in the masked realistic scenario despite succeeding in the unprotected.

The ‘near-generic’ results support our reasoned argument of Chapter 2 and the existing evidence of Chapter 4 that the \((m - 1)\)-LSB power model does not perform an equivalent function against injective targets to that of the identity (i.e. generic) power model against non-injective targets. In fact, they do not even recover the key in the optimistic scenario (though the JKS\(^m\) variant is fourth-order successful in the realistic scenario).

The impact of noise on distinguishing margins  Figure 5.7 shows the evolution of the distinguishing margins as the SNR increases. The so-called ‘near-generic’ variants fail to distinguish the key across the tested range in all scenarios, as do the standard variants in the challenging scenario.

The (successful) performance of the standard variants is fairly level in the optimistic scenario, but the performance of the MMI\(^m\)/JMI\(^m\) attacks are observed to be critically noise-dependent in the realistic scenario, so that it only recovers the key after a (very high) threshold SNR of about 128—a reminder that noise sensitivity can affect outcomes in a surprising way, which we would not necessarily be able to predict.
Figure 5.7: Theoretic relative and absolute distinguishing margins as the (Gaussian) noise varies, for two-target attacks against the PRESENT S-Box in a masked implementation.
Table 5.2: Theoretic outcomes against PRESENT in optimistic, realistic and challenging leakage scenarios without noise. (Note that MMI and JMI coincide in the univariate case, hence we have combined these into one column where appropriate).

|                  | Optimistic |                  |                  |                |                  |                  |                  |
|------------------|------------|------------------|------------------|----------------|------------------|------------------|
| 1. Standard S-Box attacks |            |                  |                  |                |                |                  |                  |
| Correct key ranking (order) | 1          | 1                | 1                |                | 1               | 1                | 1                |                |
| Standard score    | 2.52       | 2.94             | 2.79             |                | 1.44            | 3.17             | 2.07             |                |
| Relative margin   | 1.06       | 2.09             | 1.94             |                | 0.25            | 2.08             | 0.62             |                |
| Absolute margin   | 1.00       | 1.00             | 1.00             |                | 0.20            | 0.63             | 0.22             |                |
| Average critical support | 4          | 7                | 7                |                | 10              | 9                | 9                |                |
| Critical support for 90% SR | 6          | 9                | 9                |                | 15              | 12               | 14               |                |
| Critical support for 100% SR | 10         | 13               | 13               |                | 16              | 14               | 16               |                |
| 2. Standard AddRoundKey attacks |            |                  |                  |                |                |                  |                  |
| Correct key ranking (order) | 1 (2)      | 1 (2)            | 1 (2)            |                | 1 (2)           | 1 (4)            | 1 (2)            |                |
| Standard score    | 1.89       | 2.53             | 2.32             |                | 1.73            | 1.73             | 2.41             |                |
| Relative margin   | 0.00       | 0.00             | 0.00             |                | 0.00            | 0.00             | 0.00             |                |
| Absolute margin   | –          | –                | –                |                | –               | –                | –                |                |
| Average critical support | 5          | 6                | 6                |                | 10              | 6                | 8                |                |
| Critical support for 90% SR | 6          | 8                | 8                |                | 15              | 10               | 12               |                |
| Critical support for 100% SR | 8          | 12               | 12               |                | 15              | 13               | 15               |                |
| 3. Standard two-target attacks (unprotected) |            |                  |                  |                |                |                  |                  |
| Correct key ranking (order) | –          | 1                | 1                | 1              | –               | 1                | 1 (16)           | 14             |
| Standard score    | –          | 2.72             | 2.54             | 2.08           | –               | 2.66             | 2.45             | 1.32           |
| Relative margin   | –          | 1.11             | 1.02             | 0.67           | –               | 1.66             | 1.63             | -0.73          |
| Absolute margin   | –          | 0.54             | 0.33             | 0.30           | –               | 0.69             | 0.17             | -0.15          |
| Average critical support | –          | 6                | 10               | 10             | –               | 8                | 14               | –             |
| Critical support for 90% SR | –          | 8                | 13               | 13             | –               | 11               | 16               | –             |
| Critical support for 100% SR | –          | 12               | 15               | 15             | –               | 15               | 16               | –             |

Continued on next page...
Table 5.2: Theoretic outcomes against PRESENT in optimistic, realistic and challenging leakage scenarios without noise (continued).

<table>
<thead>
<tr>
<th></th>
<th>Optimistic</th>
<th></th>
<th></th>
<th>Realistic</th>
<th></th>
<th></th>
<th>Challenging</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>4. Near-generic S-Box attacks</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct key ranking (order)</td>
<td>1</td>
<td>14</td>
<td>5</td>
<td></td>
<td></td>
<td>1</td>
<td>9</td>
<td>1 (2)</td>
<td></td>
</tr>
<tr>
<td>Standard score</td>
<td>2.45</td>
<td>-1.41</td>
<td>0.50</td>
<td></td>
<td></td>
<td>2.10</td>
<td>-0.90</td>
<td>1.65</td>
<td></td>
</tr>
<tr>
<td>Relative margin</td>
<td>1.19</td>
<td>-2.83</td>
<td>-1.45</td>
<td></td>
<td></td>
<td>1.00</td>
<td>-2.41</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Absolute margin</td>
<td>0.98</td>
<td>-0.50</td>
<td>-0.35</td>
<td></td>
<td></td>
<td>0.30</td>
<td>-0.33</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Average critical support</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>12</td>
<td></td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Critical support for 90% SR</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>16</td>
<td></td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Critical support for 100% SR</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>16</td>
<td></td>
<td>16</td>
<td></td>
</tr>
<tr>
<td><strong>5. Near-generic AddRoundKey attacks</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct key ranking (order)</td>
<td>1 (4)</td>
<td>1 (16)</td>
<td>1 (16)</td>
<td></td>
<td></td>
<td>1 (4)</td>
<td>1 (16)</td>
<td>1 (16)</td>
<td></td>
</tr>
<tr>
<td>Standard score</td>
<td>1.34</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
<td></td>
<td>1.34</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Relative margin</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Absolute margin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average critical support</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Critical support for 90% SR</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Critical support for 100% SR</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>6. Near-generic two-target attacks (unprotected)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct key ranking (order)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard score</td>
<td>-1.98</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
<td></td>
<td>-1.89</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Relative margin</td>
<td>-3.59</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
<td></td>
<td>-3.77</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Absolute margin</td>
<td>-0.83</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
<td></td>
<td>-0.67</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Average critical support</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Critical support for 90% SR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Critical support for 100% SR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Continued on next page...
Table 5.2: Theoretic outcomes against PRESENT in optimistic, realistic and challenging leakage scenarios without noise (continued).

<table>
<thead>
<tr>
<th></th>
<th>Optimistic</th>
<th></th>
<th>Realistic</th>
<th></th>
<th>Challenging</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>7. Standard attacks against masking</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct key ranking (order)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Standard score</td>
<td>2.52</td>
<td>2.79</td>
<td>2.79</td>
<td>2.76</td>
<td></td>
</tr>
<tr>
<td>Relative margin</td>
<td>1.06</td>
<td>1.84</td>
<td>1.84</td>
<td>2.05</td>
<td></td>
</tr>
<tr>
<td>Absolute margin</td>
<td>0.25</td>
<td>0.66</td>
<td>0.66</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>Average critical support</td>
<td>20</td>
<td>6</td>
<td>23</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>Critical support for 90% SR</td>
<td>42</td>
<td>8</td>
<td>39</td>
<td>119</td>
<td></td>
</tr>
<tr>
<td>Critical support for 100% SR</td>
<td>105</td>
<td>11</td>
<td>59</td>
<td>185</td>
<td></td>
</tr>
<tr>
<td><strong>8. Near-generic attacks against masking</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct key ranking (order)</td>
<td>1</td>
<td>16</td>
<td>16</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Standard score</td>
<td>2.45</td>
<td>-2.07</td>
<td>-2.07</td>
<td>-0.18</td>
<td></td>
</tr>
<tr>
<td>Relative margin</td>
<td>1.19</td>
<td>-3.90</td>
<td>-3.90</td>
<td>-2.34</td>
<td></td>
</tr>
<tr>
<td>Absolute margin</td>
<td>0.25</td>
<td>-0.86</td>
<td>-0.86</td>
<td>-0.08</td>
<td></td>
</tr>
<tr>
<td>Average critical support</td>
<td>29</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Critical support for 90% SR</td>
<td>67</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Critical support for 100% SR</td>
<td>141</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5.4 Conclusion

In this chapter we have demonstrated that true generic strategies offer extremely scenario-robust (theoretical) key recovery in multi-target attacks against masked S-Boxes. This concurs with what we have already discovered about single-target generic strategies in unprotected scenarios (Chapter 4). However, they evidently carry hefty data overheads as in practice they require estimating a greater number of conditional densities/cumulative densities.

In addition, the observed gains over correlation-based methods in most cases do not appear large enough to compensate for the substantial increase in computational complexity. This is particularly significant when we consider a real-life attack scenario. One of the hardest problems of multi-target attacks is point selection [30, 92]: even if the candidate time intervals for each target value can be reduced to small windows they still need to be evaluated pair-wise until a peak is discovered. With complex distinguishers like MMI this becomes prohibitively costly.

Generic multi-target strategies, then, are theoretically robust but may not always be practical. By contrast, so-called ‘near-generic’ strategies using the \((m - 1)\)-LSB power models are not equivalently robust when the targets are injective. This is again in line with the single-target analysis of Chapter 4 and our reasoning in Chapter 2 around the inescapable requirement for meaningful prior knowledge unless the target provides a non-injective mapping itself.

Moreover, we have shown that multi-target attacks combining two (non-masked) points in the trace are generally less effective than a single-target attack against the most DPA-vulnerable of the two targets (an S-Box, in this instance).
Chapter 6

Generic strategies against
Hamming distance leakage and
DRP logic

6.1 Introduction

The Hamming weight model is very popular in the literature, but, as we discussed in Chapter 1, §1.1.4, it is generally the bit transitions as words are processed which produce data-dependencies in the leakage of common technologies such as CMOS. Hence the previous state of a targeted device plays a role of equal importance to the current state. This implicates the Hamming distance model as a reasonable approximation for the leakage function: if \( r = \{r_1, \ldots, r_m\} \) is the reference state of the circuit before the target word \( v = \{v_1, \ldots, v_m\} \) is processed, then the consumption is proportional to the number of bit flips between \( r \) and \( v \), that is to say \( L(v) = HD(v, r) = HW(v \oplus r) \). Inherent in such a model is the assumption that \( 0 \to 1 \) transitions and \( 1 \to 0 \) transitions consume the same power, and also that consumption is unaffected by the position and inter-relation of the bits in the word.

When the Hamming weight model is thought or known to be reasonable this implies the
special, simplified case that the reference state is always pre-charged to 0 (or 1), so that the number of transitions is simply the number of ‘on’ bits in the word (or the number of ‘off’ bits, which of course produces leakage proportional to the Hamming weight but with a negative coefficient). Since this scenario is both desirable in the context of DPA and reasonable in practice, it has become the standard setting for much of the literature (perhaps artificially reinforcing the superiority of correlation DPA as a strategy).

In this chapter we consider the usefulness of generic DPA in cases where this simplifying assumption does not hold. We can, broadly speaking, class them into three scenarios:

1. The previous state is known to the attacker. This might be fixed or data-dependent (for example, another intermediate value of the algorithm which can be computed from a known function). In such cases the target can be attacked with equivalent efficacy as a Hamming weight attack against Hamming weight leakage: the known state can simply be incorporated into the predictions, preserving the accuracy.

2. The previous state is unknown to the attacker but fixed. Brier et al. [19] have shown how to adapt correlation DPA in order to determine the state $R$ as an unknown of the problem in addition to $F_{k^*}(X) \oplus R$, which together reveals the secret key $k^*$. Thus it adds to the complexity of correlation DPA but does not thwart it. One of the questions we address in this chapter is whether or not generic strategies can be useful in bypassing this need to simultaneously recover $R$.

3. The previous state is unknown to the attacker and can vary. If it is random and independent of the data then the problem is equivalent to that of attacking a masked implementation, and higher-order methods will be required. Whether or not first-order attacks are possible when the unknown state is in some way data-dependent remains to be seen and is another of the questions of interest in the following.

We will explore these questions by first applying the framework of Chapter 3 to compare the ideal/theoretic capabilities (as defined in Chapter 1 §1.2.1) of mutual information (MI) based DPA with those of correlation DPA, and then performing practical attacks against simulated traces with a view to experimentally verifying our theoretic results. We
will also discuss other scenarios to which this analysis can be supposed to apply—namely, dual-rail precharge (DRP) logic and imperfect masking schemes.

### 6.2 Constant reference states

To begin with, suppose, as in [19], that the reference state is a constant but unknown machine word $R$. The device no longer leaks $\text{HW}(F_k^*(X))$ but rather $\text{HW}(R \oplus F_k^*(X)) = \text{HD}(R, F_k^*(X))$.

First observe that no attack against a linear target function such as AddRoundKey can achieve first-order success, because the ‘true key’ values are perfectly replicated under an incorrect key hypothesis, namely $k^* \oplus R$. The power consumption for a plaintext $X$ will be proportional to $\text{HD}((k^* \oplus X), R) = \text{HW}((k^* \oplus X) \oplus R) = \text{HW}((k^* \oplus R) \oplus X)$, so that when our hypothesis is $k = k^* \oplus R$ we get maximum correlation/MI (for both HW and generic models) and in fact the theoretical distinguishing vector is identical to that of a successful attack against HW leakage with a key of $k^* \oplus R$.

Targeting highly nonlinear functions such as S-Boxes avoids this predicament. For example, if $S_{\text{DES1}}$ is the first DES S-Box, we know that there is no $R'$ such that $S_{\text{DES1}}(k^* \oplus X) \oplus R = S_{\text{DES1}}((k^* \oplus R') \oplus X) \forall X \in \mathcal{X}$, so no incorrect key will produce the correct predictions. It remains to be seen whether the resemblance between the imperfect predictions (with naive power models) and the true power consumption remains strong enough for the correct key and weak enough for the alternative hypotheses for any sort of attack to be successful.

Thus motivated, we investigate the theoretic and practical outcomes of MI- and correlation-based DPA attacks against the first DES S-Box in the presence of a constant but unknown reference state. We assume that the attacker either proceeds naively with the HW power model or exploits the non-injectivity of the target function by way of an MI-based generic strategy.
Theoretic outcomes against pure-signal leakage

When there is no noise, correlation DPA (with a naive HW power model) succeeds precisely in those scenarios where the HW of the reference is 1 (or 0) and fails whenever it is 2 (see Table 6.1). Further, were we to use the absolute value of the correlation to distinguish the resulting attack would succeed whenever the HW of the reference state is 3 or 4; however, there is a substantial reduction in theoretic strength when the HW is 1 or 3, and for some reference states (absolute) correlation DPA requires almost the entire plaintext set to determine the correct key.

MI-based DPA paired with the HW power model succeeds for any reference state and gains a considerable advantage both in terms of the distinguishing scores with full information (nearest-rival scores are in the range of 3.6 to 4.5 for power model MI but just 0.5-2.7 for absolute correlation) and also in terms of the minimum input support required for success (on average, 14 to 15 for power model MI compared with 17 to 18 for absolute correlation).

The authors of [49] already observed that generic MI-based DPA is essentially unaffected by a constant reference state, and this is indeed what we observe here: the nearest-rival distinguishing score is always around 5 for the generic strategy and average support requirement around 16. This means that when $R \in \{0000_{(2)}, 1111_{(2)}\}$ (i.e. $L$ is the HW function) the generic attacks are less effective than the equivalent methods combined with a HW power model—which we would expect, according to Chapter 2, as they trade a perfect (nominal) model for the less specific generic one—but for all other reference state scenarios they gain an advantage.

Theoretic outcomes as SNR varies

We now explore whether the advantages observed in a pure-signal setting persist in the presence of noise. Consider the particular case that the device leaks the Hamming distance from 0100_{(2)}: from Table 6.1 above, we know this to be one instance in which correlation DPA still successfully recovers the key but is substantially disadvantaged (in the pure-
Table 6.1: Theoretical strength of correlation- and MI-based attacks against DES with Hamming distance leakage from a constant reference state.

<table>
<thead>
<tr>
<th>4 LSBs of reference state</th>
<th>Correlation (raw)</th>
<th>Correlation (absolute)</th>
<th>Standard MI</th>
<th>Generic MI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hamming weight 0</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct key ranking</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Standard score</td>
<td>3.61</td>
<td>5.14</td>
<td>6.59</td>
<td>6.35</td>
</tr>
<tr>
<td>Relative margin</td>
<td>2.14</td>
<td>3.56</td>
<td>5.61</td>
<td>5.08</td>
</tr>
<tr>
<td>Absolute margin</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.68</td>
</tr>
<tr>
<td>Average critical support</td>
<td>6</td>
<td>6</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>Critical support for 90% SR</td>
<td>8</td>
<td>8</td>
<td>11</td>
<td>19</td>
</tr>
<tr>
<td>Critical support for 100% SR</td>
<td>12</td>
<td>12</td>
<td>14</td>
<td>22</td>
</tr>
<tr>
<td><strong>Hamming weight 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct key ranking</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Standard score</td>
<td>2.04-4.65</td>
<td>2.56-4.94</td>
<td>5.48-5.97</td>
<td>5.81-6.46</td>
</tr>
<tr>
<td>Relative margin</td>
<td>0.38-2.28</td>
<td>0.53-2.65</td>
<td>3.60-4.47</td>
<td>4.57-5.20</td>
</tr>
<tr>
<td>Absolute margin</td>
<td>0.16-0.47</td>
<td>0.12-0.41</td>
<td>0.43-0.51</td>
<td>0.65-0.70</td>
</tr>
<tr>
<td>Average critical support</td>
<td>17-25</td>
<td>20-34</td>
<td>14-15</td>
<td>16-17</td>
</tr>
<tr>
<td>Critical support for 90% SR</td>
<td>31-49</td>
<td>33-53</td>
<td>20-22</td>
<td>19-20</td>
</tr>
<tr>
<td>Critical support for 100% SR</td>
<td>40-59</td>
<td>44-61</td>
<td>28-32</td>
<td>21-24</td>
</tr>
<tr>
<td><strong>Hamming weight 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct key ranking</td>
<td>27-32</td>
<td>54-63</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Standard score</td>
<td>0.00</td>
<td>-1.94-1.15</td>
<td>5.06-5.53</td>
<td>5.98-6.43</td>
</tr>
<tr>
<td>Relative margin</td>
<td>-3.30-1.65</td>
<td>-5.62-3.52</td>
<td>3.05-3.16</td>
<td>4.49-5.42</td>
</tr>
<tr>
<td>Absolute margin</td>
<td>-0.63-0.32</td>
<td>-0.71-0.53</td>
<td>0.31-0.33</td>
<td>0.64-0.73</td>
</tr>
<tr>
<td>Average critical support</td>
<td>-</td>
<td>-</td>
<td>17-18</td>
<td>16-16</td>
</tr>
<tr>
<td>Critical support for 90% SR</td>
<td>-</td>
<td>-</td>
<td>20-22</td>
<td>19-20</td>
</tr>
<tr>
<td>Critical support for 100% SR</td>
<td>-</td>
<td>-</td>
<td>33-36</td>
<td>22</td>
</tr>
<tr>
<td><strong>Hamming weight 3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct key ranking</td>
<td>64</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Standard score</td>
<td>-4.05-2.04</td>
<td>2.56-4.94</td>
<td>5.48-5.97</td>
<td>5.81-6.46</td>
</tr>
<tr>
<td>Relative margin</td>
<td>-6.32-3.69</td>
<td>0.53-2.65</td>
<td>3.60-4.47</td>
<td>4.57-5.20</td>
</tr>
<tr>
<td>Absolute margin</td>
<td>-1.58-1.32</td>
<td>0.12-0.41</td>
<td>0.43-0.51</td>
<td>0.65-0.70</td>
</tr>
<tr>
<td>Average critical support</td>
<td>-</td>
<td>20-34</td>
<td>14-15</td>
<td>16-17</td>
</tr>
<tr>
<td>Critical support for 90% SR</td>
<td>-</td>
<td>33-53</td>
<td>20-22</td>
<td>19-20</td>
</tr>
<tr>
<td>Critical support for 100% SR</td>
<td>-</td>
<td>44-61</td>
<td>28-32</td>
<td>21-24</td>
</tr>
<tr>
<td><strong>Hamming weight 4</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct key ranking</td>
<td>64</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Standard score</td>
<td>-3.61</td>
<td>5.14</td>
<td>6.59</td>
<td>6.35</td>
</tr>
<tr>
<td>Relative margin</td>
<td>-5.30</td>
<td>3.56</td>
<td>5.61</td>
<td>5.08</td>
</tr>
<tr>
<td>Absolute margin</td>
<td>-2.47</td>
<td>1.00</td>
<td>1.00</td>
<td>0.68</td>
</tr>
<tr>
<td>Average critical support</td>
<td>-</td>
<td>6</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>Critical support for 90% SR</td>
<td>-</td>
<td>8</td>
<td>11</td>
<td>19</td>
</tr>
<tr>
<td>Critical support for 100% SR</td>
<td>-</td>
<td>12</td>
<td>14</td>
<td>22</td>
</tr>
</tbody>
</table>
noise case, at least) relative to the generic MI-based strategy.

Figure 6.1 shows the impact of Gaussian noise on theoretic attack effectiveness, both in terms of nearest-rival distinguishing margin and in terms of the minimum support size required for first-order success. The strategy which pairs the MI-based distinguisher with the (standard) HW power model is not very robust to the addition of noise: the distinguishing margin actually falls below that of the correlation-based distinguisher, and there is a hefty penalty in terms of required support size. By contrast, the generic strategy is far more robust and even exhibits some evidence of stochastic resonance-type behaviour so that distinguishing power (as measured by our margins, which are standardised) is actually slightly enhanced by noise up to a certain threshold. Moreover, the critical support size for the generic strategy remains constant in the tested range.

**Figure 6.1:** Relative margins and critical support of theoretic attacks against Hamming distance leakage (with a reference state of $0100_{(2)}$) for varying levels of Gaussian noise.

**Practical outcomes (simulations)**

In order to get some idea of whether the perceived theoretic advantages of the generic strategy can be translated into practice, we test all three approaches against simulated traces comprising Hamming distance leakage (from $0100_{(2)}$) of the first DES S-Box with independent Gaussian noise, as the SNR varies. We use the sample correlation coefficient to estimate the correlation, as standard; we use histogram-based estimators for MI, where bin counts are chosen equal to the cardinality of the power model domain, according to the heuristic which has emerged from the literature (see, for example, [12]). That is, for the model-based strategy we use 5 bins, for the generic strategy, 16. In a pure-signal scenario
(see the dashed lines in Figure 6.2) the 5-bin estimator for MI with a model requires fewer traces than correlation DPA to identify the correct key, but the introduction of even the smallest amount of noise incurs a burden so that across the tested range it is substantially less efficient. By contrast, the 16-bin estimator for the generic strategy approaches the efficiency achieved in the pure-signal scenario as the SNR increases, and moreover substantially outperforms correlation DPA once the SNR is at least 1. We have thus confirmed that—in this instance at least—the theoretic advantages of generic strategies can be translated into practice.

![Mean traces required for key recovery](image)

**Figure 6.2:** Average number of traces required for key recovery in simulated practical attacks against Hamming-distance leakage (with a reference state of 0100(2)), for varying levels of Gaussian noise.

These results further confirm what we have argued all along: that it is not the choice of *distinguisher* which qualifies a strategy as ‘generic’ (as was originally supposed), rather the appropriate *pairing* of a distinguisher with the generic power model, as defined in Chapter 2. Strategies which combine so-called ‘generic distinguishers’ such as MI with power models still require for those models to be meaningful. In this instance, the HW power model is incorrect but adequately accurate for correlation DPA to work—so it is no surprise that MI-based DPA using the HW model also works. However, the theoretic noise sensitivity of the latter, combined with the inherent data overhead of *estimating* MI, means that there are no particular advantages to favouring that approach over correlation. By contrast, the generic strategy uses a model which is perfectly precise (although lacking perfect recall—see Chapter 2) and so able to supply genuine theoretic and practical advantages over any ‘wrong power model’ strategy.
Indeed, we have shown that generic strategies applied with little consideration for or knowledge about the true leakage can be effective even when that leakage actually depends on an unknown reference state. Correlation DPA, applied equally blindly, is far less likely to yield a successful attack. However, as [19] showed, it is possible to adapt correlation DPA in a strategy which simultaneously recovers $R$ along with $F_{k^*}(X) \oplus R$ (thereby, of course, revealing the secret key $k^*$). Whilst this simultaneous search process is more computationally costly than a standard correlation attack, generic strategies can themselves be computationally costly (for example, requiring many more bins in the case that histograms are estimated) in addition to the likely data complexity overheads. Further work (and broader cost considerations) would be required to establish which of the two methods is most practical.

6.2.1 Relationship with DRP logic

We observe an important and useful parallel between HD leakage and the expected behaviour of DPA-resistant dual-rail precharge (DRP) logic. In fact, an imperfect realisation of the logic style can be shown to exhibit data-dependent power consumption of a similar form to the HD from a constant reference state, enabling us to clarify its vulnerability to the ‘generic’ MI-based attack described by Gierlichs et al. in [49].

DRP logic attempts to eradicate the data-dependency of the power consumption by making it equal in each clock cycle. This is achieved insofar as the capacitances of the complementary output wires in each logic gate can be balanced, a difficult feat in practice [87]. Suppose the $i^{th}$ bit of an $m$-bit word $x$ is carried by a DRP logic gate driving two differential outputs with imperfectly balanced capacities ($\alpha_i, \beta_i$), so that $\alpha_i = \beta_i + \gamma_i$. The power consumption of such a circuit can be shown to be equivalent to leakage scenarios with which we are more familiar, enabling us to comment on theoretical attack capabilities.

Let us initially consider the simplified case that both capacitances are the same throughout the circuit: $\beta_i = \beta$, $\alpha_i = \beta + \gamma$, $\forall i \in \{0, \ldots, m - 1\}$. Then the data-dependent leakage is
proportional to:

\[ \text{HW}(x)\alpha + \text{HW}(\bar{x})\beta = \text{HW}(x)(\beta + \gamma) + \text{HW}(\bar{x})\beta \]
\[ = (\text{HW}(x) + \text{HW}(\bar{x}))\beta + \text{HW}(x)\gamma \]
\[ = m\beta + \text{HW}(x)\gamma \]

The constant \(m\beta\) is absorbed into the non-data-dependent component and we thus obtain the result that the leakage is proportional to the Hamming weight. Both correlation DPA and MI-based DPA using the HW power model will be theoretically capable of returning the correct key; practical success will depend on ability and resources to estimate the distinguishing vectors with sufficient precision, and, as we have already established, correlation is likely to have the advantage.

Now suppose that the capacitances are the same throughout the circuit but that the order changes, i.e. so that some gates have capacitances \((\alpha, \beta)\) and others \((\beta, \alpha)\), where \(\alpha = \beta + \gamma\). We can express this by introducing \(R = (r_0, \ldots, r_{m-1}) \in \{0, 1\}^m\) such that gate \(i\) is \((\beta, \alpha)\) if \(r_i = 1\) and \((\alpha, \beta)\) otherwise. Then the data-dependent leakage is:

\[ \text{HW}(x \oplus R)\alpha + \text{HW}(x \oplus \bar{R})\beta = \text{HW}(x \oplus R)(\beta + \gamma) + \text{HW}(x \oplus \bar{R})\beta \]
\[ = (\text{HW}(x \oplus R) + \text{HW}(x \oplus \bar{R}))\beta + \text{HW}(x \oplus R)\gamma \]
\[ = m\beta + \text{HW}(x \oplus R)\gamma \]

That is, the data-dependent leakage is proportional to the Hamming distance from \(R\), which equates to the scenario of a more conventional logic style (such as CMOS) consuming power proportional to the number of transitions from a constant, unknown reference state. We have already shown that the generic MI-based strategy remains ideally successful against such leakage, whilst correlation DPA is (depending on the state) either unsuccessful or greatly reduced in distinguishing power. This confirms that DRP logic gives rise to leakage scenarios under which first-order generic strategies could be useful, in particular, shedding light on the experimental result of [49].

In the most general case, the size of the capacitances and not just the direction of the
differences may vary over the circuit. Suppose the gates corresponding to bits \( i = 1, \ldots, m \) have capacitances \((\alpha_i, \beta_i)\) such that \( \alpha_i = \beta_i + \gamma_i \) where \( \gamma_i \) can be positive or negative.

Letting \( x = (x_1, \ldots, x_m) \) and \( \alpha = (\alpha_1, \ldots, \alpha_m) \), \( \beta = (\beta_1, \ldots, \beta_m) \), \( \gamma = (\gamma_1, \ldots, \gamma_m) \) we get a leakage function of \( x \cdot \alpha + (x \oplus 1) \cdot \beta = (x + x \oplus 1) \cdot \beta + x \cdot \gamma = 1 \cdot \beta + x \cdot \gamma \), so that the data-dependent power consumption is proportional to a weighted combination of the bits of \( x \), where the weights can take negative values. Further investigation is needed to establish the expected behaviour of our distinguishers as the relative weights become increasingly disproportionate. Of course, since generic-compatible strategies interpret the power model nominally, they are unlikely to be affected by such factors.

### 6.3 Data-dependent reference states

As mentioned previously, if the reference state varies randomly and independently of the data then the situation is equivalent to random masking and requires some sort of higher-order strategy. The case that there are a limited number of different states, assigned according to the plaintext input and fixed in repeated runs, could be produced by an incorrect implementation of a masking scheme. For the sake of simplicity we ignore, for now, the impact of noise, and investigate attack outcomes when \( R \) is of this form as the number of possible different states increases.

In the (commonly studied) case of an 8-bit micro-controller, the reference states (or masks) take values in \( \{0, 1\}^8 = \{0, \ldots, 255\} \). Since our attacks on the first DES S-Box target 6-bit key portions, our plaintext inputs are drawn from \( \{0, 1\}^6 = \{0, \ldots, 63\} \)—there could be up to 64 different input-dependent reference states. The number of possible ways that \( r \) reference states could be associated with the 64 input values is given by the Stirling number of the second kind: \( \left\{ \frac{64}{r} \right\} = \frac{1}{r!} \sum_{j=0}^{r} (-1)^{r-j} \binom{r}{j} j^{64} \), so it is no longer possible to exhaustively explore every scenario. Instead, we calculate the success rates in 1,000 random experiments for increasing numbers of different reference states, randomly assigned to approximately equal-sized subsets of the input space (see Figure 6.3). \(^1\)

\(^1\)When the reference state is constant, only the 4 bits which are replaced by the S-Box output contribute to the data-dependent leakage whilst the contribution of the remaining bits is absorbed into the static
We find that both MI-based distinguishers are better able to succeed than correlation DPA—although even the generic strategy does not achieve 100% success for attacks with more than 2 different states and for more than 6 states success rates drop to below 50%. The success of correlation DPA degrades rapidly: for attacks with about 20 different states it is no better than a random guess, whilst the generic strategy and even MI with the HW model appear to retain some advantage over guessing.

Thus, when very little is known about the leakage an attacker may well be able to recover a great deal of information just by applying a ‘blind’ MI-based strategy—though even success in the noise-free setting will be partially determined by chance, and the number of traces required for adequate estimation may be prohibitive. Such an approach may not be the best way of exploiting the available data: where resources permit, it may prove more effective or efficient to refine a correlation-based approach (or similar), investing greater effort in understanding the leakage to begin with, perhaps through profiling.

![Figure 6.3: Theoretic success (in the absence of noise) against the first DES S-Box in the presence of data-dependent reference states of length 8 bits, as the number of different states increases.](image)

component of the power consumption. However, when the state depends on the data in the manner described here, the contribution of the remaining bits does need to be taken into consideration as it becomes part of the data-dependent power consumption.
6.4 Conclusion

In this chapter we have considered a particular, highly relevant leakage scenario where generic strategies may offer certain advantages, namely, that of Hamming distance leakage from unknown reference states. An adversary naively assuming a Hamming weight power model will not necessarily (depending on the state) be able to recover the key using correlation DPA, and even when this is possible there will be a loss in efficiency due to the diminished quality of the model.

We investigate the use of generic strategies, as exemplified by MI-based DPA paired with the generic power model, comparing their theoretic and practical capabilities with those of naive correlation DPA. As expected, the success of generic MI-based DPA is robust to a constant reference state and, for once, demonstrably capable of outperforming correlation DPA.

Against a number of reference states, in the case that they are data-dependent and fixed in repeated runs, MI-based DPA appears more resilient to correlation DPA, at least while the number of different states is low (say, less than 6).

Not only do Hamming distance leakages occur naturally in CMOS logic, they can also (we have shown) arise from imperfectly-balanced DRP logic technologies—whilst the ‘multiple states’ scenario may arise as a result of an imperfectly implemented masking scheme. This goes some way to confirming the existence of real-life scenarios in which generic strategies such as MI-based DPA may prove practically useful.
Chapter 7

Linear regression-based DPA

7.1 Introduction

The techniques of linear regression (LR) were first applied to side-channel analysis in the context of profiling DPA, where the adversary is assumed to be in possession (and full control) of a device identical to that targeted and thereby able to build accurate models of the (deterministic and random) data-dependent side-channel leakage. The so-called ‘stochastic attacks’ of [97] use LR to do this: the leakage is characterised in function of several targeted bits, producing a power model for use in a second, ‘key recovery’ stage. The method was expected to be more efficient (in the profiling stage) than the Bayesian classification-style template approach [30] thanks to some reasonable simplifying assumptions. Even this very first paper recognised the potential for the technique to be adapted to the task of non-profiled key recovery, but left the particulars to further work: the precise details, and a performance evaluation, were eventually given in [37]. A two-target variant (with normalised product pre-processing) for use against masked implementations was recently introduced in [36].

We will explain the attack mechanism more fully in §7.2, but in essence it may be compared to a correlation DPA in which the power model is unknown a priori and is revealed simultaneously with the correct key in a manner which has been described as ‘on-the-fly’
profiling. Thanks to its minimal dependency on prior knowledge it has been promoted as a ‘generic’ distinguisher (informally speaking). However, as with other so-called generic distinguishers, it has been observed to require something in the way of meaningful insight about the true leakage—in this case, a suitable (non-exhaustive) basis of covariates from which to build the regression equations [119].

In this chapter we explore LR-based DPA in the context of the formalised understanding of ‘generic strategies’ proposed in Chapter 2. We show that it can be applied as a generic-compatible distinguisher even though the sense in which it accepts the generic power model is not immediately obvious. Of course, as a generic strategy (in this strict, formal sense) it is subject to all of the same limitations as other generic strategies—in particular, it can only be used effectively against non-injective target functions meeting some minimal nonlinearity-related criteria. However, it turns out that forcing it into the generic strategy paradigm ignores some very natural non device-specific intuition about the leakage which, when taken into consideration, leverages what can be learned from the (key-dependent) estimated models over and above what may be concluded from the LR-based distinguisher values alone. This motivates a novel twist on the standard operation: we show that the techniques of stepwise regression can be used in such a way as to produce successful outcomes even in scenarios where strictly generic strategies are known to fail!1

The rest of this chapter proceeds as follows: §7.2 introduces (standard) LR-based DPA, explaining the mechanism by which it distinguishes the correct key (§7.2.1 explores the power model assumptions exploited by the distinguisher and §7.2.2 demonstrates that it is among the class of generic-compatible distinguishers); §7.3 presents the stepwise regression-inspired variant; §7.4 discusses multi-target extensions of both standard and stepwise approaches; §7.5 verifies the feasibility of our methods by quantifying the asymptotic capabilities of LR-based distinguishers in a range of practically meaningful unprotected and Boolean masked scenarios, targeting functions relating to DES, AES, and PRESENT block ciphers.2

1The ‘non device-specific intuition’ in question relates to the supposition that even if the leakage function is a high degree polynomial of the target bits, yet it will remain ‘simpler’ or ‘more orderly’ (in some sense) than a well-chosen DPA target such as an S-Box; quite what we mean will become clearer when we examine the distinguishing mechanism of the attack (see §7.3).

2In the study of generic DPA, it is appropriate that feasibility should take precedence over efficiency
7.2 Linear regression-based distinguishers

The motivation for an LR-based approach begins with the observation that $L : \mathbb{F}_2^n \rightarrow \mathbb{R}$ can be viewed as a pseudo-Boolean vectorial function with a unique expression in numerical normal form [26]. That is to say, there exists coefficients $\alpha_u \in \mathbb{R}$ such that $L(z) = \sum_{u \in \mathbb{F}_2^n} \alpha_u z^u$, $\forall z \in \mathbb{F}_2^n$ ($z^u$ denotes the monomial $\prod_{i=1}^m z_i^u$ 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 $F$.

LR is a statistical method designed for just such tasks. It models the relationship between a single dependent variable $Y$ and one or more explanatory variables $Z$, 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’, $\beta$ is the $p$-dimensional vector of unknown parameters, and $\varepsilon$ is the noise or error term, that is, all remaining variation in $y$ which is not caused by $Z$. (See Appendix 7.A for more details). 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 between the outcomes and their predicted values: $R^2 = r(Z\hat{\beta}, y)^2$ where $r$ is the sample estimate of Pearson’s correlation coefficient, defined for two random variables $A, B$ as $\rho(A, B) = \frac{\text{Cov}(A, B)}{\sqrt{\text{Var}(A)\text{Var}(B)}}$. 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 $D_{\rho}(k) = \rho(L \circ F_k(X) + \varepsilon, M \circ F_k(X))$. In LR-based DPA the challenge is to simultaneously recover the true power model along with the correct key as follows:

---

as a matter of interest, since the existence and scope of such strategies has proved so difficult to establish (see Chapter 2).
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:

$$D_{LR}(k) = \rho(L \circ F_k^\ast(X) + \varepsilon, \hat{\alpha}_{k,0} + \sum_{u \in U} F_k(X)^u \hat{\alpha}_{k,u})^2. \quad (7.1)$$

7.2.1 The ‘power model’ in LR-based DPA

In the way the distinguisher is most naturally presented, the attacker’s prior knowledge is contained within $U$; it is not immediately obvious what, precisely, we might call the ‘power model’ in this context, or where it fits alongside the various types presented in Chapter 2. In fact, each $u \in U$ 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 $\{x \in \mathbb{F}_2^n | F_k(x)^u = 0\}$.

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 non-uniform changes to the output. Such, in fact, is the mechanism by which the original difference-of-means distinguisher of [63] operates!

So the LR 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 $u \in U$ which does not induce a ‘meaningful’ partition

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

\[\text{The discussion in [37] further helps to frame LR-based DPA in the context of existing methodologies.}\]
will detract from the overall distinguishing power—hence the motivation to use any prior knowledge available to refine $U$.

Intuitively, the generic instantiation should correspond to $U = \mathbb{F}_2^m \setminus \{0\}$, 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 $U = \mathbb{F}_2^m \setminus \{0\}$ 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. This re-framing is argued as follows:

If $M_k$ is an arbitrary labelling 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 $M'_k(x)^u = \sum v \in \mathbb{F}_2^m b_v F_k(x)^v$, $b_v \in \mathbb{R}$ [26]. 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 $\alpha_u$, $u \in \mathbb{F}_2^m$ whenever we re-parametrise in this way: but, crucially, the terms in the equation collectively explain the measured traces equally well—and it is in this sense that LR-based DPA is invariant to re-labelling and therefore can be discussed alongside other generic-compatible strategies (though it is not usually used in this way—particularly as restrictions on $U$ contribute to efficiency gains in the estimation stage).

Limitations

At this stage it is important (and enlightening) to clarify some senses in which LR-based DPA is not invariant to relabelling:

1. The model coefficients produced by the correct key hypothesis will only give a representation of the ‘true leakage model’ under the ‘correct’ labelling.

2. Trace efficiency at the estimation stage will depend on the complexity of the model
equations (i.e. the number of truly contributory terms) and so is likely to be min-
imised under the ‘correct’ labelling.

3. The representation of the multiple binary power models as a single, nominal power
model does not extend to the general case \( \mathcal{U} \subset \mathbb{F}_2^m \setminus \{0\} \), as the implicit restriction on
the allowable terms prevents us from re-writing the permuted values in polynomial
function of the output bits.

7.2.2 Generic operation

By the above argument, we expect the outcomes of an LR-based DPA with \( \mathcal{U} = \mathbb{F}_2^m \setminus \{0\} \)
to be influenced by the properties of \( F \) similarly to other attacks accepting the generic
power model. But it is informative to explore quite how the power model-based argument
developed in §7.2.1 plays out in this specific instance, as aspects of the precise mechanism
reveal information that is not inferable from the distinguishing vector of \( R^2 \)'s. We focus
(for clarity, and because we are more interested in feasibility than efficiency) on the
asymptotic case—which is to say we assume that sufficient trace measurements have been
collected for the attacker to ‘average out’ the noise (as per the law of large numbers) so
that we only need to consider the behaviour of the ideal distinguishing vector
\( \mathbf{D}_{\text{IDEAL}} = \{ D(L \circ F_k^* (X), M \circ F_k (X)) \}_{k \in \mathcal{K}}. \)

If the target function is non-injective, then \( L \) can be expressed as a system of \( 2^n \) equations
(in function of \( F_k (x) \)) with \( 2^m \) unknowns. In other words, it is \underline{overdetermined}. Under the
correct key hypothesis, the \underline{perfect precision} of the generic model guarantees consistency
in the system: there are only \( 2^m \) \underline{linearly independent} equations and the system has a
unique solution—and consequently an \( R^2 \) of 1. Provided the target satisfies the criteria
in §2.3.1 then a wrong-key candidate introduces inconsistency into the system so that the
resulting least squares approximation is not a perfect fit (\( R^2 < 1 \)) and the candidate is
distinguished from the true one.

\footnote{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.}
If the target function is instead injective, then the data-dependent part of the power consumption can be expressed as a system of $2^n$ equations in $2^n$ unknowns: this is always consistent and therefore always has a perfect solution. 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 $\mathcal{U} \subset \mathbb{F}_2^m \setminus \{0\}$).

In practice, the insight required to make a sufficiently specific meaningful restriction of the type exemplified above can be out of reach for an attacker who ends up resorting to guesswork [37]. But we note that the LR output gives us more than just the $R^2$s forming the distinguisher values: among other things, we have the estimated (key-dependent) model coefficients. Even though, when $\mathcal{U} = \mathbb{F}_2^m \setminus \{0\}$, these give an equally good fit, still, 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 $L$.

This motivates the idea that maybe some kind of non device-specific intuition, combined with the key-dependent regression coefficients (when $\mathcal{U} = \mathbb{F}_2^m \setminus \{0\}$), could be used to help identify the correct expression for $L$ and therefore the correct key. This would be a step removed from a generic-compatible strategy because of its reliance on said ‘intuition’ and on the assumption that $M_k = F_k$; just how large a step would depend on the nature of the intuition. In the next section we introduce a learning technique from data mining which is able to exploit a remarkably non-specific ‘sense of what the leakage function should look like’ to produce, in a wide range of leakage scenarios, asymptotically successful key
recovery against injective targets (and more efficient key recovery against non-injective targets). Because the strategy works even when provided with a full set of covariates \( U = \mathbb{F}_2^n \setminus \{0\} \), it may reasonably be described as \textit{generic-emulating}.

7.3 A stepwise regression-based distinguisher

We showed in §7.2.1 that the leakage function \( L \) can be written just as readily in polynomial function of the incorrectly predicted bits \( F_k(X) \) as it can in function of the correct bits \( F_{k^*}(X) \). But those expressions can look very different, as we now illustrate for the example scenario 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 7.1 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 and PRESENT 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.

\[ \begin{array}{c}
\text{AES S–Box} \\
\text{PRESENT S–Box} \\
\text{DES S–Box}
\end{array} \]

\text{Figure 7.1: Coefficients, in the polynomial expression for} \( L \), \text{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. 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.
Stepwise regression [57] 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 7.B for full details). The resulting regression model should therefore exclude ‘unimportant’ terms whilst retaining all of the ‘significant’ terms. In the context of LR-based 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 $\mathcal{U} = \mathbb{F}_2^m \setminus \{0\}$, 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 7.1 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 7.1 also reinforces the intuition (discussed in §2.3.1) that S-Box vulnerability increases with size: the extent to which explanatory power can be dispersed among the covariates under wrong-key re-parametrisation 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 very situations in which it could be most useful) 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.

§7.5 demonstrates the stepwise approach in action: we find that it can succeed (asymptotically) against injective targets, even with high-degree leakage, and moreover can improve the trace efficiency of attacks against non-injective targets by widening the margins by which the true key is distinguished from the alternatives.

First, though, we introduce a multi-target strategy for overcoming masked implementations, which can equally take advantage of the stepwise regression ‘trick’.

7.4 Multi-target extensions

Profiling DPA with an LR-based preliminary stage has natural extensions to multivariate attacks (for example, against masked implementations), as presented by Schindler [96]. In the case that all inputs and masks are controlled by the profiler, LR models are built for the deterministic part of the leakage at different time points in the measured trace (for example the points at which a mask and the corresponding masked S-Box output are processed), and then the joint distribution of the noise is approximated (for example, via a multivariate Gaussian distribution). The key of the attacked device can subsequently be recovered by examining the likelihood of the observed sample being produced by each of the profiled key-dependent hypothetical distributions and choosing the one for which this is greatest (the so-called ‘maximum likelihood’ method).

In the case that the attacker does not know or control the random masks of the profiled device, LR-based profiling is trickier. Similarly to template attacks [67], the task now becomes one of estimating mixture densities with unknown components. A brief sketch of a work-around is mentioned in [96] but the paper does not go into details.
A suggestion for extending non-profiled LR-based DPA to higher dimensions has been put forward by [36]. They describe a two-target attack in which the two trace measurements (i.e. relating to a mask and the corresponding masked output) are combined in a pre-processing step before performing standard LR-based DPA on the resulting univariate variable. This is, of course, similar to the way in which correlation DPA has been extended to the masked scenario. The justification for this approach is that, whilst the (conditional) leakages of the (uniformly random) mask and the masked S-Box output cannot be supposed to depend on the unmasked S-Box output, the conditional covariance between the two can. It is an estimate of this covariance which is computed as the pre-processed one-dimensional variable on the LHS of the LR equation. The authors of [36] are able to verify the effectiveness of this approach.

If, then, the S-Box output is masked with a uniformly distributed random mask $Q$, and the attacker has access to side channel leakage of both the mask and the masked output (of functional forms $L_1$, $L_2$, not necessarily the same) a two-target LR-based DPA proceeds by computing the leakage covariances conditional on the input $X$ and fitting hypothesis-dependent least-squares linear models in function of the bits of the (known) predictions for the (unmasked) S-Box output. This can be represented by the following distinguisher:

$$D_{2TLR}(k) = \rho(\text{cov}(L_1(Q), L_2(Q \circ F_k(X)))|X) + \zeta, \tilde{\eta}_{k,0} + \sum_{u \in \mathcal{U}} F_k(X)^u \tilde{\eta}_{k,u}^2. \quad (7.2)$$

The conditional sample covariance is equivalent to the (input-averaged) product-combining function demonstrated by [90] to be the most efficient pre-processing function (for correlation-based DPA) of those known in the literature to date. The same paper shows how to derive ‘optimal’ prediction functions via the conditional expectations of the covariances of known leakage functions. The functional form of the combined leakages is likely to be of higher degree than the individual leakages, as terms are multiplied together. This may have implications for LR-based DPA as a larger covariate set $\mathcal{U}$ will be needed.

\footnote{For key recovery they use as distinguisher the residual sum-of-squares $SS_{\text{resid}}$, selecting the hypothesis that minimises this quantity. It becomes clear that this directly equivalent to choosing the hypothesis that maximises $R^2$ when we consider that $R^2 = 1 - SS_{\text{resid}} / SS_{\text{total}}$ (noting that $SS_{\text{total}}$ is not hypothesis-dependent).}
to account for all the data-dependent variance in the combined leakage. Moreover, the multiplication of trace values amplifies noise, so that more traces will inevitably be required to reach success rates equivalent with those of univariate attacks in unprotected scenarios.

The question of how to extend LR-based DPA to dimensions greater than 2 has not yet been dealt with in the literature, to our knowledge, and could well be an interesting topic for further work. A solution does not immediately follow from the two-target adaptation as the sample covariance does not have a natural analogue in higher dimensions.

7.5 Asymptotic performance evaluation

In this section we examine the asymptotic performance of single- and two-target LR-related DPA attacks against AES, PRESENT and DES S-Boxes in different scenarios. Supposing leakage polynomials of increasing degree, we consider what may be achieved with varying levels of prior knowledge:

- A perfectly characterised power model, which can be exploited straightforwardly in a correlation attack. In the second-order case, we use the known leakage function to compute the functional form of the pre-processed dependent variable, as per [90]. (Corresponds to the line labelled ‘Perfect model’ in Figures 7.2 and 7.3).

- Knowledge of the degree \( d \) of the leakage polynomial (or, for two-target attacks, the polynomial expression for the combined leakage), 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., \( \mathcal{U} = \{u \in \mathbb{F}_2^n \setminus \{0\} | HW(u) \leq d\} \). (Labelled in Figures 7.2 and 7.3 as ‘MaxDeg LR’ and ‘MaxDeg SW’ respectively).

- No knowledge about the leakage polynomial, i.e., \( \mathcal{U} = \mathbb{F}_2^n \setminus \{0\} \), in which case we are interested in the performance of generic LR-based DPA and uninformed (i.e. ‘generic emulating’) stepwise LR-based DPA. (Labelled in Figures 7.2 and 7.3 as ‘Generic LR’ and ‘GenEm SW’ respectively).
Figure 7.2 shows the median nearest-rival margins of the single-target attacks against unprotected leakage. 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 LR does succeed, even against high degree leakage, and when combined with knowledge of the degree it is the best performing of all the LR-based attacks.

The third panel nicely illustrates the robustness of a generic strategy against a non-injective target (namely, the first DES S-Box), and the distinguishability gains available from both information on the degree and the use of stepwise regression. As above, stepwise regression-based DPA combined with knowledge of the leakage degree (when less than maximal) produces the largest distinguishing margins after correlation DPA with a perfect model.

**Figure 7.2:** Median distinguishing margins of single-target attacks against unprotected AES, PRESENT and DES S-Boxes as the leakage degree increases (500 experiments with uniformly random coefficients between -10 and 10).

Figure 7.3 shows the median nearest-rival margins of the two-target attacks against first-order Boolean masking. They are strikingly similar to the margins attained in the unmasked setting. Of course, the efficiency of the distinguishers will be impacted by the amplification of noise in the masked setting, so that we always expect such attacks to require more trace measurements. Nonetheless, it is interesting to note that the stepwise enhancement and prior knowledge maintain similar (relative) beneficial impacts.
Figure 7.3: Median distinguishing margins of two-target attacks against masked AES, PRESENT and DES S-Boxes as the leakage degree increases (500 experiments with uniformly random coefficients between -10 and 10).

The medians displayed in Figures 7.2 and 7.3 provide a reliable indication of the behaviour over the whole experimental sample as the variance is moderate, at least in the case of AES and DES S-Boxes. By way of illustration, Figures 7.4 and 7.5 below show the 1st 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 and the consequent limit on the degree of cryptographic nonlinearity attainable. Note, again, that the percentiles for the two-target attacks are remarkably similar to those for the single-target attacks (without masking)—indicating that the similarity is across the whole distribution of exhibited outcomes.

Of course, all of these attacks use fixed stepwise inclusion/exclusion thresholds. The margins attained may not represent the most effective attacks possible as user-defined settings vary; failure cases, in particular, may respond to more sensitive tuning.

Figure 7.4: First percentile of the distinguishing margins of single-target 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).

Figure 7.1 would suggest 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 $k^*$. 149
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. 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. We are quite able to artificially construct leakage functions which *do* frustrate stepwise regression: for example, random permutations over $\{0, \ldots, 2^m - 1\}$ (it may be relevant that, when interpreted as functions over $F_m^2$, these are the very functions to achieve a high cryptographic nonlinearity—at least when $m$ is large enough). However, we leave as an open question the precise properties of $L$ which will cause stepwise regression to fail.

### 7.6 Conclusion

In this chapter we have explained and developed the theory around LR-based DPA and argued for its use as an efficient but flexible strategy which can be viewed as a generalisation of correlation DPA—and also as an extension of the original difference-of-means approach. We have shown that it belongs to the class of ‘generic compatible’ distinguishers as defined in Chapter 2, but also that such an application imposes unnecessary restrictions on its operation.

By relaxing the generic model assumptions to incorporate some extremely minimal (and reasonable) intuition about the form of the leakage function relative to its composition with a highly-nonlinear target function, we are able to learn more about the secret key.
than that revealed by the distinguishing vector (which is comprised of coefficients of determination). The parameter estimates themselves can be used to guide us towards the ‘most likely’ key hypothesis. A data-mining technique known as ‘stepwise regression’ can be used to automate this learning process, producing simplified key-dependent approximations to the side-channel leakage which are closest under the true key (thereby producing the highest $R^2$) even when the target function is injective and the model begins with a full set of covariates $U = \mathbb{F}_2^m \setminus \{0\}$. We describe such a strategy as ‘generic-emulating’.

We have examined the asymptotic outcomes of LR- and stepwise LR-based attacks against different S-Boxes and confirmed the theoretic success of the latter against injective S-Boxes such as AES and PRESENT and the enhanced distinguishing margins by comparison with standard LR-based DPA against a non-injective S-Box from the DES block cipher. We have also demonstrated the advantage of having some information on the true leakage function (in this case, the polynomial degree). This analysis has also revealed that masking impacts only the efficiency and not the feasibility (or relative effectiveness) of the LR-based distinguishers explored, when taken in combination with a pre-processing step equivalent to that used for two-target correlation DPA against first-order masking.

### 7.A Ordinary least-squares linear regression

Suppose we are interested in modelling the relationship between a random variable $Y$ (the dependent variable) and $p$ random variables $Z_i$, $i = 1, \ldots, p$ (the explanatory variables). An LR model takes the form:

$$Y = \sum_{i=1}^{p} \beta_i Z_i + \varepsilon = Z^T \beta + \varepsilon,$$

where $Z = \{Z_i\}_{i=1}^{p}$ is the $p \times 1$ vector of explanatory variables and $\beta$ is the $p \times 1$ vector of model parameters. *Fitting the model* means estimating $\beta$ from a sample $\{y_j, z_j\}_{j=1}^{N}$.

The random variable $\varepsilon$ is the error term, capturing all factors other than $Z$ which influence $Y$. In ordinary least squares (OLS) estimation $\varepsilon$ is assumed to have mean zero and to
be uncorrelated with Z. Moreover, it is assumed to have a constant conditional variance \( \text{Var}[\varepsilon | Z] = \sigma^2_\varepsilon \) and to be uncorrelated between observations. If the errors are additionally assumed to be normally distributed (conditional on Z) then the OLS estimator is the maximum likelihood estimator and is therefore asymptotically efficient. The normal assumption also helps with inference (and, by extension, sample size computations) but is not needed for the validity of OLS.

The OLS estimator for \( \beta \) is constructed such that the residual sum of squares between the actual data and the fitted model is minimised:

\[
\hat{\beta} = \arg \min_{b \in \mathbb{R}^p} \sum_{j=1}^{N} (y_j - z_j^T b)^2 = \left( \frac{1}{N} \sum_{j=1}^{N} z_j z_j^T \right)^{-1} \cdot \frac{1}{N} \sum_{j=1}^{N} z_j y_j.
\]

Under the above-mentioned assumptions that the error term has zero mean and is uncorrelated with Z, \( \hat{\beta} \) is an unbiased estimator for \( \beta \), and the assumption that the errors have a constant variance and are uncorrelated with one another gives, via the Gauss-Markov theorem, that \( \hat{\beta} \) is the ‘best linear unbiased estimator’, that is, in the class of linear unbiased estimators it attains the smallest sampling variance. (This contributes to its efficiency, which, in the context of DPA, is understood with respect to the number of trace measurements needed for precise estimation).

The coefficient of determination \( R^2 \) is a measure of how well the estimated model approximates the observed data. It is computed as the proportion of variance in the data which is explained by the model fitted to \( \{z_j\}_{j=1}^{N} \):

\[
R^2 = \frac{\sum_{j=1}^{N}(z_j^T \hat{\beta} - \bar{y})^2}{\sum_{j=1}^{N}(y_j - \bar{y})^2},
\]

where \( \bar{y} \) is the sample mean: \( \bar{y} = \frac{1}{N} \sum_{j=1}^{N} y_j \). The higher the \( R^2 \), the better the model is considered to ‘fit’ the data.
7.B Stepwise regression

Stepwise regression is a procedure taking as inputs 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. $I_{fix} \cap I_{initial} = \emptyset$). It returns a set of indices $I_{in}$ indicating the terms selected for inclusion in the final model, which is arrived at as follows:

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 $I_{in} = I_{in} \cup \arg\min_{j \in I_{test}} pval_j$, $I_{test} = I_{test} \setminus \arg\min_{j \in I_{test}} pval_j$ 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 $I_{in} = I_{in} \setminus \arg\max_{i \in I_{test}} pval_i$, $I_{test} = I_{test} \cup \arg\max_{i \in I_{test}} pval_i$ 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 $I_{initial} = I_{fix} = \emptyset$. 

153
Chapter 8

Summary

We are confident that the preceding chapters present a meaningful contribution to the understanding of differential side-channel analysis—in particular, the behaviour and applicability of certain statistics as distinguishers in a variety of leakage scenarios and as supplied with varying degrees of prior knowledge about those scenarios.

Existing side-channel literature gives quite a lot of attention to ‘perfect knowledge’ (i.e. fully-profiled) adversary capabilities [30, 50, 92, 97], as representative of the ‘worst-case’ vulnerability of a device. Meanwhile, the study of ‘no knowledge’ attacker capabilities is a fairly recent development—the mutual information analysis proposed in 2008 by [49] being a prominent early attempt. These so-called ‘generic’ methods are of interest because they indicate in some sense the inherent vulnerability of a cryptosystem independently from the physical implementation details. However, quite what it means for a strategy to be ‘generic’ has only been vaguely specified and varies from study to study. In Chapter 2 we attempted to formalise the notion of genericity and explore the conditions under which it is possible. We confirmed that a truly generic DPA strategy is only possible when the target function is non-injective and sufficiently nonlinear in some appropriate sense.

Previous studies evaluating or comparing side-channel distinguishers have relied on the empirical outcomes of practical attacks against simulated or measured traces. In Chapter 3 we discussed the problems inherent with such an approach, namely that results are...
necessarily estimator-dependent with no means of establishing a ‘best-case’ estimation methodology by which to produce ‘optimal’ outcomes. We proposed, complementarily, to evaluate distinguishers based on the theoretic quantities to be estimated, and presented a range of outcome measures designed to be indicative of the data-complexity of the practical stage of an attack.

DPA methodologies proposed as generic in the informal sense have included a broad class of nonparametric statistics, known to be robust to having little a priori knowledge about the underlying distributions to be compared. However, past works have viewed them as robust to incorrect knowledge—an unjustifiable extension which rather indicates a misunderstanding of their appropriate use. We clarified, in Chapter 4, that even robust statistics require whatever prior knowledge they do assume to be meaningful. Of course, sometimes (i.e. when the target itself is non-injective) it is possible to apply nonparametric statistics with a (formally) generic model, as per Chapter 2. But the adapted strategies suggested for use against injective targets—such as those using the \((m - 1)\)-LSB power model—do not produce analogously effective attacks. In fact, they only succeed when the model inadvertently captures something meaningful about the true leakage after all.

Our theoretic analysis (also found in Chapter 4, and according to the framework of Chapter 3) revealed that the foremost nonparametric distinguishers already circulating in the literature (mutual information [49], Cramér–von Mises criterion [117], Kolmogorov–Smirnov test statistic [117]) perform (broadly) similarly well or badly in various situations and that (theoretic) outcomes are far more heavily influenced by the accuracy of the prior knowledge on the leakage than they are by choice of statistical methodology. Moreover, the variance ratio distinguisher of [106] is similarly able to exploit a generic power model whilst making some (reasonable) assumptions about the noise, indicating that nonparametric methods may often be more flexible than necessary—an important consideration when we recognise the well-documented trade-off between flexibility and efficiency [45] (which, in DPA terms, implies an unnecessary overhead in the number of trace measurements needed).

We similarly (in Chapter 5) analysed nonparametric multi-target strategies, particularly
interesting for use against masked implementations, and promoted in the literature as preferable over multi-target correlation DPA which requires multivariate trace measurements to be pre-processed—that is, converted to a univariate ‘trace’ in an inevitably information-losing mapping. Our analysis confirmed that generic two-target strategies were robustly successful against a wide range of (Boolean) masked leakage scenarios, but that, as with the single-target distinguishers, ‘near-generic’ strategies using the \((m - 1)\)-LSB power models did not provide equivalent robustness. Moreover, incorporating multiple targets in attacks against unprotected implementations diminished, rather than enhanced, distinguishing power.

In Chapter 6 we applied our analysis framework to the particular case of Hamming distance leakage from an unknown reference state, which we showed had a natural correspondence with the type of leakage exhibited by imperfectly-balanced DRP logic. Generic strategies were shown once more to be particularly robust in such contexts, and even maintained reasonable success rates when the unknown reference state was not constant for all inputs.

In Chapter 7 we examined (non-profiled) linear regression-based DPA as suggested in [97] and recently developed in [37]. We showed that it constitutes a generic-compatible strategy (with the usual limited scope for application), but can be adapted via the introduction of some non-device-specific information about the leakage so as to produce theoretically successful key recovery in a much broader range of scenarios. The existence of such ‘generic-emulating’ strategies—succeeding where generic strategies fail—is unprecedented in the literature; suggested work-arounds for targeting injective functions focused on \((m - 1)\)-LSB power models which are actually, we have shown, highly sensitive to the actual leakage, only succeeding if they in fact chance to capture something sufficiently meaningful about the true functional form.

8.1 Future directions

We identify several avenues for further work indicated by the findings of this thesis:
1. The approach that we have taken to focus on theoretic (rather than experimental) performance has offered valuable insight about non-profiled DPA. There would seem to be many potential applications for similar approaches to questions about profiled DPA. In particular, it would be useful to be able to formally define what is meant by a ‘template’ and to explore appropriate ways of evaluating different methods of template building from a theoretic perspective.

2. The discovery of one ‘generic-emulating’ distinguisher raises the question of whether or not other strategies can be found—using potentially quite different mechanisms—which are able to exploit similarly minimal non-device-specific intuition.

3. The focus in Chapter 7 is on the feasibility of linear regression-based strategies and of the generic-emulating stepwise variant. It would be enlightening to follow this up with theoretical and experimental investigation into the efficiency (i.e., the data complexity).

4. The linear regression-based distinguishers have been extended (by us and others [36]) to multi-target attacks via pre-processing, similar to the ways in which correlation-based DPA have been extended. However, intuitively it seems that the flexibility of such strategies may offer opportunity to incorporate multiple target points without the need to pre-process; this should be explored.

8.2 Publications

8.2.1 Relating to this thesis

Much of the work in this thesis has been published or is currently in submission, as follows:

A Comprehensive Evaluation of Mutual Information Analysis Using a Fair Evaluation Framework
Carolyn Whitnall and Elisabeth Oswald
In: Advances in Cryptology, CRYPTO 2011, LNCS 6841, pages 316–334. Springer, Au-
Abstract: The resistance of cryptographic implementations to side-channel analysis is a matter of considerable interest to those concerned with information security. It is particularly desirable to identify the attack methodology (e.g. differential power analysis using correlation or distance-of-means as the distinguisher) able to produce the best results. Such attempts are complicated by the many and varied factors contributing to attack success: the device power consumption characteristics, an attacker’s power model, the distinguisher by which measurements and model predictions are compared, the quality of the estimations, and so on. Previous work has delivered partial answers for certain restricted scenarios. In this paper we assess the effectiveness of mutual information-based differential power analysis within a generic and comprehensive evaluation framework. Complementary to existing work, we present several notions/characterisations of attack success with direct implications for the amount of data required. We are thus able to identify scenarios in which mutual information offers performance advantages over other distinguishers. Furthermore we observe an interesting feature—unique to the mutual information based distinguisher—resembling a type of stochastic resonance, which could potentially enhance the effectiveness of such attacks over other methods in certain noisy scenarios.

This paper introduces the evaluation framework described in Chapter 3 and corresponds strongly with the material in Chapter 6, although also touching briefly on some of the theory behind mutual information-based DPA as presented in Chapter 4.

A Fair Evaluation Framework for Comparing Side-Channel Distinguishers

Carolyn Whitnall and Elisabeth Oswald


Abstract: The ability to make meaningful comparisons between side-channel distinguishers is important both to attackers seeking an optimal strategy and to designers wishing
to secure a device against the strongest possible threat. The usual experimental approach requires the distinguishing vectors to be estimated: outcomes do not fully represent the inherent theoretic capabilities of distinguishers and do not provide a basis for conclusive, like-for-like comparisons. This is particularly problematic in the case of mutual information-based side channel analysis (MIA) which is notoriously sensitive to the choice of estimator. We propose an evaluation framework which captures those theoretic characteristics of attack distinguishers having the strongest bearing on an attacker’s general ability to estimate with practical success, thus enabling like-for-like comparisons between different distinguishers in various leakage scenarios. We apply our framework to an evaluation of MIA relative to its rather more well-established correlation-based predecessor and a proposed variant inspired by the Kolmogorov–Smirnov distance. Our analysis makes sense of the rift between the \textit{a priori} reasoning in favour of MIA and the disappointing empirical findings of previous comparative studies and moreover reveals several unprecedented features of the attack distinguishers in terms of their sensitivity to noise. It also explores—to our knowledge, for the first time—theoretic properties of near-generic power models previously proposed (and \textit{experimentally verified}) for use in attacks targeting injective functions.

This paper reviews the framework from Chapter 3 and covers much of the material in Chapter 4 and 5.

\textbf{An Exploration of the Kolmogorov–Smirnov Test as a Competitor to Mutual Information Analysis}

\textit{Carolyn Whitnall, Elisabeth Oswald and Luke Mather}


\textit{Abstract:} A theme of recent side-channel research has been the quest for distinguishers which remain effective even when few assumptions can be made about the underlying distribution of the measured leakage traces. The Kolmogorov–Smirnov (KS) test is a well known non-parametric method for distinguishing between distributions, and, as such, a perfect candidate and an interesting competitor to the (already much discussed) mutual
information (MI) based attacks. However, the side-channel distinguisher based on the KS test statistic has received only cursory evaluation so far, which is the gap we narrow here. This contribution explores the effectiveness and efficiency of Kolmogorov–Smirnov analysis (KSA), and compares it with mutual information analysis (MIA) in a number of relevant scenarios ranging from optimistic first-order DPA to multivariate settings. We show that KSA shares certain ‘generic’ capabilities in common with MIA whilst being more robust to noise than MIA in univariate settings. This has the practical implication that designers should consider results of KSA to determine the resilience of their designs against univariate power analysis attacks.

This paper explores Kolmogorov–Smirnov analysis from a theoretic and a practical perspective. There is overlap with Chapter 4 and brief reference to the framework of Chapter 3, but also work not covered in this thesis.

The Myth of Generic DPA... and the Magic of Learning

Carolyn Whitnall, Elisabeth Oswald and François-Xavier Standaert

In submission. (A preliminary version is available from the Cryptology ePrint Archive: Report 2012/256).

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.

The delineation of model types and definition of ‘generic DPA’ in this paper overlaps with Chapter 2 whilst the exploration of first-order linear regression-based DPA overlaps with material from Chapter 7.

8.2.2 Other publications

In addition to research output relating to this thesis I have contributed to the following paper, also in submission:

**Faster, Higher, Stronger: Practical Second and Third Order DPA Attacks**  
*Michael Tunstall, Carolyn Whitnall and Elisabeth Oswald*

In preparation at the time of submitting this dissertation.

*Abstract:* The side-channel literature describes numerous masking schemes designed to protect block ciphers at the implementation level. These schemes typically require an initialisation step during which random values are generated and/or randomised tables are written to memory. In this paper we revisit mask-biasing strategies and present a new attack which directly exploits the initialisation operation in such a way as to recover all or some of the masks as a pre-processing step. This pre-processing enables highly efficient key recovery via standard first-order DPA on (fully or partially) unmasked target values. We provide a theoretic analysis of the success and efficiency of both strategies as applied to affine and second-order Boolean masking schemes, and experimentally confirm our expectations through practical attacks against typical implementations.
Chapter 9

Bibliography


168


