Leakage Certification Made Simple

Aakash Chowdhury[1], Arnab Roy[1], Carlo Brunetta[2], and Elisabeth Oswald[1,3]

[1] University of Klagenfurt, Austria {arnab.roy, aakash.chowdhury, elisabeth.oswald}@aau.at
[2] Simula UiB, Norway carlob@simula.no
[3] University of Birmingham, UK

Abstract. Side channel evaluations benefit from sound characterisations of adversarial leakage models, which are the determining factor for attack success. Two questions are of interest: can we estimate a quantity that captures the ideal adversary (who knows the distributions that are involved in an attack), and can we judge how good one (or several) given leakage models are in relation to the ideal adversary?

Existing work has lead to a proliferation of custom quantities (the hypothetical information HI, perceived informatino PI, training information TI, and learnable information LI). These quantities all provide only (loose) bounds for the ideal adversary, they are slow to estimate, convergence guarantees are only for discrete distributions, and they have bias.

Our work shows that none of these quantities is necessary: it is possible to characterise the ideal adversary precisely via the mutual information between the device inputs and the observed side channel traces. We achieve this result by a careful characterisation of the distributions in play. We also put forward a mutual information based approach to leakage certification, with a consistent estimator, and demonstrate via a range of case studies that our approach is simpler, faster, and correct.

Keywords: Side channels, Evaluation, Leakage Certification, Mutual Information Estimation

1 Introduction

The mutual information enables to quantify the amount of information that we obtain about one random variable by observing another random variable. This is a useful concept in the context of side channels, because it enables us to quantify how much information we get about a secret (key-dependent) device state by observing e.g. the device power consumpion. As a consequence, the mutual information appears across various areas in side channel research, such as in proofs about the security of masking (e.g. [1]), in the context of side channel distinguishers (e.g. [2]), and in the context of reasoning about the quality of so called device leakage models (e.g. [3]) — the latter application is the focus of our work.

1.1 Evaluating device security via leakage certification

Device leakage models are important ingredients in side channel attacks. Side channel attacks are highly configurable, but they always require the extraction of information of small portions of the secret key from some observed side channel traces (they follow a divide-and-conquer principle). The extraction of key information from the observable side channel traces can be achieved with a wide range of statistical and machine learning tools, which use as inputs a (key-dependent) leakage model and the observed side channel traces. It is well known that the use of an accurate leakage model is necessary for optimal information extraction [4].

The best leakage model (from an adversary's point of view) would evidently be equal to the distribution of the side channel that the device emits. We call an adversary *ideal* if they know this distribution and therefore have the best model. In order to understand the worst case security of a device, an evaluator wishes to assess the strength of this ideal adversary.

In the context of physical side channels such as the power consumption, the EM emanation, or device timing characteristics, the exact distribution of the observable side channel is unknown—both adversaries and evaluators can only work with estimations.

State of the art. An evaluator thus seeks to understand how good their leakage model is, which is a task that was described by Durvaux et al. [3] as *leakage certification*, drawing on the earlier work of Renauld et al [5]. In a series of follow on works [6,7,8] the initial approach was further refined. To be more precise, Durvaux et al [6] consider both aspects of leakage certication: reasoning about the ideal adversary and comparing leakage models. The latest two papers [7,8] only focus on reasoning about the ideal adversary. The main challenge, as perceived in these works is to provide information bounds, because the true distributions are unknown. Their idea works as follows.

1. The evaluator estimates a quantity called the *perceived information* (via the ePI or gPI, or LI), which measures a relationship between the device leakage model and the actual device leakage. The quantity can be estimated by sampling from the model and from the real device. It is supposed to lower bound the ideal adversary.
2. The evaluator estimates a quantity called the *hypothetical information* (via the eHI, or TI), which is defined as the mutual information between the device leakage model and the key-dependent state. If the model is defined to be the empirical distribution, then, assuming enough samples are available, this quantity will converge to the mutual information between the key-dependent state and the observed leakage (under a range of assumptions). It is supposed to upper bound the ideal adversary
3. The evaluator uses the resp. estimated quantities to get a lower and an upper bound for the ideal adversary. To test the quality of one leakage model, or compare leakage models, the resp. quantities are estimated using the given model(s).

In [7,8] proofs are provided for convergence bounds, assuming univariate discrete side channel observations.

Problem 1. Physical side channels are typically neither discrete nor univariate. The argument that side channels such as power and EM are measured by *digital* oscilloscopes (i.e. devices that use an analogue to digital converter) misses two points. Firstly, modern digital oscilloscopes offer sophisticated signal amplification and de-noising settings which produce real-valued outputs: assuming that devices are only used in their most basic setting underestimates real-world adveraries. Secondly, implementations that implement masking countermeasures are often analysed after further software processing, including filtering, and mean-free product-combining [9], which again create real-valued outputs.

Under the assumption of discrete and uniform side channel traces, there is in fact a consistent MI estimator availabe in the literature, see e.g. [10]. No workaround for the estimation problem is necessary. *Only if one variable is assumed to be continous, or a mixture distribution, the estimation of MI becomes an interesting problem.* This problem was however also solved in a series of works culminating in a consistent estimator for MI that can work on any type of input data (discrete, continuous, and mixtures), by Gao et al. [11].

Last the disconnect between the experiments presented in these works (based on continuous side channel traces, clearly visible in the code that the authors helpfully supply) and the theory (assuming discrete side channel traces) is not analysed. Both [7,8] acknowledge that there is an issue because of discretisation via referencing the paper by Paninski [12], but they don't discuss what the impact concretely is.

Problem 2. The core idea of the existing work is to estimate the HI and the PI and to use them as bounds to reason about the ideal adversary . However, already Durvaux et al. [6] noticed that the PI is undefined if models are bad approximations. Recently Masure et al. [8] show that the gHI (which is one particular way of estimating the HI) is not guaranteed to be an upper bound for the PI (when estimated via the gPI), and that the eHI suffers from serious bias especially in multivariate settings. We remark at this point, that the non-parametric estimators that were put forward by the authors become computationally infeasible as the number of dimension increases.

New quantities (TI and LI) were introduced to remedy efficiency problems, but they come with a number of assumptions as well, in particular we still need discrete data and they require parametric estimation. Summarising, the literature shows that the HI, PI, LI, TI are limited, biased, slow to estimate, and they offer loose bounds for the quantity of interest.

1.2 Gaps which our Contributions Seek to Close

The existing work in the side channel community in the context of leakage certification has failed to account for the fact that in many situations evaluators are confronted with real valued side channel traces, it has misunderstood some of the results about mutual information estimation, and it has missed some of the progress that has been made regarding mutual information estimation. Our work corrects some misunderstandings, and puts leakage certification on a modern basis.

Contributions. After a brief introduction of some notation and a review of the side channel setting in Sect. 2, we review the state of the art of estimating mutual information in Sect. 3. Our review recaps on the salient characteristics of estimators, it introduces the HI and eHI from previous work, and the relatively recent estimator by Gao et al. [11] that we will use in the context of our approach to leakage certification.

In Sect. 4 we explain the PI and recap on the notion of regret that was recently put forward by Masure et al. [8] and which formalises the idea in the previous works [6,7]. We use this to formalise how to compare two given leakage models, and then we show that the PI can be understood as quantifying an average information loss (via the Kullback-Leibler divergence). We interprete this result as evidence that leakage certification can be made much simpler by directly quantifying model information via the mutual information. We then turn our attentention to the HI and show that discretisation implies that it is not necessarily an upper bound for the ideal adversary.

In Sect. 5 we introduce our new simple method of leakage certification, which is based on the idea that we can estimate the mutual information that characterises the ideal adversary via estimating the mutual information between the device inputs and the side channel traces. We show in Sect. 6 when this is possible: we can do this for all leakage functions, models and noise distributions that have been used in the side channel literature.

We move towards the practical aspects of our proposal in Sect. 7 where we discuss our fast implementation of the estimator by Gao, and confirm experimentally our theoretical result on discretisation. Finally we look at a range of case studies in which we challenge the bounds provided by estimating the HI and PI vs. directly esimating the ideal adversary. We find that in multivariate settings the results from the HI/PI estimators are poor, and when comparing leakage models, they can even be misleading. Our method remains efficient and correct in multivariate scenarios and it is able to correctly assess different leakage models.

Our main results contribute a novel approach for leakage certification which:

- uses a strongly consistent (thus unbiased) efficient estimator for all types of side channel traces that can be observed in practice (discrete, continuous, mixtures, and even probabilistic functions),
- naturally extends to the multivariate setting,
- enables to characterise the ideal adversary, and
- enables to compare arbitrary device leakage models.

2 Preliminaries

We aim to keep this section as brief as possible, and offer deeper explanations only for those concepts that our results are based on.

2.1 Notation

Following convention, we represent random variables with upper case letters, and their realisations with the corresponding lower case letters and sets are denoted with calligraphic typefaces. For two functions g and h, $g \circ h$ denotes the composition of the functions.

We denote the probability density function (pdf) and cumulative distribution function (cdf) of a continuous random variable with f and F respectively. For a discrete random variable, p will denote its probability mass function (pmf). Whenever necessary, in a pdf, cdf or pmf we will make the corresponding random variable explicit in the subscript (e.g. f_X or F_X). In particular $p_{(X,Y)}$ refers to the joint distribution (pmf in this case) of the variables X and Y.

For any random variable X, $\mathbb{E}(X)$ and resp. $\mathbb{E}_X$ denote the expectation. For simplicity we denote the conditional expectation of random variable $X|Y = y$ by $\mathbb{E}_{X|y}$.

We refer to an estimated quantity by using the sample size n in the subscript, e.g. I_n refers to a mutual information estimate obtained from a sample with size n, $f_{X,n}$ or $p_{X,n}$ denote the estimated pdf or pmf corresponding to a random variable X using n samples.

The indicator function for a realisation x of X, is denoted as $\mathbb{I}_{X=x}$. We use $\mathcal{N}(\mu, \sigma)$ to denote the Gaussian/normal distribution with mean μ and standard deviation σ. We use $\mathcal{L}(0, \sigma)$ to denote a Laplacian distribution. We use R to denote the random variable corresponding to the device noise. For any d-dimensional vector $(x_1, \ldots, x_d) \in \mathbb{R}^d$ the ℓ_∞ or max norm is defined as $\max\{|x_i| : i = 1, \ldots, d\}$. Discretised distributions are denoted by putting brackets around them, e.g. $[X]$. For a real valued variable x, $[x]$ denotes the discretised value.

When working with functions we overload notation, and use the same variable for both the function, as well as the result of the function, and we may adapt the inputs to the context, e.g. $L(X, K)$ is a function, we also understood L as a random variable, i.e. l is the realisation of L with some concrete inputs x, k.

2.2 The side channel setting

In the side channel setting we work with random variables that represent inputs/intermediates/outputs of cryptographic processes and leakage observations: we use $x \in \mathcal{X}$ for the input, which is mapped according to the cryptographic process via the application of some (cryptographic) target function(s) C and an (unknown) key k^* $\in \mathcal{K}$ to an intermediate $y \in \mathcal{Y}$. Implementations process cryptographic keys in "chunks", thus K and X have small support. The intermediate value is then mapped via a (noisy) device leakage function (we discuss properties of them subsequently) to the observable side channel trace t.

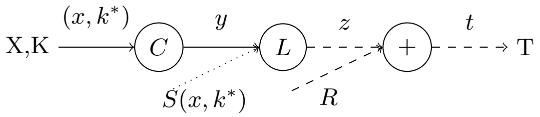


Fig. 1: Relationships between variables

A side channel trace t is a vector of leakage points. Each point corresponds to the physical processes that happen inside the device. Some of the physical processes depend on the input and key and we capture their contribution to the observable traces with the leakage function L. Other processes are independent of the input and the key and we capture them via the independent noise variable R.

An evaluator is assumed to be able to observe (and even control) inputs/outputs $x \in \mathcal{X}$ and the key k^* of the device. The also know the function(s) C, and they can observe the side channel leakage traces $t \in \mathcal{T}$. An adversary can typically either observe or control x, they know C and they observe t. Like an evaluator, the adversary does not know L and thus uses a so-called leakage model M for extracting key information.

Leakage models. A leakage model is a function M that maps x, k under a target function C to $\mathbb{R}^d$. A model can be assumed based on device knowledge, or it can be estimated from real trace data. For example, a very popular standard leakage model is the Hamming Weight function, i.e. $M(x, k, C) = HW(C(x, k))$. Non-parametric estimated models are often derived by building histograms for either pairs (x, k), or by building histograms for the target function, i.e. $C(x, k)$. Models can be univariate, i.e. $d = 1$, but they can also be multivariate. In either case, the evaluator isolates some “points of interest” in each trace and uses these points for model building.

A hidden, but important, point for the quantities HI and PI (and all their estimators, we will introduce them in the coming sections) is that the model M and the leakage L should be defined over the same space. This is because instead of working with the unknown joint distribution (Y, T), these quantities work with the known distribution (Y, M) but using observed traces. Intuitively, this point should hold for a good model, but clearly it might not hold for a model that is a poor approximation.

Leakage functions. An important detail is that the leakage function L may be either a deterministic or a probabilistic function of multiple variables. A deterministic function is fully determined by its’ inputs. A probabilistic function includes an element of chance.

The leakage function L for a specific step in the execution of an algorithm can be simple. For instance, it can be determined by the number of bits changing within a register, or on a bus, in the case of a memory instruction, in which case it can be understood as a deterministic function. In other words, for a given input x and a fixed key k^*, it will alway produce the same value $L(x, k^*)$, and the distribution of $L(x, k^*)$ is completely determined by the current state (x, k^*).

But the leakage function for a specific step in the execution of an algorithm can also be complex. For instance, in a dedicated hardware implementation of a non-linear function, the power consumption depends on a complex interaction between many gates, which can result in data dependent glitches, cross-talk, etc. In this case, for a given input x and a fixed key k^*, it can produce different values upon repeat execution. The distribution of L thus depends on x, k^*, and some unknown randomness $S(x, k^*)$ that depends on x, k^*. We provide Fig. 1 as a visual aid to understand the relationship between the variables, based on the functions that act on them (the dashed lines indicate the random processes and variables, the dotted line visualises that L might depend on some input and key dependent randomness S).

We wish to emphasise the need to capture *all* types of leakage functions in the context of leakage certification because an evaluator does not know the leakage function(s) that a device exhibits and thus needs a methodology that always returns correct results.

In the rest of this paper, T *should always be understood as continuous variable (or a mixture with a continuous component).* Whenever estimators require discrete inputs, we make this explicit by writing $[T]$ to indicate that discretisation of T must take place. Whenever the probabilistic nature of the leakage is not relevant, i.e. a statement holds irrespective of S and thus irrespective of whether L is deterministic or probabilistic, we drop S in the text for readability.

3 Estimating Mutual Information—State of the Art

The mutual information (MI) quantifies what we can learn about a variable X upon observing another variable Y. In the context of evaluating side channel security, it is clear that we can use the MI to quantify how much we can learn about a secret (key-dependent) device state upon observing the device's side channel.

In the context of leakage certification, we will (in the next section) also use it to reason about the quality of a device leakage model. Intuitively, a device leakage model is "better" if it is "closer" to the real device leakage.

In the following, we first recap how the mutual informatino can be defined, and then how the mutual information can be estimated.

3.1 Defining Mutual Information

For general random variables X, Y (with marginal distributions P_X, P_Y and joint distribution P_{XY}), the mutual information (MI) is defined via the Radon-Nikodym derivative [13]:

$$I(X;Y) = \int_{\mathcal{X}\times\mathcal{Y}} \log \frac{dP_{XY}}{dP_X P_Y} dP_{XY}.$$

The definition via the Radon-Nikodym derivative links the mutual information also with the *Kullback-Leibler divergence* D_{KL} as it can be expressed as $I(X;Y) = D_{KL}(dP_{XY}||dP_X dP_Y)$ (see [13]).

If either both variables are discrete, or both variables are continuous, then the MI can be expressed via the marginal and joint or conditional entropies[4], leading to the well known "2H" and "3H" expressions (owing to how many entropies are in the formulae) for MI, see Equ. (1).

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) && (1) \\ &= H(X) + H(Y) - H(X,Y) && (2) \end{aligned}$$

If one variable is discrete and one is continuous, or if one variable is a mixture, then the conditional density in the 2H formula, and the joint density in the 3H formula, may not be well defined unless the involved distributions satisfy specific conditions, see [14][5]. Consequently, in situations where the distributions are unkown, and thus one cannot verify that the conditional/joint entropies are well defined, the conservative choice is to utilise an estimator that estimates the mutual information via the Radon-Nikodym derivative.

3.2 MI Estimation

The crucial property of any MI estimator is how well it "approximates" the true MI. This property is called the convergence of the estimator, and it describes the behaviour of the estimator when we supply it with increasing amounts of data. There exist different notions of convergence. The weakest notion is convergence in probability, and estimators that have this property can be biased. A stronger notion is convergence in mean, which implies asymptotic unbiasedness. Bias in an estimator refers to the possibility that the estimator's expected value remains different from the true quantity being estimated. Bias is an undesireable property, although if the bias can be described, it can typically also be corrected for. The rate at which an unbiased estimator convergences is of practical interest as well.

Statistical estimators often benefit from assumptions about the distribution of the quantity that they are estimating. If such assumptions can be justified and they are incorporated in the estimator design, then we call such estimators "parametric". In the case of leakage certification we do not wish to make any assumptions and thus we are interested in *non-parametric* estimators.

[4] We remind the reader that $H(X) = \mathbb{E}_X[-\log f(X)]$ if X continuous, and $H(X) = \mathbb{E}_X[-\log p(X)]$, if X is discrete; the definitions are extended in the natural way for conditional and joint distributions

[5] Observe that in such cases, we have a term that corresponds to a discrete entropy which is always positive, and a term that corresponds to a differential entropy which can be negative. Furthermore the conditional distribution in the 2H formulae might not exist.

3.3 Non-parametric MI estimation

We first provide a conceptual overview of the existing estimation techiques, starting from the oldest techniques and leading up to the most recent advancements.

There are different approaches to estimating the MI (non-parametrically). One can either estimate the entropies in the 2H/3H formulas, or one can estimate the Radon-Nikodym derivative.

Entropy based MI estimation. In the context of MI estimation based on the 2H/3H formulae, there exist two fundamentally different families of (non-parametric) entropy estimators: one family is based on direct density estimators and the other family is based on k-Nearest Neighbour (k-NN) estimators. Density based estimators directly estimate the densities in the 2H/3H formulae, whereas the k-NN based estimators estimate the distribution of the k-nn distance as a proxy for the density itself [15]. The before mentioned limitations of 2H/3H estimators (i.e. both variables must either be discrete or continuous) initially applied to both approaches. However k-NN estimators were further developed and, in a series of works starting with [16], approaches were developed that aim to estimate the MI directly via estimating the Radon-Nidoym derivative (thus without estimating entropies, but still requiring that the variables have a global joint density).

A complementary approach based on using deep learning was published by [17] in 2018. It was suggested to be used for side channel tasks in [18]. However, it was shown later in [19] that the claimed convergence results were erroneous.

In the side channel literature, based on the simplifying assumption of having discrete traces, the study in [20] use an integral estimate [21]. Györfi and van Meulen [22] showed that the integral estimator of entropy (with histogram density estimate) is strongly consistent only if the (conditional) distribution satisfies specific conditions. Hall and Morton [23] (again under certain conditions regarding the distribution) showed that a histogram-based estimator provides mean-square convergence when the dimension of X is 1 or 2. The family of integral estimators does not generalise to the multivariate setting (either their efficiency drops significantly or the convergence guarantee does not extend to the multivariate setting). In the purely discrete setting, the plug-in estimator produces the best results in terms of convergence as proven in [10]. This convergence result was not known in the side channel literature, and instead the eHI was developed as a means to bound the MI. Bronchain et al. [7] put forward the notions of the HI and eHI, see equations (3) and (4), where we make the use of a model M that approxmites T explicit for the sake of clarity.

$$\mathrm{HI}(X;[T];[M]) = \mathrm{H}(X) + \sum_{x\in\mathcal{X}} p_X(x) \sum_{t\in[\mathcal{T}]} p_{(X,[M])}(t|x) \log_2 p_{(X,[M])}(x|t) \quad (3)$$

$$\mathrm{eHI}_n(X;[T]) = \mathrm{H}(X) + \sum_{x\in\mathcal{X}} p_X(x) \sum_{t\in[\mathcal{T}]} \tilde{e}_n(t|[x]) \log_2 \tilde{e}_n(x|[t]) \quad (4)$$

The HI defines a quantity that measures the relationship between a variable X (which in [7] is set to be either the key variable, K or the intermediate $Y = C(X,K)$), and the observed (discrete or discretised) traces $[T]$ under a given model density $[M]$. It is easy to see that $\mathrm{HI}(X;[T];[M]) = I(X;[M])$. And if $[M] = [T]$ the HI can be used to derive the MI. The empirical HI (eHI) uses the empirical distribution $\tilde{e}_n(x,[t])$, which can be estimated from the observed traces $[T]$. Bronchain et al. [7] show that with some assumptions the eHI converges in probability to the MI, which is a result that was already proven for the comparable plug-in estimator by Antos and Kontoyiannis [10] in 2001.

Nearest Neighbour Estimator for MI. Motivated by the need for a non-parametric MI estimator that applies even to high-dimensional/multivariate problems, Krasov et al. [16] introduced the idea of using a k nearest neighbour (short k-NN) based estimator (also known as KSG estimator in the wider statistical literature).

A recent contribution by Gao et al. [11] made a further significant step by estimating the Radon-Nikodym derivative[6] whilst requiring only **local** joint densities: in other words, their estimator does no longer require the existance of a joint density for the entire probability space. Their estimator essentially deals with two cases that can occur for the joint distribution: either the sample (x,y) is discrete (this can be detected by checking the k-nn distance), then one can use the plug-in estimator for the Radon-Nikodym derivative; or the sample (x,y) is locally continuous, in which case they estimate the Radon-Nikodym derivative based on (5). They furthermore show that if either x or y are mixed, then the continuous case applies. Consequently, their estimator can deal with any form of mixtures. The GKOV estimator is defined as given in Equation (5).

$$I_n(X;Y) = \frac{1}{n}\sum_{i=1}^{n}\hat{I}_i = \log n + \frac{1}{n}\sum_{i=1}^{n}(\psi(\tilde{t}_i) - \log(n_{x,i}+1) - \log(n_{y,i}+1)) \quad (5)$$

Here, $\psi(u)$ is the digamma function $\psi(u) = \frac{d}{du}\log \mathit{\Gamma}(u) \approx \log u - \frac{1}{2u}$. The details of how to compute the quantities $n_{x,i}, n_{y,i}$ and $\tilde{t}_i$ can be found in Algorithm 1.

With a suitable choice of the function k_n the GKOV estimator has the same convergence rate as existing pmf/pdf based mutual information estimators, it provides strong convergence (covergence in mean, asymptotic unbiasedness) in all settings, and it can be generalised to multivariate variables.

4 Leakage Certification using HI and MI

We now return to the task of leakage certification. The idea of leakage certification is to assess the quality of a leakage model or multiple models without needing to conduct a full key recovery attack. Not needing to perform a full

[6] This estimator thus is not based on estimating the joint or conditional density.

Algorithm 1 Non-parametric $I(X;Y)$ estimation for mixed r.v.s (X,Y)[11]

```
Require: $\{x_i, y_i\}_{i=1}^n$ and $t_n = t$
for $i = 1, \dots, n$ do
    $d_{i,xy} = t$th smallest distance from$\{d_{ij} = \max\{\|x_j - x_i\|, \|y_j - y_i\|\} : i \neq j\}$
    if $d_{i,xy} = 0$ then
        $\tilde{t}_i = |\{j : t_{ij} = 0\}|$
    else
        $\tilde{t}_i = t$
    end if
    $n_{x,i} = |\{j : \|x_j - x_i\| \leq d_{i,xy}\}|$
    $n_{y,i} = |\{j : \|y_j - y_i\| \leq d_{i,xy}\}|$
    $\alpha_i = \psi(\tilde{t}_i) - \log(n_{x,i} + 1) - \log(n_{y,i} + 1)$
end for
return $\frac{1}{n}\sum_i \alpha_i + \log(n)$
```

key recovery is advantageous, because a full attack takes time (and is therefore costly), or it may even be infeasible in the time that evaluators have.

In this section we first explain the state of the art, which is based on (estimating) quantities such as the HI, PI, (most recently TI, and LI), and the problems that arise when working with these quantities.

4.1 Assessing Model Quality using the HI and PI

In a series of works starting with Renauld et al. [5] the *perceived information* PI (6) was put forward as a measure for the quality of a leakage model. In [7] they propose two types of estimators of PI, PI_n and ePI_n (we denote it later as ePI), which we provide in the equations (7) and(8), respectively.

$$\mathrm{PI}(Y;[T];[M]) = H(Y) + \sum_{y\in\mathcal{Y}} p_Y(y) \cdot \sum_{t\in[\mathcal{T}]} p_{(Y,[T])}(t|y) \log_2 p_{(Y,[M])}(y|t) \tag{6}$$

$$\mathrm{PI}_n(Y;[T]) = H(Y) + \sum_{y\in\mathcal{Y}} p_Y(y) \cdot \sum_{t\in[\mathcal{T}]} p_{(Y,[T])}(t|y) \log_2 \tilde{e}_n(y|[t]) \tag{7}$$

$$\mathrm{ePI}_n(Y;[T]) = H(Y) + \sum_{y\in\mathcal{Y}} p_Y(y) \cdot \sum_{i=1}^{n_t(y)} \frac{1}{n_t(y)} \log_2 \tilde{e}_n(y|[t] = i) \tag{8}$$

In the definition of the ePI, the variable $n_t(y)$ is the cardinality of the set $\{t \in [\mathcal{T}] : Y = y\}$, and $\tilde{e}$ refers to an estimated empirical model.

The idea in [6] is to compare the (estimated) PI with the MI, which gets approximated by the HI. Durvaux et al. then state that if HI and PI are close (for a given leakage distribution, intermediate, and model), then the model may be a good representation of the unknown device leakage. They point out problems with just basing an analysis on the difference between the MI and the PI, and develop a moment-based characterisation. It was explained in [24] that this moment based analysis is not always a suitable workaround, and in the follow on work [7] the focus was on comparing the MI (via the eHI) and ePI. However, we found a significant gap between the theroetical and experimental properties that was addressed in [7]. In [7, Theorem 6] it was proven that PI_n is a lower bound of the MI, but this does not imply that the ePI is a lower bound too (or that it even converges to the PI). In their experimental results on multidimensional traces the lower bound is indeed not attained by the ePI, and the assumption is introduced that $T|Y = y$ follows a Gaussian distribution; hence a parametric estimator is now assumed. Hence the bounds provided by the HI and PI, which are used to bound the ideal adversary, are not reliable or require parametric assumptions.

In the latest work, the process of comparing MI and PI, has been formalised via the concept of the *Regret* for a model M in [8, Definition 4]:

$$\mathsf{R}(M) = I(Y;T) - \mathrm{PI}(Y;T;M). \tag{9}$$

With the regret, we provide a natural extension to compare two leakage models in the subsequent definition.

Definition 1. *MI/PI-based model quality. Given two (discrete) leakage models $[M]_1$ and $[M]_2$, we say that $[M]_1$ is a better leakage model than $[M]_2$ for a (discrete) trace distribution $[T]$ if*

$$\mathsf{R}(M_1) < \mathsf{R}(M_2).$$

Definition 1 must be implemented using estimators for the MI and the PI in practice when the distributions of traces and models are unknown.

Definition 2. *eHI/PI-based model quality. Given two (discrete) leakage models $[M]_1$ and $[M]_2$, a suitably large discrete trace set $[T]$, and the estimators eHI and ePI, then we say that $[M]_1$ is a better leakage model than $[M]_2$ if*

$$\mathrm{eHI}_n(Y;[T]) - \mathrm{ePI}_n([M]_1;[T]) < \mathrm{eHI}_n(Y;[T]) - \mathrm{ePI}_n([M]_2;[T]).$$

4.2 The PI judges model strength via the KL divergence

The PI defines a quantity that is supposed to capture the information about the device state when utilising side channel observations and interpreting them via some leakage model. In the definition of the PI, the joint distributions of $(Y, [T])$ and $(Y, [M])$ are being used in such a way that one cannot deal with arbitrary models. Already Durvaux et al. [6] and then also Bronchain et al.[7] give examples where the PI leads to problematic results. Clearly a better understanding of the PI quantity is needed and we now develop an alternative representation of the PI.

Remark 1. The PI between the three variables $Y, [M], [T]$, and all distributions defined for all $y \in \mathcal{Y}$, can be written as:

$$PI(Y;[T];[M]) = I(Y;[T]) - \mathbb{E}[D_{KL}((Y|t,T)||(Y|t,M))].$$

Proof. We will use the simple substitution $p_{Y,[M]}(y|t) = p_{Y,[M]}(y|t)\frac{p_{Y,[T]}(y|t)}{p_{Y,[T]}(y|t)}$, and find that

$$
\begin{aligned}
PI(Y;[T];[M]) &= H(Y) + \sum_{y\in\mathcal{Y}} p_Y(y) \cdot \sum_{t\in[\mathcal{T}]} p_{(Y,[T])}(t|y) \log_2 p_{Y,[M]}(y|t)\frac{p_{Y,[T]}(y|t)}{p_{Y,[T]}(y|t)} \\
&= H(Y) + \sum_{y\in\mathcal{Y}} p_Y(y) \cdot \sum_{t\in[\mathcal{T}]} p_{(Y,[T])}(t|y) \left(\log_2 p_{(Y,[T])}(y|t) + \log_2 \frac{p_{Y,[M]}(y|t)}{p_{Y,[T]}(y|t)}\right) \\
&= H(Y) + \sum_{y\in\mathcal{Y}} p_Y(y) \cdot \sum_{t\in[\mathcal{T}]} p_{(Y,[T])}(t|y) \log_2 p_{(Y,[T])}(y|t) \\
&\quad + \sum_{t\in[\mathcal{T}]} p_{[T]}(t) \cdot \sum_{y\in\mathcal{Y}} p_{(Y,[T])}(y|t) \log_2 \frac{p_{Y,[M]}(y|t)}{p_{Y,[T]}(y|t)} \\
&= I(Y;[T]) - \mathbb{E}[D_{KL}((Y|t,[T])||(Y|t,[M]))].
\end{aligned}
$$

This representation makes it perhaps clearer that in order for the PI to be well defined, we need that if $p_{(Y,[M])}(y|t) = 0$ then also $p_{(Y,[T])}(y|t) = 0$ otherwise we have $p_{(Y,[T])}(y|t)\log_2 0$, which is not well defined. But this may occur for models that are bad representations of the unknown leakage, implying that the PI is not ideally suited to deal with models that are a poor approximation.

The remark makes apparent that the PI is indeed a quantity that is smaller or equal to $I(Y;[T])$, because the expected value of the KL divergence is positive. If $M = L$ then the expected value of the KL divergence is 0, and thus the PI equals to $I(Y;[T])$. If $M \neq L$ then the KL divergence is larger than zero and thus the PI measures (intuitively speaking) the amount of information that is lost on average if a specific model M is used.

Let us consider this alternative definition jointly with the MI-PI based Definition 1. This means the evaluator judges the model quality via the regret, i.e. they compute (with suitable estimators)

$$
\begin{aligned}
\mathsf{R}([M]) &= I(Y;[T]) - PI(Y;[T];[M]) \\
&= I(Y;[T]) - I(Y;[T]) + \mathbb{E}[D_{KL}((Y|t,T)||(Y|t,M))] \\
&= \mathbb{E}[D_{KL}((Y|t,T)||(Y|t,M))].
\end{aligned}
$$

This shows clearly that the model quality only depends on the average KL divergence between the joint distributions $(Y|t,[M])$ and $(Y|t,[T])$. But why would we then not simply aim to quantify the dependence between the joint distribution of the model M and the observable traces T with the KL divergence in the first place?

4.3 The Curse of Discretisation in the context of the eHI

The convergence guarantee for the eHI towards the MI requires the assumption that the traces are discrete. However, as we argued before, this assumption cannot be applied in general to side channel observations and it becomes invalid as soon as de-noising and other trace processing methods are used. Existing work so far has only been able to say that "this is potentially problematic". We now study the effect of discretisation on the MI.

Discretisation divides the range of a continuous random variable X into possibly an infinite number of intervals. Drawing on [25, cf. Proposition 1] we now provide a concrete mathematical characterisation for the MI between the a discrete and a discretised continous random variable.

The paper [25] considers two (continuous) random variables X, Y and the use of a simple partitioning of the space $X \times Y$ into rectangles. Typically, such a partitioning $\mathcal{P}$ is a product partitioning i.e. $\mathcal{P} = \mathcal{I} \times \mathcal{J}$ where $\mathcal{I}$ and $\mathcal{J}$ are partitioning of X and Y respectively [7]. We denote the discretised random variables obtained from such partitioning as $X^{\mathcal{I}}, Y^{\mathcal{J}}$.

We can now show that the MI which is based on the discretised leakage is smaller or equal to the MI based on the non-discretised leakage. This implies that an evaluator who discretises traces for the estimation of mutual information will underestimate the strength of an adversary who works with the non-discretised traces.

Proposition 1. *Let X, Y be two random variables with pmf p_X and pdf f_Y respectively. Let $\mathcal{P} = \mathcal{I} \times \mathcal{J}$ be the product partitioning of $X \times Y$ as described above (the partitioning $\mathcal{I}$ is defined by the discrete X). Then $I(X;Y) \geq I(X^{\mathcal{I}}; Y^{\mathcal{J}})$.*

Proof. We assume that the joint distribution exists. As explained in [25, Section II], for the product partition $\mathcal{P}$ we can write that

$$I(X;Y) = I(X^{\mathcal{I}}; Y^{\mathcal{J}}) + D_{\mathcal{P}}(X;Y)$$

where $D_{\mathcal{P}}(X;Y)$ is the residual divergence, see [25, cf. Proposition 1] for the definition.

It is shown in [25] that the residual divergence $D_{\mathcal{P}}(X;Y) \geq 0$ for any partition (including the specific partition that is given by a discrete X). Thus the result follows. □

Proposition 2 in [25, Section II] goes on to the develop that the residual divergence converges to zero asymptotically for increasingly finer product partitions. Consequently, in practice when we set the number of partitions to finite, the residual divergence is strictly larger than zero, and thus we always loose information upon discretisation:

$$I(X;T) > I(X;[T]) = \lim_{n \to \infty} \mathbb{E}[\mathrm{eHI}(X;[T])]. \tag{10}$$

In Sect. 7.4 we provide practical experiments that show the effect of Prop. in action. Proposition 4.3 also implies that the eHI is not necessarily an upper bound to the MI in the context of any arbitrary continuous traces (it also depends on the bias that it has, which is different in different settings).

[7] In the side channel community, a similar method is often implemented by partitioning the leakage into intervals, which then define the bins for histogram based estimation techniques—this is also the method used in Bronchain et al.[7] for the eHI.

4.4 Beyond HI and PI

The most recent work by Masure et al. [8] acknowledges that there are issues with both the HI and PI and clarifies mistakes in previous work. They also propose another MI estimator called the *Training Information* (TI) which is similar to the eHI. Instead of plugging in the so-called empirical distribution, they suggest to use both model distribution and the empirical distribution. The TI is therefore more akin to the PI and they show that the TI is an upper bound to the PI. Next they define the *Learnable Information* (LI) which is the supremum of the PI over all models of a given class of models. We mentioned before in this section that they define the *Regret*(R) which is the difference between the MI and the PI.

Thus also in the latest work, the idea of HI/PI based leakage certification prevails, and the new quantities TI and LI are mainly introduced for efficiency. But do we really need all these quantities to quantify the strength of a leakage model if in any case the regret only depends on the KL divergence between the joint distributions $(Y, [M])$ and $(Y, [T])$?

In this work we show that a much simpler, mathematically sound approach is possible for assessing the model quality by directly using MI. The results from statistics literature provides us a method to practically and efficiently estimate the MI (in a non-parametric way) for leakage certification. Furthermore, the method guarantees mathematically that the estimated MI is arbitrarily close to the actual MI which in practice is not known due to the unknown device leakage function or its output distribution.

5 Assessing Model Quality directly with the MI

In the previous section we have established that in the existing work, the regret of a model is in fact determined by the KL divergence between the joint distributions (Y, T) and (Y, M). We now develop a more intuitive approach to leakage certification, that also comes with better convergence guarantees for the required practical estimations.

We know that the best model (from an adversarial point of view) is the model that coincides with the device leakage function, i.e. $M = L$. If $M = L$ then we have that $I(M; T) = I(L; T)$. In contrast if $L \neq M$ (in the sense that L and M are independent), the the mutual information between them is zero $I(M; T) = 0$. If a model M is an approximation of L, then the mutual information should be somewhere between 0 and $I(L; T)$. This motivates the use of the mutual information to assess model quality.

Mutual information as a proxy for a distance metric. Informally speaking, we would like to use the mutual information as a metric by which we can judge how "close" a given model is to the "best model", or to compare two models. Although the mutual information does not satisfy the definition of a metric (the triangle inequality does not hold), we now explain why it is still a useful measure to compare models.

Notice that the variation of information does satisfy the definition of a metric. Given two distributions X and Y it is defined as $d(X, Y) = H(X) + H(Y) - 2I(X; Y)$ (we may understand the entropies here as either discrete or differential, depending on the nature of the X and Y). It is easy to see that if $I(X_1; Y) < I(X_2; Y)$, then $d(X_1, Y) > d(X_2, Y)$. Translating this to the scenario of leakage certification, we find that if the mutual information between model M_1 and the traces T is smaller than the mutual information between model M_2 and the traces, the the model M_1 is a better model than M_2.

Definition 3 (MI-based model quality). *Given two leakage models M_1 and M_2, we say that M_1 is a better leakage model than M_2 for a trace distribution T if $I(M_1; T) > I(M_2; T)$.*

Definition 4 (Estimating MI-based model quality). *Given two leakage models M_1 and M_2, and a consistent estimator I_n we say that M_1 is a better leakage model than M_2 for a trace distribution T if $I_n(M_1; T) > I_n(M_2; T)$.*

5.1 Reasoning about the Ideal Adversary

As a special case of comparing leakage models, an evaluator wishes to compare their model to the "best possible leakage model" which is evidently when $M = L$, which leads to the *best MI*, which we call I^b, see (11).

$$I^b = I(L(Y); T). \tag{11}$$

The challenge is that the evaluator does not know L, and that for different C, there will be different L. Hence this quantity needs to be efficiently estimated for each point in a given side channel trace.

An evaluator can however estimate the mutual information between the input and key and the observable traces, and we call this quantity I^{k}:

$$I^{\mathsf{k}} = I((X, K); T). \tag{12}$$

The connection between I^b and I^{k} is via the unknown leakage function L and the cryptographic target function C, which maps the key and input value to an intermediate value $Y = C(X, K)$. Using the data processing inequality, see [13], we know that $I^{\mathsf{k}} \leq I^b$ (observe that the variables in Fig. 1 follow a Markov chain).

From the data processing inequality we can also infer that equality holds if $L \circ C$ is one-to-one, which we cannot expect to hold in practice. However, the data processing inequality is a very crude tool to reason about these two quantities, and we show using a different proof technique in Sect.6 that equality holds under much more realistic conditions.

The fact that $I^b = I^{\mathsf{k}}$ under realistic conditions is essential for judging model quality in practice: given two (or more) leakage models M_i, we can estimate

$I(M_i;T)$, and compare this to I^{k}. In this way we cannot only compare the quality of leakage models relative to each other, we can also assess how "far" they are from the best model.

6 Proof that $I^b = I^{\mathsf{k}}$ in many Realistic Scenarios

In this section we show that $I^b = I^{\mathsf{k}}$ under some mild conditions, which we can expect to hold in many practical settings. This equality implies that in many practical cases I^b can be obtained via estimating I^{k} and thus without the need to know or even estimate L. With the help of this theoretical result, we implement a practical estimator [11] in the side channel setting, described in details in Sect. 7.

Based on the characteristics of the leakage functions (explained in Sect. 2.2), we select three possible cases for the subsequent proofs:

- L is discrete and deterministically depends on the realisations of X and K. which means, $T = L \circ C(X,K) + R$.
- L is discrete and probabilistically depends on the values of X and K. i.e., $T = L(S, C(X,K)) + R$, where, S follows a discrete distribution.
- Lastly, L is continuous and probabilistically depends on the realisations of X and K. i.e., $T = L(S, C(X,K)) + R$, where, S follows a continuous distribution.

Now, consider $Z = L \circ C(X,K)$, when L is deterministic and $Z = L(C(X,K),S)$, when L is probabilistic, then the MI for the ideal adversary, I^b, can be represented as $I(T;Z) = H(T) - H(T|Z)$, while the MI between the random inputs and the observable trace can be written as:

- $I^{\mathsf{k}} = I(T;(X,K)) = H(T) - H(T|(X,K))$
- $I^{\mathsf{k}} = I(T;(X,K,S)) = H(T) - H(T|(X,K,S))$

Clearly, I^b and I^{k} only differ from each other in the conditional entropy term. Consequently, our argument is be based on establishing the conditions under which these two conditional distributions are equal. A basic assumption in this section is thus that the conditional entropy exists.

6.1 Equality of I^b and I^{k} when L is discrete

Characterising the Conditional Distributions We first study the conditional distribution of $T|Z$. It is easy to see that this conditional distribution is completely defined by the distribution of R:

$$\begin{aligned} F_{T|Z}(t|z) &= P(T \leq t | Z = z) \\ &= P(Z + R \leq t | Z = z) \\ &= P(z + R \leq t) \text{ (as, } Z \text{ is independent of } R) \\ &= F_R(t - z)\ \forall t \in \mathbb{R} \end{aligned} \tag{13}$$

Consequently, the pdf $f_{T|Z}$ of the conditional variable $T|Z$ is given by the pdf of R.

We now consider the conditional distribution of $T|(X, K)$ when L is deterministic.

$$\begin{aligned} F_{T|(X,K)}(t|(x,k)) &= P(T \leq t|(X,K) = (x,k)) \\ &= P(L \circ C(X,K) + R \leq t|(X,K) = (x,k)) \\ &= F_R(t - L \circ C(x,k)) \ \forall t \in \mathbb{R} \end{aligned} \tag{14}$$

It follows again that the pdf of $T|(X, K)$ is given by the pdf of R. This observation has been formalised before in [26, Corollary 3.]. Note that, by using the same technique as above it is also obvious that when L is discrete and probabilistic,

$$F_{T|(X,K,S)}(t|(x,k,s)) = F_R(t - L(s, C(x,k))) \ \forall t \in \mathbb{R} \tag{15}$$

Now, with these properties of conditional distributions, we show that I^b is equal to I^{k} for both cases when L is deterministic and probabilistic, respectively.

Proposition 2. *If L is discrete and $T = L \circ C(X, K) + R$, then for any well-defined*[8] *function $C(\cdot)$, the following equality will hold*

$$I^b = I(T; Z) = I(T; (X, K)) = I^{\mathsf{k}}$$

Proof. We recall that $Z = L \circ C(X, K)$, and suppose it has m realisations. It is clear that the probability of $Z = z_i$ is given by the number of pairs (x, k) that map to z_i. Thus, we have

$$\begin{aligned} p_Z(z_i) = P(Z = z_i) &= P\{(X,K) = (x,k) : L(C(x,k)) = z_i\} \\ &= \sum_{(x,k):\ L(C(x,k))=z_i} p_{X,K}(x,k) \quad \text{for } i = 1,2,..,m \end{aligned} \tag{16}$$

(Here, every pair (x, k) maps to exactly one z_i, because C is well defined). We use this observation to rewrite I^b :

[8] An assignment of values y to elements $x \in \mathcal{X}$ is said to be a *well-defined* function $f: X \to Y$ if it satisfies the following three properties:

- Totality: For every $x \in \mathcal{X}$, $\exists$ y such that $f(x) = y$.
- Existence: For every $x \in \mathcal{X}$, $f(x) \in \mathcal{Y}$.
- Uniqueness: For every $x \in \mathcal{X}$, there is only $y \in \mathcal{Y}$ such that $f(x) = y$.

$$\begin{aligned}
I^{\mathsf{k}} &= I(T;(X,K))\\
&= H(T) - \sum_{(x,k)\in(\mathcal{X}\times\mathcal{K})} p_{(X,K)}(x,k)\mathbb{E}_{T|(x,k)}\left[-\log_2(f_{T|(X,K)}(t|(x,k)))\right]\\
&= H(T) - \sum_{i=1}^{m}\sum_{(x,k):\ L(C(x,k))=z_i} p_{(X,K)}(x,k)\mathbb{E}_R\left[-\log_2(f_R(t-L\circ C(x,k)))\right] \text{ (by Eq.(14))}\\
&= H(T) - \sum_{i=1}^{m} p_Z(z_i)\mathbb{E}_R\left[-\log_2(f_R(t-z_i))\right] \text{ (by Eq. (16))}\\
&= H(T) - \sum_{i=1}^{m} p_Z(z_i)\mathbb{E}_{T|z_i}\left[-\log_2(f_{T|Z}(t|z_i))\right] = I^b \text{ (by Eq.(13))}
\end{aligned}$$

Proposition 3. *Suppose, R follows a distribution with the location and scaling parameters μ and σ (> 0) respectively. Let X, K denote the plaintext and key (both independently drawn and distributed uniformly), and the leakage function L is discrete and $T = L(S, C(X,K)) + R$. If, the differential entropy of R is independent of location shift[9] (i.e., $H(R) = \phi(\sigma)$, where ϕ depends only on the pdf f_R), then the following equality holds:*

$$I^b = I(T;Z) = I(T;(X,K,S)) = I^{\mathsf{k}}$$

Proof. First, we compute I^b:

$$\begin{aligned}
I^b = I(T;Z) &= H(T) - H(T|Z)\\
&= H(T) - \sum_{z\in\mathcal{Z}} p_Z(z)\mathbb{E}_{T|z}\left[-\log_2(f_{T|Z}(t|z))\right]\\
&= H(T) - \sum_{z\in\mathcal{Z}} p_Z(z)\mathbb{E}_R\left[-\log_2(f_R(t-z))\right] \text{ (by (13))}
\end{aligned}$$

Second, we derive I^{k}:

$$\begin{aligned}
I^{\mathsf{k}} &= I(T;(X,K,S))\\
&= H(T) - \sum_{x,k,s} p_{(X,K,S)}(x,k,s)H(T|(x,k,s))\\
&= H(T) - \sum_{x,k,s} p_{(X,K,S)}(x,k,s)\mathbb{E}_{T|(x,k,s)}\left[-\log_2(f_{T|(X,K,S)}(t|(x,k,s)))\right]\\
&= H(T) - \sum_{x,k,s} p_{(X,K,S)}(x,k,s)\mathbb{E}_R\left[-\log_2(f_R(t-L(s,C(x,k)))\right] \text{ (by (15))}
\end{aligned}$$

[9] An illustration of location independent entropy:
Suppose, X_1 and X_2 follow normal distribution with different means μ_1 and μ_2, respectively but have same variance σ^2. Then, $H(X_1) = H(X_2) = \frac{1}{2}\log_2(2\pi e\sigma^2)$

Clearly, we already know from the entropy condition that $H(R) = \phi(\sigma)$, when $R \sim f_R(t-z)$ or when $R \sim f_R(t - L(s, C(x,k)))$. Hence, we can say that $\mathbb{E}_R\left[-\log_2(f_R(t-z))\right]$ is equal to $\mathbb{E}_R\left[-\log_2(f_R(t - L(s, C(x,k))))\right]$, which implies $I^b = I^k$.

6.2 Equality of I^b and I^{k} when L is continuous

Characterising the Conditional Distributions The continuity of L is due to some randomness of the continuous variable S that depends on X, K and the target function C but importantly we still have the independence between $Z = L(Y)$ and R. To derive the distribution of $T|Z$ (and then $T|(x,k,s)$) we need a little bit more machinery than before because L is continuous (was not covered by [26, Corollary 3.]).

The distribution of a function of two random variables (given their joint distribution) can be derived by a technique that is known as "change of variables" [27]. The trick works as follows, given two variables (X_1, X_2) and two functions u_1 and u_2 such that $Y_1 = u_1(X_1, X_2)$ and $Y_2 = u_2(X_1, X_2)$, with inverses $X_1 = v_1(Y_1, Y_2)$ and $X_2 = v_2(Y_1, Y_2)$; the joint pdf of (Y_1, Y_2) is given by $f_{(Y_1,Y_2)}(y_1, y_2) = |J| \cdot f_{(X_1,X_2)}(x_1, x_2)\big|_{\{x_1 = v_1(y_1,y_2), x_2 = v_2(y_1,y_2)\}}$. The value $|J|$ is the absolute value of the Jacobian $J = \left|\frac{\partial(x_1,x_2)}{\partial(y_1,y_2)}\right| = \begin{vmatrix} \frac{\partial(x_1)}{\partial(y_1)} & \frac{\partial(x_1)}{\partial(y_2)} \\ \frac{\partial(x_2)}{\partial(y_1)} & \frac{\partial(x_2}{\partial(y_2)} \end{vmatrix}$. Knowledge of the joint distribution (Y_1, Y_2) enables to derive the distributions of Y_1 (and Y_2 respectively) by marginalisation.

We first derive the distribution of $T|Z$. Hence we apply the change of variables technique to derive the distribution of $T = Z + R$, Z, and choose $Y_1 = Z + R$, $Y_2 = Z$. Hence $|J| = 1$, and this gives

$$\begin{aligned} & f_{T,Z}(t,z) = 1 \cdot f_{R,Z}(t-z, z) = 1 \cdot f_Z(z) \cdot f_R(t-z) = f_Z(z) \cdot f_R(t-z) \\ & \Rightarrow f_{T|Z}(t,z) = \frac{f_{T,Z}(t,z)}{f_Z(z)} = \frac{f_Z(z) \cdot f_R(t-z)}{f_Z(z)} = f_R(t-z) \end{aligned} \tag{17}$$

Using the same trick, we can also derive the pdf of $T|(x,k,s)$, which will give us $f_R(t - L(s, C(x,k)))$. To achieve this, we have to consider the following change of variables for each pair $(x,k) \in (\mathbb{X}, \mathbb{K})$:

$$(R, S) \to (T, S) : T = L(S, C(x,k)) + R$$

And the Jacobian of the transformation $J = \frac{1}{\left|\frac{\partial(t,s)}{\partial(r,s)}\right|} = 1$ under the condition that the mapping $L:\ S \to L(S, C(x,k))$ is one-to-one, which is a criterion for the existence of the partial derivative $\frac{\partial(t)}{\partial(s)}$ (for details see [27]).

Using this property of conditional distribution we now proof the equality between I^b and I^{k} exactly as same as we did in Proposition 3.

Proposition 4. *Suppose, R follows a distribution with the location and scaling parameters μ and σ (> 0) respectively. Let X, K denote the plaintext and key* *(both independently drawn and distributed uniformly), and the leakage function L is continuous and $T = L(S, C(X, K)) + R$. If, the differential entropy of R is independent of location shift (i.e., $H(R) = \phi(\sigma)$, where ϕ depends only on the pdf f_R), then the following equality holds:*

$$I^b = I(T; Z) = I(T; (X, K, S)) = I^{\mathsf{k}}$$

Proof. We are going to use the same proof technique as in Proposition 3 only by replacing the summation with the integration:

$$\begin{aligned} I^b = I(T; Z) &= H(T) - H(T|Z) \\ &= H(T) - \int_z f_Z(z) H(T|z) dz \\ &= H(T) - \int_z f_Z(z) \mathbb{E}_{T|z} \left[-\log_2(f_{T|Z}(t|z)) \right] dz \\ &= H(T) - \int_z f_Z(z) \mathbb{E}_R \left[-\log_2(f_R(t - z)) \right] dz \end{aligned}$$

We now derive I^{k} as in the following:

$$\begin{aligned} I^{\mathsf{k}} &= I(T; (X, K, S)) \\ &= H(T) - \sum_{x,k} \int_s f_{(X,K,S)}(x, k, s) H(T|x, k, s) ds \\ &= H(T) - \sum_{x,k} \int_s f_{(X,K,S)}(x, k, s) \mathbb{E}_{T|(x,k,s)} \left[-\log_2(f_{T|(X,K,S)}(t|x, k, s)) \right] ds \\ &= H(T) - \sum_{x,k} \int_s f_{(X,K,S)}(x, k, s) \mathbb{E}_R \left[-\log_2(f_R(t - L(s, C(x, k)))) \right] ds \end{aligned}$$

Based on the entropy criteria of R it is known that $H(R) = \phi(\sigma)$ irrespective of whether $R \sim f_R(t - z)$ or $R \sim f_R(t - L(s, C(x, k)))$. Therefore, we have

$$\begin{aligned} \mathbb{E}_R \left[-\log_2(f_R(t - z)) \right] &= \phi(\sigma) = \mathbb{E}_R \left[-\log_2(f_R(t - L(s, C(x, k)))) \right] \\ \Rightarrow I^b &= H(T) - \phi(\sigma) = I^{\mathsf{k}} \end{aligned}$$

Remark 2. From Proposition 2, we see when L is discrete and there is no internal randomness S, the equality $I^{\mathsf{k}} = I^b$ holds for any arbitrary distribution of noise R. However, Propositions 3,4 indicate that under the presence of the internal randomness S to find the quality we need the distributional assumption(entropy is independent of location shift) of the random noise R.

Must we check the condition of the distribution of *R*? We wish to point out that the distributional assumption (entropy is location independent) about the noise R holds for **all** the distributions that so far have been mentioned in the existing side channel literature (e.g., [2]). In particular, the entropy criteria is applicable to distributions like Gaussian, Laplacian, Cauchy, Uniform, etc.

However, it is possible to check this assumption efficiently if this is desireable. For a given set of traces and intermediate values one can check the entropy condition by performing for instance a The Kolmogorov-Smirnov test for goodness of fit [28] on samples of the leakage partition $(T|Y = y)$ to determine which distribution they come from.

7 Practical MI Estimation Using the GKOV Method

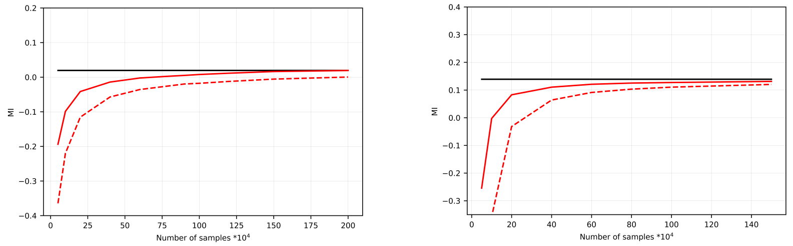
The recently proposed GKOV estimator [11] is convergent in mean and thus is asymptotically unbiased for all combinations of random variables. In contrast to previous nearest neighbour estimators, the number of nearest neighbours that are considered in the estimator is now a function of the sample size n (thus denoted as t_n), rather than a constant. The estimator is also efficient for multivariate settings. Hence, depending on the scenario that is considered in an evaluation, the GKOV estimator can be calculated for each point in a leakage trace independently of all other points (univariate setting), or over multiple points (multivariate setting). We know provide some details about implementing the estimator efficiently, and we investigate if there are criticial parameter choices.

7.1 Fast implementation of Alg.1

The authors of [11] provide a nice Python implementation of their estimator[10]. However, we developed a much more efficient and generic implementation that works for high dimensional data, which is important for side-channel analysis. For our C++ implementation, we used the popular machine learning library mlpack. The library offers several in-built distance metrics including the option of providing a custom distance metric. From the available options of efficient nearest neighbour search algorithms we used VPTree and BallTree. Note that the search algorithm may depend on the choice of distance metric. For example, the ℓ_∞ metric is not compatible with the KDTree search algorithm. This is not a limitation of mlpack but a consequence of the mathematical requirements of a specific search algorithm.

For calculating distances of each sample point from all other points which is necessary beyond the NN search, we have used OpenMP to parallelize the computation. Note that the OpenMP library can also be used by mlpack if it is available on the system. A particular observation on this part of our experiment is that for multidimensional leakage, computing the ℓ_∞ norm with an unrolled loop is more efficient than using the looped version or the mlpack library function for the same. For example, with the dimension $m = 2$, computing the ℓ_∞ norm as

[10] https://github.com/wgao9/mixed_KSG/blob/master/mixed.py



(a) L linear, $R \sim \mathcal{L}(0, 8)$

(b) L non-linear, $R \sim \mathcal{N}(0, 10)$

$t_n = \log(n)$:▬ , $t_n = \log_{10}^2(n)$:- - -, $I((X, K); T)$:▬

Fig. 2: Convergence experiments for different choices of t_n.

```
max( abs(data(i,0)-data(j,0)), abs(data(i,1)-data(j,1)) );
```

is more efficient than using the library function

```
arma::norm(data.row(i)-data.row(j), "inf");
```

For all experiments we have used an Intel(R) Core(TM) i7-8700 CPU 3.20GHz system having 6 CPU cores and Ubuntu operating system.

7.2 GKOV in a Multivariate Setting

The estimator by Gao et al.[11] does elegantly generalise to multiple points because its' only configuration parameter is the function t_n (based on the sample size). This is a significant advantage over previous t–NN estimators. The only remaining computational challenge is measuring the distance of all sample points L_j from the sample point L_i where $j \neq i$ for each i. A number of efficient algorithms for finding nearest neighbours are part of common machine learning libraries in both C/C++ and Python, and our implementation, as explained before, takes advantage of an existing machine learning library.

In contrast, the computational cost for estimating the mutual information in a multivariate setting using a histogram method (pdf estimation method) requires to adapt the choice of bins. For finding a "good" binning strategy one may need to compute I_n for range of values of the tuple $(b_1, b_2, \ldots, b_m) \in \mathbb{Z}^m$, where b_i denotes the number of bins along each dimension. This naturally increases the cost of estimating the mutual information using a histogram method.

We will include a range of multivariate experiments in the next section, where we include estimators from previous work.

7.3 Establishing Practical Choices for t_n

The parameter t_n, which is a function of the number of side channel observations n, is chosen by observing the convergence of the sequences $t_n \log n/n$ and

$(t_n \log n)^2/n$ (the sequences can be found in the main theorem statement of [11]). In our experiments we selected t_n equal to $\log n$ and $\log_{10}^2 n$. Figure 2 shows some representative experimental results for the GKOV estimator as implemented via (Alg. 1) in different situations. To create these plots, we performed a number of simulations where we varied both device leakage functions and noise distributions. Each simulation is performed multiple times, and we show the average over the outcomes. To provide a baseline for comparison, we also calculated the MI in all scenarions, which was possible because in simulations we know all distribution parameters.

The results in Fig. 2 illustrate that for both choices of function t_n, the convergence rate is similar, with a small advantage for $t_n = \log n$. In the remaining practical experiments, we will thus show results for $t_n = \log n$.

It is important to bear in mind that unlike a plug-in (histogram) estimator that requires data dependent parameter tuning, the choice of the parameter t_n can be pre-determined based only on the sample size n. Furthermore the choice of t_n only affects the rate of convergence, i.e. the efficiency of the estimation unlike histogram based estimators, where a wrong choice can lead to bias.

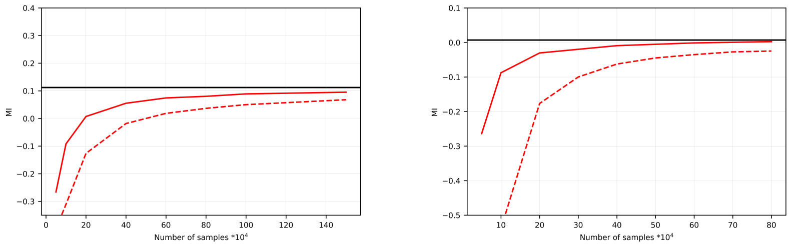
An observation is that the GKOV estimator approaches the true MI from below. There is no formal proof for this in [11], but in all our experiments we observed this behaviour. This implies that if an MI quantity is close to zero, then the GKOV estimator will take negative values, until enough samples are available and it crosses the zero line and is positive. This behaviour is not a sign of bias (note that [11] shows the asymptotic unbiasedness of their estimator).

7.4 Practical Demonstration of the Adverse Effect of Discretisation

Having established a suitable practical configuration for the GKOV estimator, we now use it to demonstrate the information loss that is incurred by the discretisation of traces (see Sect. Sect. 4.3 for the theoretical discussion) with some practical experiments.

Both experiments are based on simulated side channel observations. The first experiment is based on assuming a non-linear device leakage function with Laplacian noise. The second experiment is based on a linear device leakage function and uses Gaussian noise. Because we know all distributions, we can compute the mutual information theoretically, and we include this quantity in the traces (the black line). For each experiment, we then estimate the mutual information using GKOV for both the traces as they are generated, and for the traces after discretisation. Figures 3a and 3b show the outcomes: in both cases the mutual information estimate that is based on the discretised traces does not reach the true (higher) mutual information value: $I((X,K);[T]) < I((X,K);T)$. This is exactly what our theoretical analysis showed would happen.

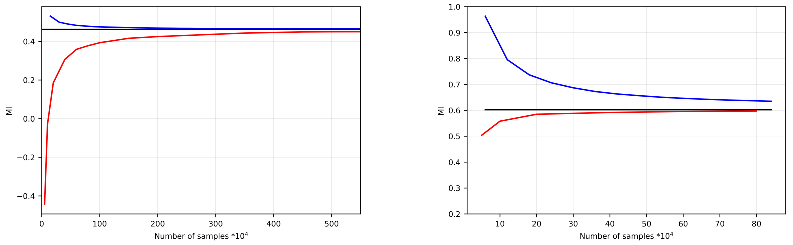
We have explained before that discretisation is inevitable if the eHI is used to estimate mutual information. Figure 4a shows the eHI vs the GKOV estimate when naturally discrete variables are in play (note that we add discrete noise in this experiment). We see that eHI and GKOV both approach the true mutual information, which we were able to calculate directly because all distributions



(a) L nonlinear, $R \sim \mathcal{L}(0, 10)$ (b) L linear, $R \sim \mathcal{N}(0, 10)$

$I((X,K);T)$:— , $I((X,K);[T])$:- - -, $I((X,K);T)$:—

Fig. 3: Discretisation lowers the mutual information



(a) L nonlinear, $R \sim$ discrete $\mathcal{L}(0, 5.64)$ (b) L nonlinear, $R \sim \mathcal{N}(0, 4)$

eHI:— ,$I((X,K);T)$:— , $I((X,K);T)$:—

Fig. 4: Discretisation implies eHI does not coverge to the true mutual information

are know here to us. Figure 4b shows, for the same device leakage function, but with continuous noise added, that the eHI is no longer able to reach the true mutual information. We also include a further non-parametric discrete mutual information estimator (the plug-in estimator which is comparable to eHI as it comes with the same notion of convergence as discussed before in Sect. 3), which has the same problem when discretisation occurs.

8 Case Studies in Evaluation Settings

So far we (mostly theoretically) argued that leakage certification using the HI-PI method is unnecessarily complicated and prone to misleading results. We made the theoretical case that it is possible to instead estimate salient mutual information quantities, in particular I^{k} elegantly via the GKOV estimator. In this section we show via a range of experiments, which are reflecting evaluation scenarios, that theory indeed translates into practice.

Like in previous work, we use simulations to produce fully controlled experiments, so that the mutual information can both be calculated as well as estimated (the black reference line is the true, theoretically calculated quantity). Simulations also enable to make experiments scalable in terms of using different device leakage functions, types of noise, noise parameters, etc. and to efficiently examine multivariate settings. To complement simulations we also show some results based on data that was sampled from a 32-bit device.

8.1 Simulation setup

In all experiments we consider a single bijective target function, which is the AES SubBytes mapping, $y = C(x, k) = SubBytes(x \oplus k)$. In our simulations, we vary the device leakage function as well as the type and magnitude of the noise distribution, and we consider univariate and multivariate analyses.

In the univariate simulations we utilise as device leakage functions:

HW: $L = \mathrm{HW}(Y)$ (Hamming weight of Y),
HD: $L = \mathrm{HD}(Y, C(Y))$ (Hamming distance between Y and $C(Y)$),
non-linear: $L = \text{DES-Sbox}\,(\text{6LSB}(Y))$ (The first DES Sbox applied to the 6 least significant bits of Y), and

In the multivariate simulations the simulated trace points are either based on either HW or HD leakage of some bits of Y (this is only to speed up experiments). For instance, the bivariate simulations are based on the tuples $(\mathrm{HW}(\text{4LSB}(Y)), \mathrm{HW}(\text{4MSB}(Y))$ or $\mathrm{HW}(\text{4LSB}(Y)), \mathrm{HD}(\text{4MSB}(Y))$ and the independent noise (R_1, R_2) is either bivariate $\mathcal{N}$ or bivariate $\mathcal{L}$ with $\sigma = 4$.

The noise R follows either a Gaussian ($\mathcal{N}(0, \sigma)$), a Laplacian ($\mathcal{L}(0, \sigma)$) or a discerete-Laplacian (discrete $\mathcal{L}(0, \sigma)$) distribution. In our experiments we consider $\sigma \in [2.8, 10]$.

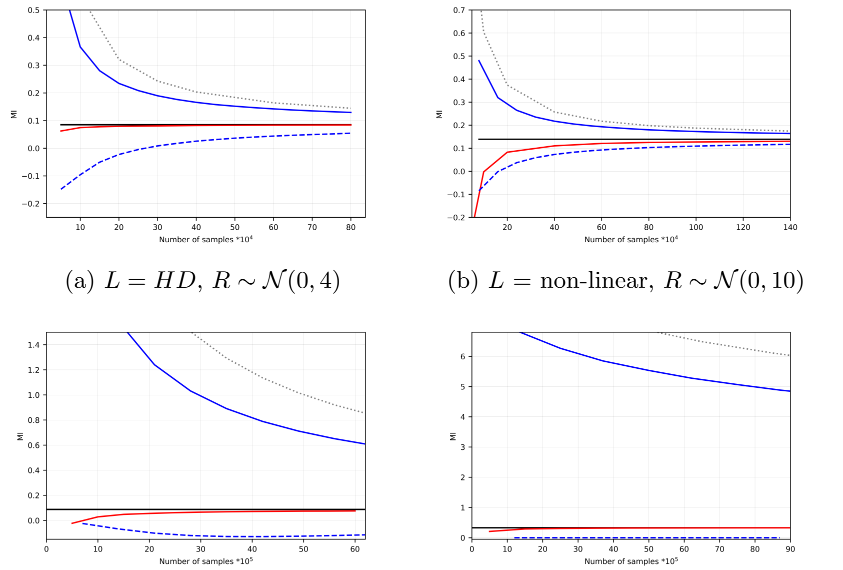
In order to compute the eHI, and the ePI, with use the scripts that were provided by the authors of [7]. It is important to bear in mind that the ePI and eHI are only defined for use with two discrete random variables, and the scripts of [7] include a step where traces are discretised.

8.2 Leakage Certification: Assessing the Ideal Adversary with the eHI-ePI vs $I^{\mathbf{k}}$

We ran a large number of experiments and include a representative subset of outcomes in Figures 5a-5d.

The experiments clearly demonstrate that the GKOV estimator (red line) quickly converges to the true mutual information value (black line), irrespective of the dimensionality of the leakage. In stark contrast, the eHI is biased and the bias increases dramatically with the number of dimensions, which is in line with [29].

In the context of leakage certification based on studying the difference eHI − ePI, we can observe that this difference increases when we move towards multivariate side channel observations. It is obvious from the experiments that a more trace efficient estimator for the eHI and ePI is needed, i.e. the result of



(a) $L = HD$, $R \sim \mathcal{N}(0, 4)$

(b) $L =$ non-linear, $R \sim \mathcal{N}(0, 10)$

(c) $L = (HW, HD)$, $R \sim \mathcal{N}(0, 4)$

(d) $L = (HW, HW, HD)$, $R \sim \mathcal{N}(0, 2)$

$I_n{}^{\text{hist}}$: ······ , eHI:▬ , $I((X,K);T)$: ▬ , $I_n{}^{\text{GKOV}}$:▬ , ePI:- - -

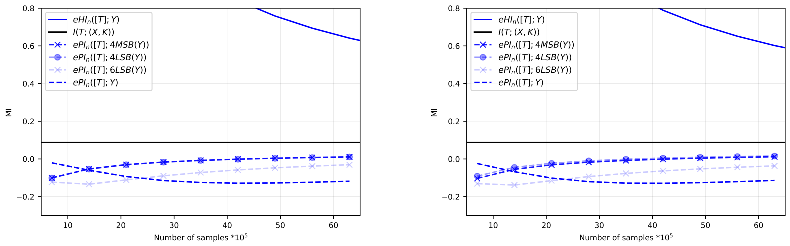
Fig. 5: Ideal Adversary: eHI-ePI vs I^{k}

[29] but these more trace efficient estimators still require discrete data or they are parametric.

In contrast, the direct estimation of I^b via I^{k} (we showed that this is mathematically sound in the theoretical part before), provides consistent results with fewer traces even in multivariate settings with only a mild assumption on the noise distribution. In an evaluation, the GKOV estimator should be used to estimate I^b via $I^{\mathsf{k}} = I((X, K); T)$.

In the results we include up to three dimensional side channel observations. This is only because we were unable to run ePI, eHI and the histogram based estimator for four shares — they require to explicitely build a multivariate pmf, which makes any higher order analysis computationally extremely expensive. But our experiments for GKOV on four dimensions again demonstrated quick convergence to the true MI value.

For completeness we also included the convergence of the histogram-based plug-in estimator: which is proven to have a weaker form of convergence in [10]. We can see that its performance is particularly bad, and it also appears to show bias (which is expected given Paninski [12]).



(a) $L = (HW, HW), R \sim \mathcal{N}(0, 4)$ (b) $L = (HW, HD), R \sim \mathcal{N}(0, 4)$

Fig. 6: Comparing models via eHI − ePI

8.3 Leakage Certification: Comparing Leakage Models

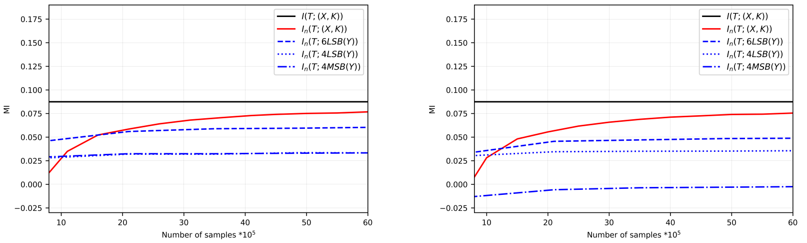
We now revisit the question of comparing two or more leakage models, in a set of controlled experiments. For this we define leakage models M in relation to some "true device leakage" L, whereby the models incorporate progressively less information of L. We achieve this by truncating Y and then we apply the device leakage function. Precisely, we consider the following leakage models:

- the model $M(Y) = L(6\text{LSB}(Y))$ is based on using just the six least significant bits of Y,
- the models $M(Y) = L(4\text{LSB}(Y)$ and $M(Y) = L(4\text{MSB}(Y))$ are based on using the four least or most significant bits of Y.

Note that the intermediate Y is the output of the AES SubBytes operation, thus $L(Y)$ has 8 bits. Consequently the 6LSB model should be a better predictor than the 4LSB or the 4MSB model. We expect that any measure, i.e. eHI − ePI and $I^{\mathsf{k}} - I(M;T)$, will correctly rate the 6LSB model as better than any of the 4LSB models.

With this in mind we examine the outcomes of our first bi-variate simulation that are given in Fig. 6a. The idea of the regret function was that a smaller regret indicates a better model. However, the ePI that is furthest away from the eHI, and thus gives the largest regret is ePI$(6LSB(Y);T)$ which is not what we should be seeing. The second bi-variate simulation is based on two points where one leaks the HW and one leads the HD: this is given in Fig. 6b. We see once more that the model that uses the most information is not closest to the eHI. We also have the exact MI value plotted as a black line which demonstrates the bias of eHI. These two experiments show that comparing models via the difference eHI − ePI is not sound.

We now perform the same model comparisons using GKOV. Figure 7a shows that the GKOV estimator approaches the exact MI as expected. Consequently $I^{\mathsf{k}} = I((X,K);T)$ and we expect that a better model is closer to I^{k}. We can see in Fig. 7a that the models stack up as they should: the 6LSB model is better than the 4LSB models. The experiment using two points that leak slightly



(a) $L = (HW, HW), R \sim \mathcal{N}(0, 4)$ (b) $L = (HW, HD), R \sim \mathcal{N}(0, 4)$

Fig. 7: Comparing models via $I^{\mathsf{k}} - I(M;T)$

differently confirm these observations: the GKOV estimator converges quickly to the exact MI and the MI estimates for the different models appear in the order that they should. In particular, when the we set the second component in the bivariate experiment to HD (and thereby introduce a further discrepancy to the the model prediction which is based on HW) we see that also the 4 bit models can be further discriminated. Summarising, leakage certification with the GKOV estimator delivers clear and correct results.

8.4 Leakage certification of a real device

Finally, we use a data set that was acquired from executing a two-share AES SubBytes implementation. The implementation runs on an ARM Cortex M3 processor core from NXP. We use a custom measurement board, which provides good measurements. We use our scope in a basic setting to avoid any trace processing (de-noising) and extract discrete measurements, where each point is represented by 8 bits. This means that eHI and ePI can work on naturally discrete traces, which is what they were designed for. However, we apply them to two trace points at a time, thus we do the analysis in a bivariate setting.

Figure 8 shows the result of this experiment. It is striking to see that there are many trace points that are highlighted by the GKOV estimate for $I^{\mathsf{k}} = I((X,K);T)$ as showing dependency: these are all the points where the red line is significantly higher than zero. However, the blue lines do not bound the red line, which is what they are supposed to do. Although also the eHI − ePI difference indicates that this trace shows data dependencies, the quanitites do not highlight all the trace points correctly, which is a problem.

9 Conclusions

If the process of leakage certification is to show an actual advantage in an evaluation setting, this process would need to reliably and efficiently identify data dependencies in observed side channel traces, it would need to be able to assess

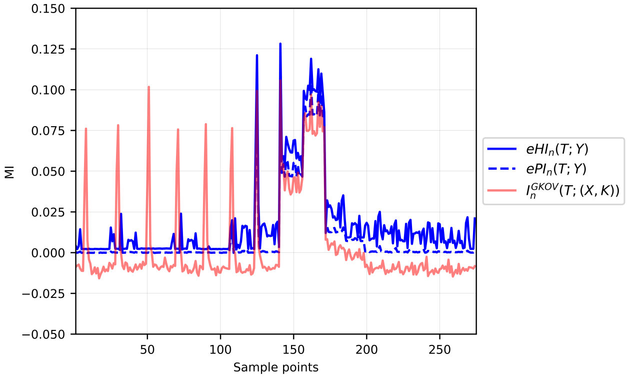


Fig. 8: Bi-variate discrete real device leakage

the ideal adversary, and it would need to be able to correctly assess leakage models of varying quality.

In this paper we have examined the existing method based on the HI and PI quantities which are meant to bound $I(Y;T)$. We have found that the bias of the eHI, which converges to $I(Y;T)$ increases such that it becomes a useless upper bound. We also found a different way to express the PI that shows that the difference $\mathrm{MI} - \mathrm{PI}$ could as well be directly expressed via the Kullback-Leibler divergence applied to a given model and some traces.

With this we asserted that a much more natural approach to leakage certification is to directly estimate mutual information quantities of interest. These quantities can only be understood in relation to the mutual information that characterises the ideal adversary $I^b = I(L(Y);T)$. This quantity seems impossible to estimate (because L is unkown) however, we show through a careful characterisation of the distributions that in many realistic settings the equality $I^b = I^{\mathsf{k}} = I((X,K);T)$ holds. Consequently we can characterise the ideal adversary in practice by the estimation of I^{k}. With this we show in a range of experiments that our proposal for leakage certication produces consistent and efficient results, whereas the state of the art approach based on the PI fails.

Our results are good news for all those practitioneres who wish to assess an implementation (with respect to the ideal adversary) or who wish to show that their leakage models are sound: our approach is sound, simple to understand, and efficient to implement.

References

1. Grosso, V., Standaert, F.: Masking proofs are tight and how to exploit it in security evaluations. In Nielsen, J.B., Rijmen, V., eds.: Advances in Cryptology - EUROCRYPT 2018. Volume 10821., Springer (2018) 385–412

2. Heuser, A., Rioul, O., Guilley, S.: Good is not good enough - deriving optimal distinguishers from communication theory. In Batina, L., Robshaw, M., eds.: Cryptographic Hardware and Embedded Systems - CHES 2014 - 16th International Workshop, Busan, South Korea, September 23-26, 2014. Proceedings. Volume 8731 of Lecture Notes in Computer Science., Springer (2014) 55–74
3. Durvaux, F., Standaert, F., Veyrat-Charvillon, N.: How to certify the leakage of a chip? In Nguyen, P.Q., Oswald, E., eds.: Advances in Cryptology - EUROCRYPT 2014 - 33rd Annual International Conference on the Theory and Applications of Cryptographic Techniques, Copenhagen, Denmark, May 11-15, 2014. Proceedings. Volume 8441 of Lecture Notes in Computer Science., Springer (2014) 459–476
4. de Chérisey, E., Guilley, S., Rioul, O., Piantanida, P.: Best information is most successful mutual information and success rate in side-channel analysis. IACR Trans. Cryptogr. Hardw. Embed. Syst. **2019**(2) (2019) 49–79
5. Renauld, M., Standaert, F.X., Veyrat-Charvillon, N., Kamel, D., Flandre, D.: A Formal Study of Power Variability Issues and Side-Channel Attacks for Nanoscale Devices. In: EUROCRYPT. (2011) 109–128
6. Durvaux, F., Standaert, F.X., Del Pozo, S.M.: Towards Easy Leakage Certification. In Gierlichs, B., Poschmann, A.Y., eds.: Cryptographic Hardware and Embedded Systems – CHES 2016, Berlin, Heidelberg, Springer Berlin Heidelberg (2016) 40–60
7. Bronchain, O., Hendrickx, J.M., Massart, C., Olshevsky, A., Standaert, F.: Leakage certification revisited: Bounding model errors in side-channel security evaluations. In Boldyreva, A., Micciancio, D., eds.: Advances in Cryptology - CRYPTO 2019 - 39th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 18-22, 2019, Proceedings, Part I. Volume 11692 of Lecture Notes in Computer Science., Springer (2019) 713–737
8. Masure, L., Cassiers, G., Hendrickx, J., Standaert, F.X.: Information bounds and convergence rates for side-channel security evaluators. Cryptology ePrint Archive, Paper 2022/490 (2022) `https://eprint.iacr.org/2022/490`.
9. Prouff, E., Rivain, M., Bevan, R.: Statistical analysis of second order differential power analysis. IEEE Trans. Computers **58**(6) (2009) 799–811
10. Antos, A., Kontoyiannis, I.: Convergence properties of functional estimates for discrete distributions. Random Structures & Algorithms **19** (10 2001) 163 – 193
11. Gao, W., Kannan, S., Oh, S., Viswanath, P.: Estimating mutual information for discrete-continuous mixtures. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17, Red Hook, NY, USA, Curran Associates Inc. (2017) 5988–5999
12. Paninski, L.: Estimation of Entropy and Mutual Information. Neural Computation **15**(6) (2003) 1191–1253
13. Thomas M. Cover, J.A.T.: Elements of Information Theory. Wiley (2005)
14. Nair, C., Prabhakar, B., Shah, D.: On entropy for mixtures of discrete and continuous variables. arXiv preprint cs/0607075 (2006)
15. L. F. Kozachenko, N.N.L.: Sample estimate of the entropy of a random vector. Problems in Information Transmission **23** (1987)
16. Kraskov, A., Stögbauer, H., Grassberger, P.: Estimating mutual information. Phys Rev E Stat Nonlin Soft Matter Phys **69** (07 2004) pp. 066138
17. Belghazi, M.I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Courville, A., Hjelm, D.: Mine: mutual information neural estimation. In: ICML 2018. (June 2018) ArXiv.
18. Cristiani, V., Lecomte, M., Maurine, P.: Leakage assessment through neural estimation of the mutual information. In: ACNS 2020. Volume 12418 of Lecture Notes in Computer Science., Springer (2020) 144–162

19. McAllester, D., Stratos, K.: Formal limitations on the measurement of mutual information. In Chiappa, S., Calandra, R., eds.: Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics. Volume 108 of Proceedings of Machine Learning Research., PMLR (26–28 Aug 2020) 875–884
20. Batina, L., Gierlichs, B., Prouff, E., Rivain, M., Standaert, F.X., Veyrat-Charvillon, N.: Mutual Information Analysis: a Comprehensive Study. J. Cryptology **24**(2) (2011) 269–291
21. Beirlant, J., Dudewicz, E., Györfi, L., Dénes, I.: Nonparametric entropy estimation. an overview. INTERNATIONAL JOURNAL OF MATHEMATICAL AND STATISTICAL SCIENCES **6**(1) (1997) 17–39
22. Györfi, L., van der Meulen, E.C.: Density-free convergence properties of various estimators of entropy. Computational Statistics and Data Analysis **5**(4) (1987) 425–436
23. Hall, P., Morton, S.: On the estimation of entropy. Annals of the Institute of Statistical Mathematics **45** (02 1993) 69–88
24. Gao, S., Oswald, E.: A novel completeness test for leakage models and its application to side channel attacks and responsibly engineered simulators. In Dunkelman, O., Dziembowski, S., eds.: EUROCRYPT 2022, Part III. Volume 13277 of Lecture Notes in Computer Science., Springer (2022) 254–283
25. Darbellay, G., Vajda, I.: Estimation of the information by an adaptive partitioning of the observation space. IEEE Transactions on Information Theory **45**(4) (1999) 1315–1321
26. Heuser, A., Rioul, O., Guilley, S.: Good is not good enough. In Batina, L., Robshaw, M., eds.: Cryptographic Hardware and Embedded Systems – CHES 2014, Berlin, Heidelberg, Springer Berlin Heidelberg (2014) 55–74
27. Roussas, G.G.: Chapter 6 - transformation of random variables. In Roussas, G.G., ed.: An Introduction to Probability and Statistical Inference (Second Edition). Second edition edn. Academic Press, Boston (2015) 207–243
28. Massey Jr, F.J.: The kolmogorov-smirnov test for goodness of fit. Journal of the American statistical Association **46**(253) (1951) 68–78
29. Masure, L., Cassiers, G., Hendrickx, J., Standaert, F.X.: Information bounds and convergence rates for side-channel security evaluators. Cryptology ePrint Archive, Paper 2022/490 (2022) `https://eprint.iacr.org/2022/490`.