Leakage Certification based on Consistent MI Estimation

Arnab Roy, Aakash Chowdhury, and Elisabeth Oswald

University of Klagenfurt, Austria
{arnab.roy, aakash.chowdhury, elisabeth.oswald}@aau.at

Abstract. The mutual information between two variables is a key metric in the context of side channel attacks; in particular it is used to judge the quality of device leakage models. In practice the mutual information can only be estimated, and existing methods in the side channel community are based on density estimation. Estimating the mutual information based on estimating distribution densities is a challenge unless assumptions about the underlying distributions can be made — this is undesirable in the side channel setting because the underlying distributions are unknown.
We suggest a radically different approach to the mutual information estimation in the side channel setting based on a recently proposed k-Nearest Neighbour estimator. We prove that the mutual information between the key and the observed side channel can be efficiently estimated without the need for any density estimation, even in multivariate settings, and we mathematically characterise the impact of some assumptions/restrictions of previous work on the estimation process. To complement our theoretical results, we offer a wide range of experimental results that compare our proposal with the state of the art estimators used in the side channel community. Finally we show in experiments the advantages of our proposed method for judging the quality of leakage models, in comparison to the existing techniques.

1 Introduction

Side channel theory and practice require metrics to quantify the information leakage about secret cryptographic keys. The mutual information (MI) is such a metric, and it appears as part of security proofs in the context of masking, e.g. [1], in the context of optimal distinguishers, e.g. [2], and as a tool to quantify the quality of a leakage model, e.g. [3].

Evaluating device security via leakage certification. Attacks that extract and exploit information leakage are highly configurable, but they always require the extraction of information of small portions of the secret key from some observed side channel leakage (they follow a divide-and-conquer principle). The extraction of key information can be achieved with a wide range of statistical and machine learning tools: it is well known that the use of an accurate device

leakage model is necessary for optimal information extraction [4]. In order to understand the worst case security of a device, an evaluator wishes to assess the ideal adversary, who is in possession of the exact device leakage model, in relation of a practical adversary, who is in possession of an estimated leakage model.

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. An evaluator thus seeks to understand how good their (estimated) device leakage model is, which is a task that was formalised by Durvaux et al. [3] as *leakage certification*. In a series of follow on works [5, 6] the initial approach was refined by introduction of two metrics: the empirical hypothetical information (eHI) that captures the amount of information that could be extracted if the device followed exactly the so-called empirical distribution (an estimated probability mass function), and the empirical perceived information (ePI) that captures the relationship between a model and some observable leakage. The idea is that the eHI and ePI used jointly enable to judge a device leakage model in comparison to the ideal adversary who has access to the true leakage model: the eHI is shown to be an upper bound for the ideal adversary, and the ePI represents the best practical adversary. These statements only hold if all variables are discrete and univariate.

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 adversaries. Secondly, implementations that implement masking countermeasures are often analysed after further software processing, including filtering, and mean-free product-combining [7], which again create real-valued outputs.

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

- uses a strongly consistent (thus unbiased) estimator for all leakage functions that can be observed in practice (discrete, continuous, and even probabilistic functions),
- naturally extends to the multivariate setting, and
- has the same asymptotic convergence rate as existing estimators (which are limited to the discrete univariate setting).

More specifically we study the mathematical relationships between different mutual information quantities (defined in the existing side channel literature) to highlight the critical and adverse role that discretisation of measurements plays in the context of estimating the MI. Discretisation of continuous side channel measurements is necessary in previous work because all MI estimators can only

work with discrete variables. We show that the mutual information between the key and the observed traces is the quantity that captures the strength of the best adversary, and that this quantity can be estimated from the traces with a recently published estimator by Gao et al. [8]. We provide experiments that show the non-asymptotic convergence rate of the Gao estimator across different side channel scenarios. Finally we show that the use of existing metrics like the eHI and ePI in multivariate settings delivers biased outcomes.

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. We abuse notation and treat random variables and their corresponding sets synonimously. 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); for an arbitrary event we use $\mathbb{P}$ to denote its probability. 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).

For any random variable X, $\mathbb{E}(X)$ and resp. $\mathbb{E}_X$ denote the expectation. For a real valued variable x, $[x]$ denotes the integral part of the value.

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. The symbols log and ln denote the logarithm with base 2 and base e respectively. 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. $[X]$.

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 with image space L, which is also understood as a random variable, i.e. t 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 $x \in X$ for the input, which is mapped according to the cryptographic process via the application of some (cryptographic) target function(s) $\mathcal{C}$ and an (unknown) key $k^* \in K$ to an intermediate $y \in Y$. Implementations process cryptographic keys in "chunks", thus K and X have small support.

An adversary is assumed to be able to observe inputs/outputs $x \in X$ of the device and the side channel leakage trace $t \in T$ that corresponds to the execution of a cryptographic algorithm using the input and the key k^* that is embedded in the device. A side channel trace is a vector of leakage points. Each point corresponds to the physical processes that happen inside the device (at that point in time/step in the execution) and some independent noise R.

Leakage functions. An important, but in the existing literature often ignored, detail is that the observed leakage T may be either a deterministic or a probabilistic function of multiple variables[1]. The secret key $k^* \in K$ and the input X interact via the target function $\mathcal{C}$ (a step in the computation of the cryptographic algorithm) and leak via the (unknown) leakage function L. The leakage function for a specific step in the execution of an algorithm can be simple (e.g. Hamming weight for a bus transfer), in which case it can be modelled as a discrete deterministic random variable:

$$T = L(\mathcal{C}(X, K)) + R.$$

This means that the output of the leakage function is fully defined by inputs (e.g. key, and plaintext byte), and the same inputs will always give the same outputs.

But for many steps in an execution the leakage function depends on some complex interaction between many components in the device, and is influenced by probabilistic processes (due to glitches, cross talk, couplings, etc.). It is also possible that the measurement setup itself impacts on the leakage, or that some post-processing to increase trace quality is used. As a result, leakage functions are often better understood to be continuous probabilistic random variables. This means that the output of the function still depends on the inputs, but supplying the same inputs can lead to different outputs (e.g. glitches impact on the power consumption).

We model such a probabilistic leakage function by some unknown internal randomness S. Note that S can be discrete or continuous, and it is different from R. Unlike R, the random variable S is not independent of (X, K) i.e. the leakage density function is $f(x, k, s)$. For a target device we can then model the observed leakage as

$$T = L(S, \mathcal{C}(X, K)) + R.$$

[1] A deterministic function is fully determined by its' inputs. A probabilistic function includes an element of chance.

Previous work on leakage certification has exclusively considered deterministic leakage functions, but we consider deterministic and probabilistic leakage functions in our work. This is important because we do want a mutual information estimation process to be strongly consistent *for all observable leakage functions*, even if we don't understand their true nature. 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 discrete and deterministic or continuous and probabilistic, we drop S in the text for readability.

In the rest of this paper, T should always be understood as continuous (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.

2.3 MI definitions and estimation considerations

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

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

If either both variables are discrete, or both variables are continuous, the MI can be expressed via the marginal and joint or conditional entropies, leading to the well known "2H" and "3H" expressions for MI:

$$I(X;Y) = H(X) - H(X|Y) \tag{1}$$
$$= H(X) + H(Y) - H(X,Y) \tag{2}$$

If one variable is discrete and one is continuous then the conditional density in the 2H formula, and the joint density in the 3H formula, may not exist[2], and thus the MI cannot always be derived in general using these formulae.

One workaround for this problem is to discretise the continuous variable, so that both variables are discrete. However the choice of a discretisation function is complex and results in biased estimators [10]. The same problem persists when considering mixture distributions, i.e. when one variable is a mixture of a discrete and a continuous variable.

Consequently, MI estimators need to cope with different constellations of random variables (discrete or continuous) and there exist three cases:

- two continuous random variables (referred to as cont. MI)
- two discrete random variables (referred to as discrete MI)
- at least one mixed random variable (referred to as mixed MI)

[2] Nair et al. provide a number of conditions that must hold for the conditional density to exist [9], and if so explain a natural extension which recover the 2H and 3H expressions.

The crucial property of any 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 more and more samples from the unknown distribution. 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.

2.4 Non-parametric MI estimation

In the side channel setting the true distribution of T is unknown; and thus in some evaluation tasks (such as leakage certification) we do not wish to use estimators that require assumptions about the distributional properties of T. Such assumption-free estimators are called non-parametric estimators in statistics.

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 [11]. The before mentioned limitations (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 [12], 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 in 2018 [13] and suggested to be used for side channel tasks in [14]. However, it was shown later in [15] that the claimed convergence rates were erroneous.

Ultimately, a recent contribution by Gao et al. [8] made a further significant step by estimating the Radon-Nikodym derivative 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 (6). 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.

Entropy based MI estimation. The most widely applied technique of MI estimation in side-channel analysis is via entropy estimation e.g. via the 2H formula:

$$I_n(X, Y) = H_n(X) - H_n(X|Y). \tag{3}$$

In general, the estimation of $H(X|Y)$ i.e. computing $H_n(X|Y)$ requires estimating $H(X|y)$ for each $Y = y$ as well as pdf or pmf of Y. In the context of side-channel analysis one random variable (r.v.), namely Y is discrete, it assumes finitely many values, and is typically uniformly distributed. Thus, it suffices to estimate only $H(X|Y = y)$ for each possible value $y \in Y$. In fact, the uniformity of Y makes 2H estimation a natural choice in side channel analysis. Evidently, the convergence of the MI estimator based on the 2H approach depends fully on the convergence of the entropy estimator.

In the side channel literature, based on the simplifying assumption of having discrete traces, the study in [16] use an integral estimate [17]. Györfi and van Meulen [18] 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 [19] (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 [20]. This convergence result appears to not be known in the side channel literature, and instead the eHI was developed as a means to bound the MI.

Defining eHI and ePI. Bronchain et al. [6] put forward the notions of ePI and eHI (see equations (4) and (5)); these quantities are based on the 2H entropy estimate for MI. These estimators are based on estimating an *empirical pdf*, denoted by $\tilde{e}_n$, which is based on the discretized leakage $[t] \in [T]$, and a key dependent variable Y, and is derived in the form of a histogram (thus this is a non-parametric estimator).

$$\mathrm{eHI}_n(Y;[T]) = H([T]) + \sum_{[t]\in[T]} p_{[T]}([t]) \cdot \sum_{y\in Y} \tilde{e}_n(y|[t]) \log_2 \tilde{e}_n([t]|y) \tag{4}$$

$$\mathrm{ePI}_n(Y;[T]) = H([T]) + \sum_{[t]\in[T]} p_{[T]}([t]) \cdot \sum_{y\in Y} p_n(y|[t]) \log_2 \tilde{e}_n([t]|y) \tag{5}$$

Bronchain et al. make the common assumption (also used in previous and related papers like [3, 5, 6]) that the noise distribution is Gaussian (e.g. $R \sim \mathcal{N}(\mu, \sigma)$). Assuming that the distribution of the key is uniform, Bronchain et al. [6] show that the eHI converges in probability to the $I([T];Y)$ (with $Y = \mathcal{C}(X, K)$).

The result on the eHI_n (of [6]) only provides weaker convergence in probability and uses the uniformity assumption. Thus we argue that mathematically it is more appealing to use the plug-in MI estimator of [20] in the purely discrete case, or to use GKOV (because the Gaussian noise assumption implies that model free MI estimation delivers the desired result).

Algorithm 1 Non-parametric $I(X;Y)$ estimation for mixed r.v.s (X,Y)[8]

Require: $\{x_i, y_i\}_{i=1}^n$ and $t_n = t$
for $i = 1, \ldots, n$ **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**
    $\tilde{d}_i = |\{j : d_{ij} = 0\}|$
  else
    $\tilde{d}_i = t$
  end if
  $n_{x,i} = |\{j : \|x_j - x_i\| \leq d_{i,xy}\}|$
  $n_{y,i} = |\{j : \|y_j - y_i\| \leq d_{i,xy}\}|$
  $\alpha_i = \psi(\tilde{d}_i) - \log(n_{x,i} + 1) - \log(n_{y,i} + 1)$
end for
return $\frac{1}{n}\sum_i \alpha_i + \log(n)$

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. [12] introduced the idea of using a k nearest neighbour (short k-NN) based estimator (also known as KSG estimator in the wider statistical literature). Recently, a generalization of the KSG estimator was proposed by Gao, Krishnan, Oh and Vishwanath [8] (that we will refer to as GKOV estimator), which is applicable to a mixture of continuous and discrete random variables. The GKOV estimator is defined as

$$I_n(X;Y) = \frac{1}{n}\sum_{i=1}^{n} \hat{I}_i = \log n + \frac{1}{n}\sum_{i=1}^{n}(\psi(\tilde{k}_i) - \log(n_{x,i} + 1) - \log(n_{y,i} + 1)). \quad (6)$$

Here, $\psi(u)$ is the digamma function $\psi(u) = \frac{d}{du}\ln\Gamma(u) \approx \ln u - \frac{1}{2u}$. The details of how to compute the quantities $n_{x,i}, n_{y,i}$ and $\tilde{t}_i$ can be found in algorithm 1. An *important feature* of GKOV estimator at least for the purpose of side channel analysis, is that the random variables involved in MI can be mixed e.g. one discrete and the other one continuous.

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.

3 Defining the MI for the Ideal Adversary

Recall that an adversary can observe inputs $x \in X$ and traces $t \in T$. The ideal adversary would possess access to a predictive leakage model that is identical to the device leakage model L. They would then use this model, applied to an intermediate step $\mathcal{C}$ of the algorithm, giving rise to predictions $L(\mathcal{C}(X,K))$ in a concrete attack vector. In an evaluation we wish to capture this adversary by computing a suitable MI metric.

3.1 Characterising the ideal adversary

The mutual information between the observed traces and the pair (input, key) measures how much information about the key is contained in the observable trace. We call this I^{k}:

$$I^{\mathsf{k}} = I((X, K); T). \tag{7}$$

The cryptographic target function $\mathcal{C}$ maps the key and input value to an intermediate value $Y = \mathcal{C}(X, K)$. Consequently, because of the data processing inequality we know that

$$I^{\mathsf{k}} \leq I^{c} = I(Y, T). \tag{8}$$

Equality holds if and only if $\mathcal{C}$ is one-to-one. In an attack, the ideal adversary would have access to the true device leakage model. If the device leakage function is statistically independent of $\mathcal{C}$ then we can utilise the data processing inequality again and find:

$$I^{\mathsf{k}} \leq I^{c} \leq I^{b} = I(L(Y), T). \tag{9}$$

Consequently, the mutual information I^{b} is an upper bound to the other MI quantities of interest. From the data processing inequality it would seem that only if the device leakage was bijective, equality could be achieved. However, in Sect.4 we will charactise the precise condition under which $I^{b} = I^{\mathsf{k}}$, which is looser than requiring the device leakage to be bijective.

The situation where $I^{b} = I^{\mathsf{k}}$ is the key comparing two (or more) estimated leakage models: the closer (in absolute terms) an estimated model is to I^{k} the better it captures the leakage characteristics of a device.

3.2 The curse of discretisation

The existing metrics in the side channel community (the eHI and ePI) only have guarantees if T is a discrete random variable. However, as we argued before, this assumption cannot be applied in general to observations from side channels such as power and EM, and it becomes invalid as soon as de-noising and other trace processing methods are used. We now study the effect of discretisation on the MI that characterises the ideal adversary.

Discretization divides the range of a continuous random variable X into possibly an infinite number of intervals. Drawing on [21, 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 [21] 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 [3]. 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.

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 [21, 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 [21, cf. Proposition 1] for the definition. It is shown in [21] 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. □

With this proposition, it follows immediately that if we consider an adversary that discretises traces, they loose information:

$$I^b \geq I^d = I(L(Y), [T]). \tag{10}$$

4 Characterising the best MI

We wish to reason about the MI that characterises the ideal adversary, but this requires knowledge of the true device leakage L, which is unknown. In this section we show that if $\mathcal{C}$ is bijective, then, $I^b = I^{\mathsf{k}}$ assuming some mild conditions on R. This equality implies that in many practical cases I^b can be estimated via I^{k} and thus it I^b can be established without the need to know or even estimate L. The result of this section is purely theoretical, but there exists a practical estimator [8], which we explain and analyse in the side channel setting, in Sect. 5.

In the following proofs we distinguish between the two cases where:

$$\begin{gathered} L(\mathcal{C}(X,K)) \text{ is a discrete function of } (X,K), \text{ or} \\ L(S, \mathcal{C}(X,K)) \text{ is a continuous function of } (X,K). \end{gathered}$$

The entropy for the ideal adversary $I^b = I(T;L) = H(T) - H(T|L)$ and the maximum entropy $I^{\mathsf{k}} = I(T;(X,K)) = H(T) - H(T|(X,K))$ only differ in the conditional entropy term. Consequently, our overall argument will be based on establishing under which conditions these two terms are equal, which requires reasoning about the conditional distributions. A basic assumption in this section is thus that the conditional entropy exists.

[3] 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.[6] for the eHI.

4.1 Characterising the Conditional Distributions

Let L be discrete. We first study the conditional distribution $T|L$. It is easy to see that this conditional distribution is completely defined by the distribution of R:

$$\begin{aligned}\mathbb{P}(T \leq t|L = l) &= \mathbb{P}(L + R \leq t|L = l) &(11)\\ &= \mathbb{P}(l + R \leq t) = \mathbb{P}(R \leq t - l) &(12)\\ &= F_R(t - l)\end{aligned}$$

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

We now consider the second conditional distribution $T|(X, K)$. Recall that we assume that $Y = \mathcal{C}(X, K)$ is one-to-one. Then the conditional distribution $T|(X, K)$ is also equal to the distribution of R:

$$\begin{aligned}\mathbb{P}(T \leq t|Y = y) &= \mathbb{P}(L(Y) + R \leq t|Y = y) &(13)\\ &= \mathbb{P}((L(y) + R \leq t) = \mathbb{P}(R \leq t - L(y)) &(14)\\ &= F_R(t - L(y))\end{aligned}$$

It follows again that the pdf of $T|(X, K)$ is given by the pdf of R. This observation has been formalised before in [22, Corollary 3.].

Let L be continuous. The continuity of L is due to some randomness S that depends on X, K and the target function $\mathcal{C}$ but importantly we still have the independence between L and R. To derive the distribution of $T|L$ (and then $T|Y$) we need a little bit more machinery than before because L is continuous (this case is not covered by [22, 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". 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 $|J| \cdot f_{(X_1,X_2)}$. The value $|J|$ is the absolute value of the Jacobian $J = \left|\frac{\partial(v_1,v_2)}{\partial(x_1,x_2)}\right|$. 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 $T|L$. Hence we apply the change of variables technique to derive the distribution of $L + R, L$, and choose $Y_1 = L + R$, $Y_2 = L$. Hence $|J| = 1$, and this gives $f_{L+R,L} = 1 \cdot f_{L,R} = 1 \cdot f_L(l) \cdot f_R(r) = f_L(l) \cdot f_R(t-l)$. Cleary the pdf of the marginal distribution f_{L+R} is then given as $f_R(t-l)$, which implies that $f_{T|L=l} = f_R(t - l)$.

We then derive the pdf of $T|(Y, S)$ by using exactly the same trick, and this gives us $f_R(t - L(y, s))$.

4.2 Equality of I^b and I^{k} if $H(R)$ is location independent

Intuitively, it should be expected from the results in the previous subsection that $I^{\mathsf{k}} = I^b$ (for $\mathcal{C}$ one-to-one) if the conditional entropy terms satisfy some conditions. We formalise the conditions in Prop. 2.

Proposition 2. *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 let $Y = \mathcal{C}(X, K)$, and $\mathcal{C}$ be one-to-one. If $H(R) = \varphi(\sigma)$ where φ depends only on f_R, then $I^{\mathsf{k}} = I^b$.*

Proof. We consider the two cases, for a discrete L and a continuous L separately. *Case 1:* Let L be discrete.

First, we derive I^b:

$$\begin{aligned} I^b = I(T, L) &= H(T) - H(T|L) \\ &= H(T) - \sum_l p_L(l) H(T|L = l) \\ &= H(T) - \sum_l p_L(l) \mathbb{E}_{T|l}(-\log(f_{T|l}(t|l)) \\ &= H(T) - \sum_l p_L(l) \mathbb{E}_R(-\log(f_R(t - l))) \end{aligned}$$

The equalities in the first line all follow from standard definitions. The equality in the second line follows from the well known fact that the entropy is the expected value of the logarithm of the resp. distribution. The equality in the third line follows from the characterisation of the conditional distribution from the previous section.

Second, we derive I^{k} using $Y = \mathcal{C}(X, K)$ (and $\mathcal{C}$ being one-to-one):

$$\begin{aligned} I^{\mathsf{k}} = I(T, Y) &= H(T) - H(T|Y) \\ &= H(T) - \sum_y p_Y(y) H(T|Y = y) \\ &= H(T) - \sum_y p_Y(y) \mathbb{E}_{T|y}(-\log(f_{T|y}(t|y)) \\ &= H(T) - \sum_y p_Y(y) \mathbb{E}_R(-\log(f_R(t - L(y))) \end{aligned}$$

We can see that $I^b = I^{\mathsf{k}}$ iff $\mathbb{E}_R(-\log(f_R(t - l))) = \mathbb{E}_R(-\log(f_R(t - L(y)))$. This is the case when the entropy of R does not depend on its location, i.e. iff $H(R) = \phi(\sigma)$ (the entropy is only a function of the spread σ, but not the location μ).

Case 2: Let L be continuous.
First, we derive I^b:

$$\begin{aligned} I^b = I(T,L) &= H(T) - H(T|L) \\ &= H(T) - \int_l f_L(l) H(T|L=l) dl \\ &= H(T) - \int_l f_L(l) \mathbb{E}_{T|l}(-\log(f_{T|l}(t,l))) dl \\ &= H(T) - \int_l f_L(l) \mathbb{E}_R(-\log(f_R(t-l))) dl \end{aligned}$$

The reasoning for the equalities in the first two lines is identical to the reasoning in the discrete case. The equality in the third line is again based on the characterisation of the conditional distribution that we developed in the previous subsection.

Second, we derive I^{k} :

$$\begin{aligned} I^{\mathsf{k}} = I(T,(Y,S)) &= H(T) - H(T|(Y,S)) \\ &= H(T) - \sum_y p_Y(y) \int_s \mathbb{E}_{T|(y,s)}(-\log(f_{T|(y,s)}(t,(y,s)))) ds \\ &= H(T) - \sum_y p_Y(y) \int_s \mathbb{E}_R(-\log(f_R(t-L(y,s)))) ds \end{aligned}$$

The various equalities are all based on the same arguments as in the previous cases. To take the continuity of L into account, we must also run over the randomness S in our argument. This does not change the final step however, which is to observe that iff the differential entropy $\mathbb{E}_R(-\log(f_R(t-l)))$ equals $\mathbb{E}_R(-\log(f_R(t-L(y,s))))$ then we have that $I^{\mathsf{k}} = I^b$. The equality between the differential entropies will hold if the entropy of the distribution does not depend on its location, but only its spread: i.e. iff $H(R) = \phi(\sigma)$.

□

Note that the r.v. S in Proposition 2 was introduced to define the output distribution of L. In practice it is implicit to a device and is not required to estimate it explicitly in the process of leakage certification (or MI estimation). Proposition 2 can easily be extended for a distribution of R characterized with only scaling parameter $\sigma > 0$.

Theorem 1 (Estimation of I^b via I^{k}). *If the entropy of the noise distribution (of a device) is location independent then $I^b = I(T; L \circ \mathcal{C}(X,K))$ can be computed via $I^{\mathsf{k}} = I(T;(X,K))$.*

We note that the noise condition covers all distributions that so far have been observed in practice or that have been assumed in comprehensive studies such as [2]. In particular, the noise condition applies to distributions such as

the Normal distribution, Laplace distribution, Cauchy distribution etc. Subsequently, we provide two specific applications of our theorem to the most widely used noise assumptions in the side channel literature, and we then explain that our result also applies in the multivariate setting.

4.3 Multivariate Analysis

No assumptions were necessary in the previous section regarding the dimensionality of T, thus our analysis naturally applies also to situations where multiple trace points are considered jointly. Working with joint distributions is also of interest when considering countermeasures, such as masking, where intermediate values are split up into multiple shares.

Consider a simple two share setting, where for an intermediate value we just have one additional random variable, denoted by M. Because the mask is part of the target function, and is chosen independently of all other quantities, it does not change our analysis from before.

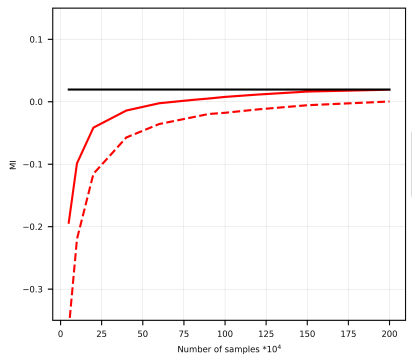
5 Practical MI Estimation Using the GKOV Method

The recently proposed GKOV estimator [8] 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).

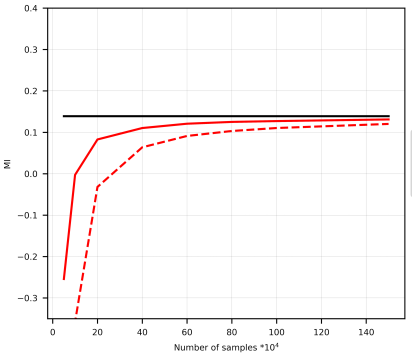
5.1 Fast implementation of Alg.1

For our `C++` implementation of MI estimation, 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



(a) HW leakage, with $\mathcal{L}$ noise



(b) Nonlinear leakage with $\mathcal{N}$ noise

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

Fig. 1: Convergence experiments for different choices of t_n.

for the same. For example, with the dimension $m = 2$, computing the ℓ_∞ norm as

```
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.

5.2 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 [8]).

In our experiments we selected t_n equal to $\log n$ and $\log_{10}^2 n$. Figure 1 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. 1 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.

A final observation is that the GKOV estimator approaches the true MI from below. There is no formal proof for this in [8], 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 any bias, and recall that [8] shows the asymptotic unbiasedness of their estimator.

5.3 Multivariate setting

The computational cost for estimating the mutual information in a multivariate setting using a histogram method (pdf estimation method) is high and a known problem. 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.

In contrast the estimator by Gao et al.[8] 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.

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

5.4 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 with some practical experiments. The first experiment is based on simulated traces, which are generated based on non-linear device leakage with Laplacian noise. The second experiment is similar, but we generate the traces based on a linear device leakage model and use Gaussian noise. For each experiment, we estimate the MI using GKOV for both the traces as they are generated, and for the traces after discretisation. Figures 2 and 3 show the outcomes: in both cases the MI between the discretised traces and the key is smaller than the MI between the traces and the key: $I([T];(X,K)) < I(T;(X,K))$.

6 Experiments: MI Estimation considering One Discrete and One Continuous Random Variable

We now examine, in a range of practical experiments, the behaviour of the GKOV estimator in comparison to the behaviour of the eHI and ePI. 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. 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.

6.1 Simulation setup

In all experiments we consider a single bijective target function, which is the AES SubBytes mapping, $y = \mathcal{C}(c,k) = SubBytes(x \oplus k)$. In our simulations, we vary the device leakage model 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: $L = \mathrm{HD}(Y, \mathcal{C}^{-1}(Y))$ (Hamming distance between Y and $\mathcal{C}^(Y)$),
non-linear: $L = \text{DES-Sbox}\,(6\mathrm{LSB}(Y))$ (The first DES Sbox applied to the 6 least significant bits of Y), and
wHW: $L = \sum_i wt_i \cdot Y_i$ (Weighted Hamming weight: a weighted linear function of the bits of Y).

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

In the multivariate simulations the simulated trace points are either based on either HW or HD leakage of some bits of Y. For instance, the bivariate simulations are based on $(HW(4LSB(Y), HW(4MSB(Y))$ or

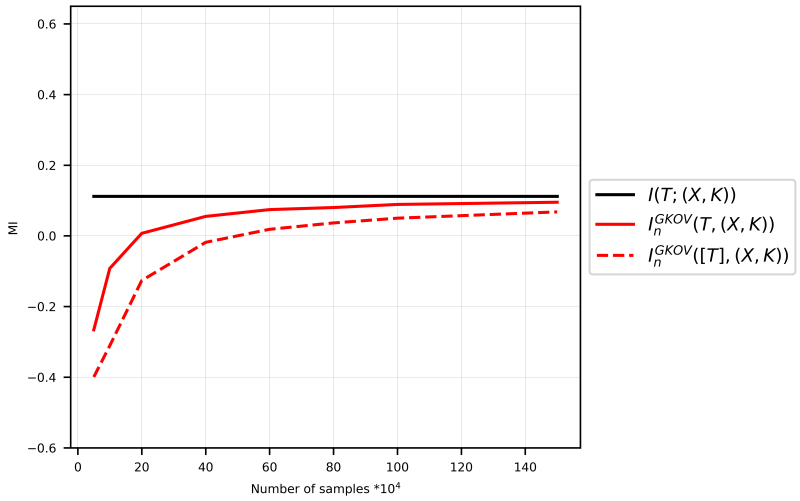


Fig. 2: MI, nonlinear leakage, with Laplacian noise

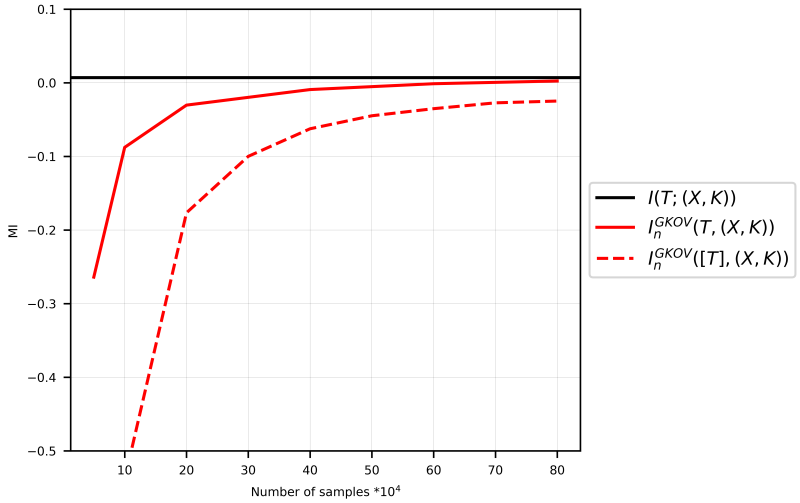


Fig. 3: MI, linear leakage, with Gaussian noise

$HW(4LSB(Y), HD(4MSB(Y))$ and the independent noise (R_1, R_2) is considered from the set $(\mathcal{N}, \mathsf{Lap})$ by taking $\sigma = 4$.

In order to compute the eHI, and the ePI, with use the scripts that were provided by the authors of [6]. 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 [6] include a step where traces are discretised.

We are able to compute the exact MI because we know all distributions, and always include the exact MI as a black reference line in the plots.

Experimental results. We ran a large number of experiments that combine different device leakage functions and different noise distributions in a univariate and various multivariate settings. They all produced the same conclusions, and thus we include a subset of experiments in Figures 4-7 that are representative of the outcomes.

The experiments clearly demonstrate that the GKOV estimator quickly converges to the true mutual information value, irrespective of the dimensionality of the leakage. In stark contrast, the eHI, as explained in the previous work is a biased upper bound, and the bias increases dramatically with the number of dimensions, which is in line with the follow up to Bronchain et al. in [23]. We were unable to run ePI, eHI and the histogram based estimator for four shares — their requirement to explicitely build a multivariate pmf 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 [20]. We can see that its performance is particularly bad, and it also appears to show bias (which is expected given Paninski [10]).

7 Leakage Certification

The question that inspired much work on MI estimation in the side channel community was the question how can an evaluator know, if a leakage model that they intend to use in an attack is a good model? And consequently, given two models, how do they compare?

We now revisit this question in a set of controlled experiments. For this we define leakage models L' in relation to some "true device leakage" L that incorporate progressively less information of L. We achieve this by loosing some bits of the intermediate value, and then we apply the same leakage function. Precisely, we consider the following models:

- the model $6LSB$ is based on using just the six least significant bits of Y,
- the models $4LSB$ and $4MSB$ are based on using the four least or most significant bits of Y.

The intermediate Y is the output of the AES SubBytes operation, thus it has 8 bits. Clearly, the $6LSB$ model should be a better predictor than the $4LSB$ or the $4MSB$ model.

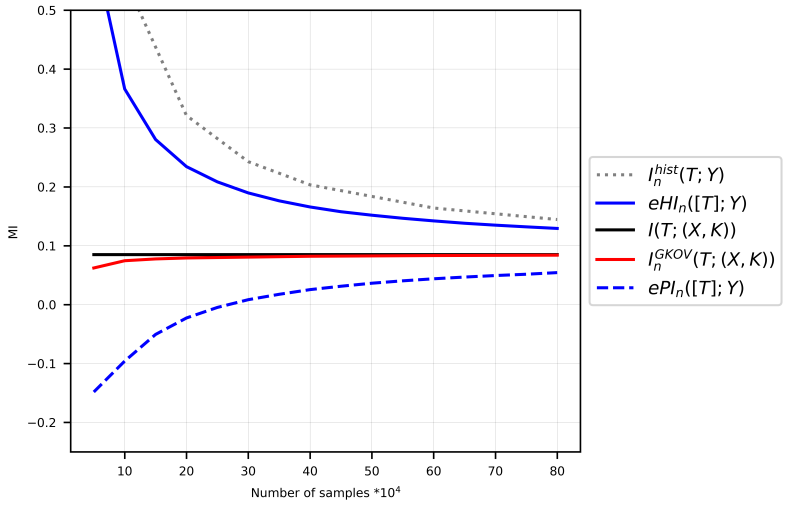


Fig. 4: Estimator behaviour, HD leakage, with Gaussian noise

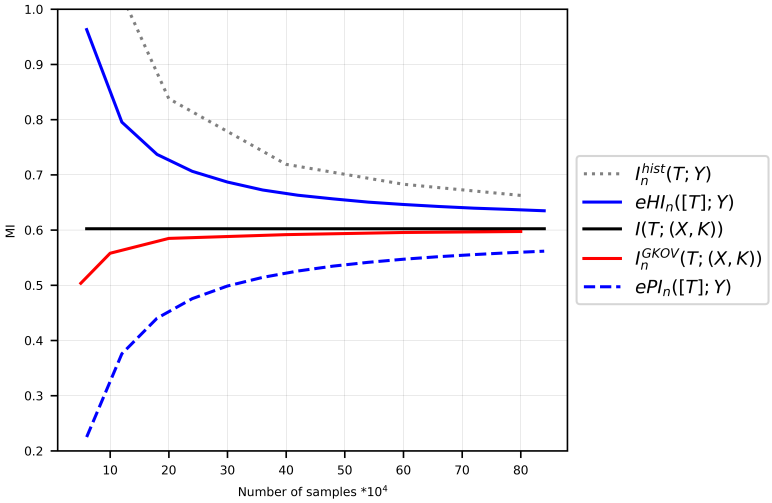


Fig. 5: Estimator behaviour, Non-linear leakage, with Gaussian noise

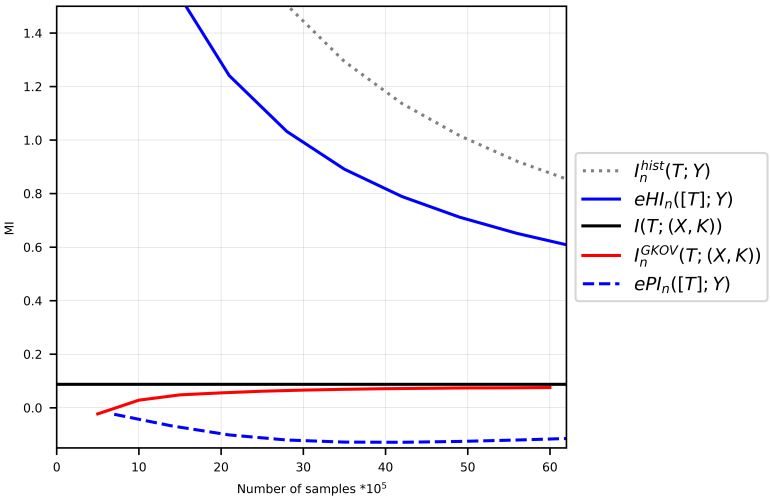


Fig. 6: Estimator behaviour, (HW, HD) leakage, with Gaussian noise

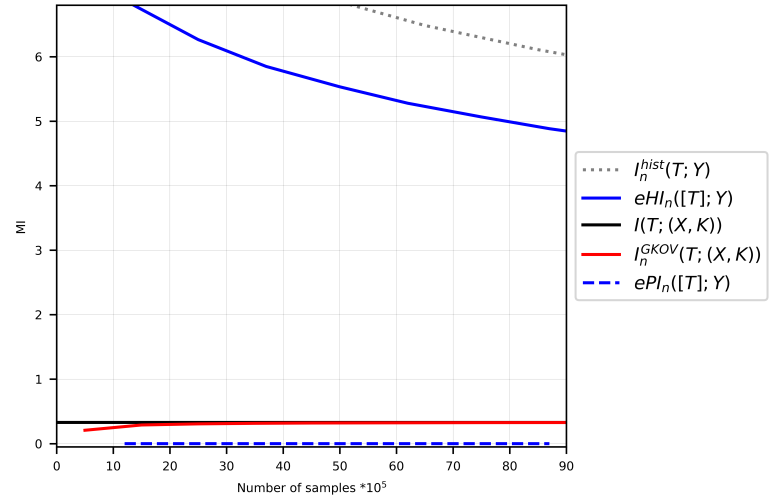


Fig. 7: Estimator behaviour, (HW, HW, HD) leakage, with Gaussian noise

7.1 Leakage certification using eHI and ePI.

We recall briefly, that the eHI metric was proven to be an upper bound for $I(T, Y)$. The ePI metric was designed to capture the relationship between a leakage model L' and the true device leakage model L. The idea is then that an evaluator can judge the quality of L' by computing the eHI and then relating this to the ePI (which depends on L'). A better model should should have an ePI that is closer to the eHI. Thus, we would expect that the ePI$(T, 6LSB)$ is closer to the $eHI(T, Y)$ than the $ePI(T, 4LSB)$ or the $ePI(T, 4MSB)$.

In the papers that originally define and discuss the eHI and ePI, experiments were provided that confirm that the eHI is always larger than the ePI and that the true MI sits somewhere in the middle. But these experiments were based on univariate discrete leakage. We want to challenge the eHI and ePI estimators in a multivariate setting, and thus look at leakage certification outcomes when considering two bivariate scenarios: one in which both points leak the HW and one where one point leaks the HW and one point leaks the HD (we provided the detailed description in Sect. 6.1 before).

With this in mind we examine the outcomes of our first bi-variate simulation that is based on two points leaking the HW: this is given in Fig. 8. The ePI that is furthest away from the eHI is ePI$(T, 6LSB(Y))$ 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. 9. 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. The bias of eHI is once more noticeable.

7.2 Leakage certification using GKOV.

The idea that a better model should exploit more information implies that the mutual information between the model and the traces should be higher: given two models L' and L'', L' is a better model than L'' if $|I^{\mathsf{k}} - I(T, L')| < |I^{\mathsf{k}} - I(T, L'')|$. The GKOV estimator seems the ideal tool to estimate the respective MI quantities.

We now look at the scenarios from the previous subsections and compute the respective MI quantities between the different models using the GKOV estimator. We plot the exact MI as a black line. Figure 10 shows that the GKOV estimator approaches the exact MI as expected from the theoretical convergence proof as well as our previous experiments, confirming once more that it is a consistent estimator. Consequently $I^{\mathsf{k}} = I(L, (X, K))$ as estimated by GKOV is then the baseline for comparison. We can see in Fig. 10 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 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

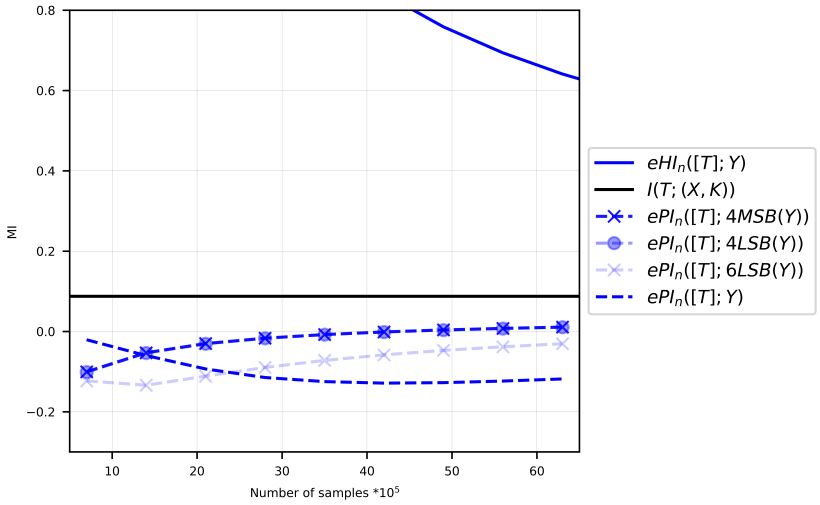


Fig. 8: Experiment: leakage certification using eHI and ePI for bi-variate (HW-HW) leakage with Gaussian noise.

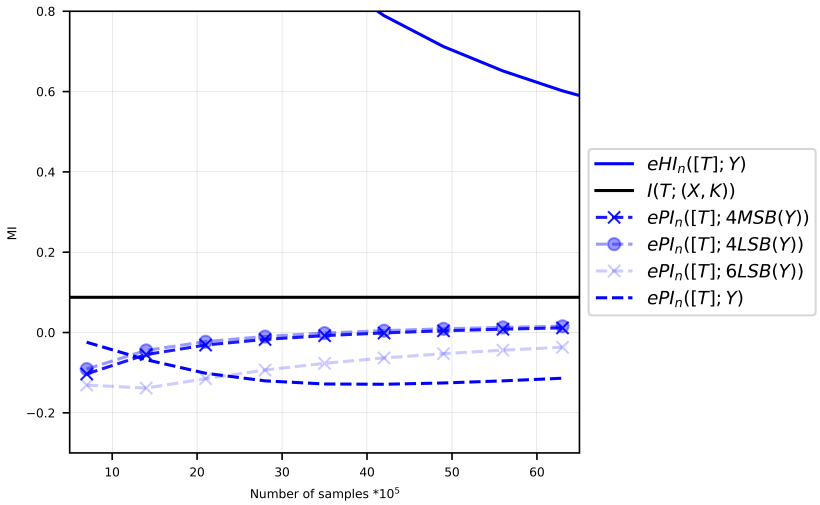


Fig. 9: Experiment: leakage certification for eHI and ePI bi-variate (HW-HD) leakage with Gaussian noise.

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.

Using real data. Finally, we use the 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.

The purpose of this experiment is to confirm with a real world dataset that we should expect ePI, and eHI to show a bias in a multivariate setting (even though it is discrete) and therefore potentially misleading outcomes. Figure 12 shows the result of this experiment. We notice that, e.g. between the sample points 100 and 150, there are a number of points where eHI outcome indicates leakage, where GKOV and ePI do not.

8 Conclusions

The estimation of the mutual information between two or more variables is required in the context of assessing practical implementations with respect to their information leakage. In the past years, progress was made in the side channel community to deal with the hazards of non-parametric MI estimation, which lead to the introduction of the notions of eHI and ePI. More progress has been made in the machine learning community, which has lead to the introduction of a non-parametric MI estimator, GKOV, that is convergent and asymptotically unbiased even when applied to mixtures.

Our paper shows that the GKOV estimator is an ideal tool in the side channel setting: we proved that the density free maximum MI between the traces and an intermediate $I(T, Y) = I(T, (X, K))$ (if $Y = C(X, K)$, and C is one-to-one) is equal to the MI that characterises the ideal adversary $I(T, L(Y))$. The ideal adversary is the adversary who knows the device leakage function L (it is an ideal adversary because such knowledge does not exist in practice). Thus the information leakage for a worst case side channel attack can be estimated via $I(T, (X, K))$ in theory, and using the GKOV estimator in practice.

Our main results also show that the bias of eHI and ePI increases quickly when using them in a multivariate setting. This is in addition to the fact that the computational effort to compute the eHI and ePI increases exponetially and therefore at present any form of higher multivariate analysis is computationally infeasible. We challenge the eHI and ePI also in the leakage certification setting and find that comparing models using GKOV is advantageous.

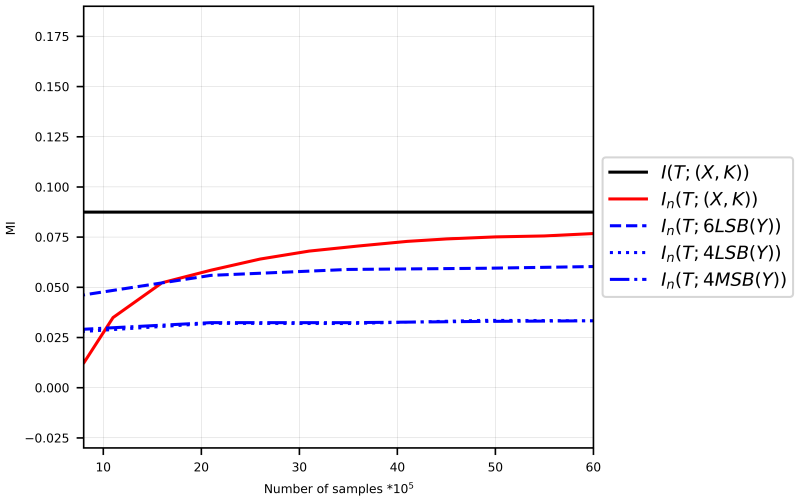


Fig. 10: Experiment: leakage certification for bi-variate (HW-HW) leakage with Gaussian noise.

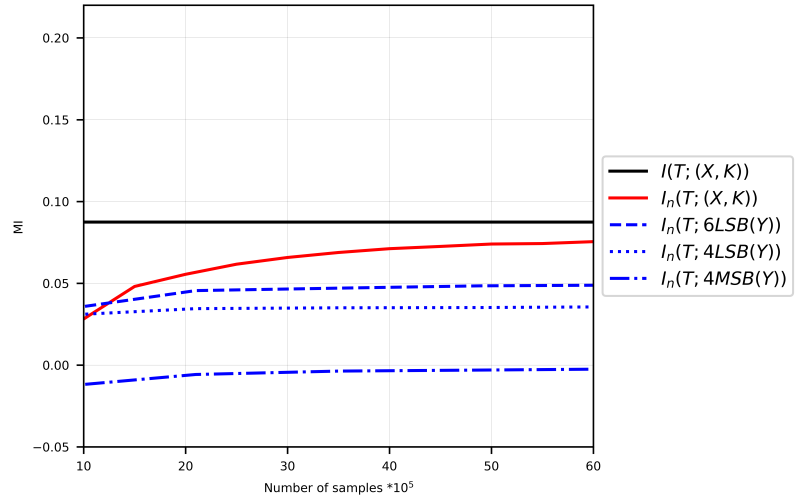


Fig. 11: Experiment: leakage certification for bi-variate (HW-HD) leakage with Gaussian noise.

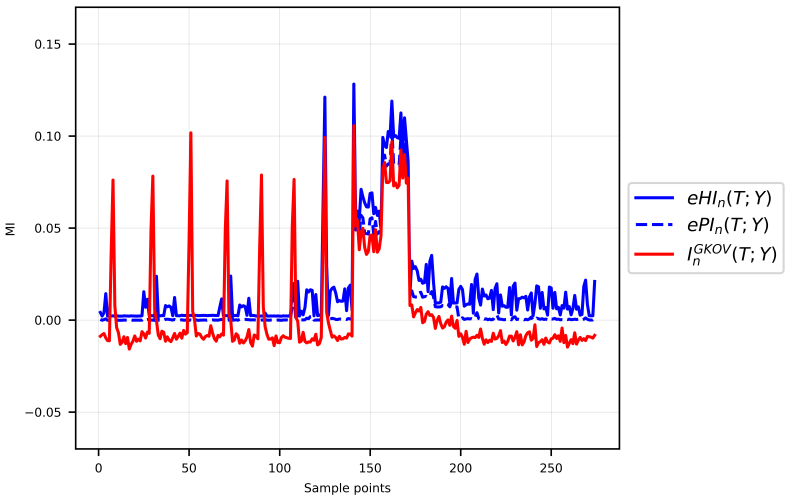


Fig. 12: Experiment: comparison of estimators using bi-variate discrete real device leakage

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. 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
6. 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
7. Prouff, E., Rivain, M., Bevan, R.: Statistical analysis of second order differential power analysis. IEEE Trans. Computers **58**(6) (2009) 799–811
8. 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
9. Nair, C., Prabhakar, B., Shah, D.: On entropy for mixtures of discrete and continuous variables. arXiv preprint cs/0607075 (2006)
10. Paninski, L.: Estimation of Entropy and Mutual Information. Neural Computation **15**(6) (2003) 1191–1253
11. L. F. Kozachenko, N.N.L.: Sample estimate of the entropy of a random vector. Problems in Information Transmission **23** (1987)
12. Kraskov, A., Stögbauer, H., Grassberger, P.: Estimating mutual information. Phys Rev E Stat Nonlin Soft Matter Phys **69** (07 2004) pp. 066138
13. 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.
14. 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
15. 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
16. 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
17. 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
18. 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
19. Hall, P., Morton, S.: On the estimation of entropy. Annals of the Institute of Statistical Mathematics **45** (02 1993) 69–88
20. Antos, A., Kontoyiannis, I.: Convergence properties of functional estimates for discrete distributions. Random Structures & Algorithms **19** (10 2001) 163 – 193
21. 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
22. 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
23. 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`.