On the Success Rate of Side-Channel Attacks on Masked Implementations

Information-Theoretical Bounds and Their Practical Usage

Akira Ito
NTT Corporation
Tokyo, Japan
akira.ito.as@hco.ntt.co.jp

Rei Ueno
Tohoku University
Sendai, Japan
rei.ueno.a8@tohoku.ac.jp

Naofumi Homma
Tohoku University
Sendai, Japan
naofumi.homma.c8@tohoku.ac.jp

ABSTRACT

This study derives information-theoretical bounds of the success rate (SR) of side-channel attacks on masked implementations. We first develop a communication channel model representing side-channel attacks on masked implementations. We then derive two SR bounds based on the conditional probability distribution and mutual information of shares. The basic idea is to evaluate the upper-bound of the mutual information between the non-masked secret value and the side-channel trace by the conditional probability distribution of shares given its leakage, with a help of the Walsh–Hadamard transform. With the derived theorems, we also prove that the security of masking schemes: the SR decreases exponentially with an increase in the number of masking shares, under a much more relaxed condition compared with the previous proof. To validate and utilize our theorems in practice, we propose a deep-learning-based profiling method for estimating the conditional probability distribution of shares to estimate the SR bound and the number of traces required for attacking a given device. We experimentally confirm that our bounds are much tighter than the conventional bounds on masked implementations, which validates the relevance of our theorems to practice.

KEYWORDS

masking; power/EM analysis; side-channel attack

1 INTRODUCTION

1.1 Background

Boolean masking is one of the major countermeasures against side-channel attacks on symmetric ciphers. In a masking scheme, an n-bit secret variable Z is represented by the sum of d random numbers over $\mathbb{F}_2^n$ as $Z = S_1 \oplus S_2 \oplus \cdots \oplus S_i \oplus \cdots \oplus S_d$, where S_i is called share. The number of masking shares d is an essential parameter which determines the security and implementation cost. If the masking is correctly and soundly implemented, masking with d shares is secure against $(d-1)$-th-order side-channel attacks. This can be intuitively explained by the $(d-1)$-th order probing model—any $(d-1)$-th probing attacker cannot recover the secret variable Z from the information of at most $(d-1)$ shares [4, 16, 43]. Many studies have been devoted for the implementation and verification of masking schemes for a given masking order $d-1$ (*e.g.*, [2, 3, 5, 22, 26, 27, 33, 39, 43, 51, 52]); and $(d-1)$-th order masking (*i.e.*, masking with d shares in this paper) which is considered sufficiently secure and practical has been commonly selected and implemented nowadays.

However, it is natural to assume that attackers would attempt any d-th order attack on the $(d-1)$-th order masked implementation. This implies that attention should be devoted towards investigating how much more difficult a d-th order attack is than the corresponding $(d-1)$-th order attack. It is considered that the number of traces required for successful attacks (*i.e.*, the attack cost) and the implementation cost increase exponentially and quadratically by d, respectively [12]; hence, the implementer should determine a minimum d such that the attack cost is greater than the theoretical and/or practical attack limits. This, in turn, leads to a high demand for estimating the cost of a d-th order attack from an implemented device and its conditions (*e.g.*, the signal-to-noise ratio (SNR) of the side-channel measurement). The attack cost can be evaluated by the upper-bound of the success rate (SR) for a given number of traces, which is equivalent to the lower-bound of the number of traces to achieve an SR [49]. A pioneering study by Duc *et al.* has derived the relation between the masking order/number of masking shares and the SR upper-bound [20]. However, its precision and tightness should be improved for a practical use. In TCHES 2019 [13], treating a side-channel attack as a communication channel, de Chérisey *et al.* showed an information-theoretical SR upper-bound for non-masked implementations that can be calculated using mutual information between a secret variable Z and side-channel trace $\boldsymbol{X}$, denoted by $I(Z;\boldsymbol{X})$. Their information-theoretic approach is promising because they experimentally showed that their bounds are much tighter than those in previous studies. However, they are only applicable to non-masked implementations and are unavailable for estimating the security of masked implementations during the design. As is described in Section 1.3, other previous studies also have some difficulties and limitations (*e.g.*, generality, precision, and tightness) in their practical use for cases of masked implementations.

1.2 Contributions

Information-theoretical SR upper-bounds. This study presents a new information-theoretical approach for deriving SR upper-bounds of side-channel attacks on masked implementations. We first extend the communication channel model of [13] to side-channel attacks on masked implementations, and then derive two SR upper-bounds: Theorem 4.5 and Theorem 4.7. Theorem 4.5 shows that an SR bound derived from the conditional probability distribution of the i-th share S_i given its side-channel leakage $\boldsymbol{L}_i$ (*i.e.*, $p_{S_i|\boldsymbol{L}_i}$), which assumes that the conditional probability distribution $p_{S_i|\boldsymbol{L}_i}$ is known in some manner. Theorem 4.7 shows another SR bound derived from the mutual information between share and its leakage

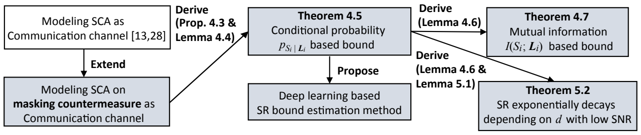

Modeling SCA as Communication channel [13,28]
Extend
Modeling SCA on masking countermeasure as Communication channel
Derive (Prop. 4.3 & Lemma 4.4)
Theorem 4.5
Conditional probability
$p_{S_i \mid \boldsymbol{L}_i}$ based bound
Propose
Deep learning based SR bound estimation method
Derive (Lemma 4.6)
Theorem 4.7
Mutual information
$I(S_i; \boldsymbol{L}_i)$ based bound
Derive (Lemma 4.6 & Lemma 5.1)
Theorem 5.2
SR exponentially decays depending on d with low SNR

Figure 1: Overview of our contributions.

$I(S_i; \boldsymbol{L}_i)$. The SR upper-bound of Theorem 4.5 provides a tighter value than that of Theorem 4.7 because $p_{S_i|\boldsymbol{L}_i}$ is far more informative than $I(S_i; \boldsymbol{L}_i)$. By contrast, the SR upper-bound of Theorem 4.7 is a natural extension of the previous study [13] and can be evaluated more easily than that of Theorem 4.5. The ease of analysis comes from the fact that the mutual information $I(S_i; \boldsymbol{L}_i)$ can be estimated/approximated by the SNR of side-channel measurement (*i.e.*, $I(S_i; \boldsymbol{L}_i) \leq \log(1 + \mathrm{SNR})/2$). Given a target device to be profiled, the evaluation of Theorem 4.7 is performed only by the SNR of the side-channel measurement, whereas that of Theorem 4.5 is required to estimate $p_{S_i|\boldsymbol{L}_i}$ experimentally with a higher computational cost, as described below. Note that these theorems bound the SRs of attacks with an optimal distinguisher that maximizes the SR [8, 29] (the SR of any other attack is lower than that of the attack with the optimal distinguisher). This indicates that the bounds are valid for any type of attack that exploits the leakage(s) of an intermediate value (*i.e.*, the output of a selection function).

Theoretical analyses on security of masking schemes. On the basis of Theorem 4.5, we provide a concrete proof of security of masking schemes as Theorem 5.2: the SR decreases exponentially with an increase in d, and SR converges to $1/2^n$ as $d \to \infty$, where n is the bit length of the target intermediate value (or the bit length of an attacked partial key). This proof also indicates that the number of traces required for an attack success increases exponentially with d. Although Duc *et al.* has already proved this statement in [20] using a noisy leakage model, their proof is valid only if $I(S_i; \boldsymbol{L}_i) \leq 2^{-2n+1}$. This indicates that, for example, in the case of AES (*i.e.*, $n = 8$), the proof is valid only if $I(S_i; \boldsymbol{L}_i) < 2^{-15} \approx 3.05 \times 10^{-5}$. Suppose that the mutual information $I(S_i; \boldsymbol{L}_i)$ can be bounded by the SNR as $I(S_i; \boldsymbol{L}_i) \leq \log(1+\mathrm{SNR})/2$, as discussed and experimentally evaluated in [13]. Their proof makes sense only if the SNR is less than 6.1×10^{-5}; the SNR value appears too low to consider practical side-channel attacks[1]. Thus, in the existing study [20], the security of masking schemes (here, such an exponential property) was unclear under practical conditions in terms of SNR. On the other hand, our proof provided in this paper is valid if $I(S_i; \boldsymbol{L}_i) < 1/(2\ln(2)) \approx 0.72$, independently of n. Our proof first validates the masking security in general and practical cases that were previously unknown. Moreover, we also analyze and discuss the convergence of $\mathrm{SR} \to 1/2^n$ as $d \to \infty$ to further investigate the security of masking schemes.

Accurate SR estimation. This paper also shows that Theorem 4.5 can be used for a practically (at least non-trivially) tight evaluation of the SR (or the number of traces) of the masked implementation with the conditional probability of a NON-masked implementation. More precisely, to utilize Theorem 4.5 for practical SR evaluation, this paper proposes a deep learning (DL)-based method for estimating (*i.e.*, profiling) the conditional probability $p_{S_i|\boldsymbol{L}_i}$ from a non-masked implementation on the same device as the masked ones. A combination of Theorem 4.5 and DL-based estimation enables the evaluation of the upper-bound of SR and the lower-bound of the number of traces for attack success more tightly and precisely than ever before.

Figure 1 illustrates the overview of the above contributions. The practicality and effectiveness of proposed bounds/method are demonstrated in Section 7 through an experimental attack on AES in comparison with an actual optimal attack [8, 29] in addition to a numerical evaluation.

1.3 Related studies

Analysis based on specific attacks. In [9], Chari *et al.* first showed that the measurement complexity of a single-bit differential power analysis (DPA) on the masked implementations increases exponentially with d. However, for this proof, they assumed that the measurement noise is Gaussian distributed and the leakages of all the shares are sufficiently noisy. As this study was the first report on the security proof of masked implementations to the best of authors' knowledge, their study has been followed by many researchers in order to, for example, generalize/extend their analysis to other attacks and relax the assumption about noise.

In [35], Mangard *et al.* derived an analytical relation between SR and the number of attack traces as an equation/identity. In [57], Zhang *et al.* presented a function named Cross Entropy Ratio (CER) which could be used as a loss function in a deep-learning based side-channel analysis (DL-SCA). Zhang *et al.* proved that DL-SCA can always successfully recover the secret key if an infinite number of attack traces are available and $\mathrm{CER} < 1$. Their formulation/derivation assumed that the intermediate values corresponding to a correct key and wrong keys are independent of each other. However, some counterexamples were found as in [31, 50], meaning that the assumption does not generally hold.

In [55], Zaid *et al.* presented another loss function for DL-SCA named Ranking loss, and showed that Ranking loss is a lower-bound of SR. Ranking loss can be computed only empirically, but not analytically, and the computation of ranking loss requires the conventional experimental/empirical evaluation of SR [32]. In this sense, there are few merits of using ranking loss instead of other conventional experimental evaluations.

Analysis based on central limit theorem (CLT). This type of analysis utilizes the fact that the distribution of a score function can be approximated as a normal distribution based on the CLT if the distinguisher in the side-channel attack is additive [34, 44, 56, 58]. In addition, the result in [45] is generalized to some non-additive distinguishers in [28].

Although the above approaches provide a precise approximation of SR (when the SNR is low), the major drawback is the lack of generality because their approach depends on a specific distinguisher and/or leakage model (power model). In addition, the error decay of the CLT-based approximation is slower than linear to m

[1] In [13], de Chérisey *et al.* proved that more than 10,000 traces are required for a reliable attack success on non-protected AES implementations if SNR is worse than 10^{-4}. However, many previous studies have reported that attack on non-protected AES implementations have been successful within 10,000 traces. In particular, in [7], Bronchain and Standeart reported that the noise level of some low-end off-the-shelf devices can be too low for a secure masking.

(the number of traces), and their results are valid only for a sufficiently large number of traces because of the characteristics of CLT. In addition, it should be noted that these analyses are asymptotic approximations of SR but not upper-bound. Although CLT-based approximations can be used to evaluate the SR bound on the basis of the Berry–Esseen theorem, the Berry–Esseen theorem incurs an approximation error of order $1/\sqrt{m}$ in the derived bound [54]. Therefore, CLT-based bound could be relatively inadequate in the context of SR evaluation as in the large deviation principle [11], although CLT-based approximation could achieve some meaningful results.

Analysis based on noisy leakage model. The noisy leakage model is an attacker model in which the attacker can obtain the value of the wires in a target circuit[2] with "noise" [19, 41]. In [20], Duc *et al.* modeled the "noise" by a statistical distance (or total variation distance) between a conditional probability distribution of intermediate values given side-channel leakage and the occurrence probability distribution of intermediate values. Here, the noisy leakage model is reduced to a random probing model [19]. Because the SR upper-bound can be easily evaluated using the random probing model, their modelling enables an indirect evaluation of the SR upper-bound in the noisy leakage model. With this result, Duc *et al.* to further prove that SR decreased exponentially with an increase in the masking order.

However, the reduction of the noisy leakage model to the random probing model is only meaningful if the noise is significantly (and sometimes impractically) large. In [13], it was experimentally demonstrated that Duc *et al.*'s bound only gave a trivial evaluation result (*i.e.*, an attack success required at least one trace) unless the SNR of the side-channel measurement was less than 10^{-4}. In many practical attack conditions, the SNR frequently exceeds 10^{-4}; thus, it is difficult to utilize this bound in designing actual cryptographic modules for many practical applications.

Analysis based on mutual information. In [13], de Chérisey *et al.* gave an SR upper-bound using the mutual information between intermediate value and side-channel leakage by modelling side-channel attack as a communication channel. Here, de Chérisey *et al.* only focused on non-masked implementations. In [10], Cheng *et al.* improved the bounds and evaluated a first-order masked implementation with some extensions. However, the analysis is limited to first-order ($d = 2$ in this study) and is dependent on the numerical/experimental evaluation of mutual information under some assumptions on leakage. In this sense, their analysis result is experimental under specific assumptions rather than fully information-theoretic. Note that their attack model or communication channel model is different from ours (they assumed that information of two shares is leaked at one point with an additive composition).

In [37], Masure *et al.* reported that the cross-entropy (CE) loss function, which is the most major loss function in DL(-SCA), is asymptotically equivalent to the perceived information and can be used for estimating the mutual information. Masure *et al.* also showed that de Chérisey *et al.*'s bound could be estimated/evaluated using DL-SCA. However, the reported method required an actual attack (DL-SCA) on the target module to evaluate the SR; therefore, it cannot be applied to evaluate the SR/number of traces for any masking order d without experimental attacks.

Summary. The studies mentioned above have significantly contributed to the analysis of the theoretical aspects of side-channel attacks; however, there are still difficulties and/or limitations in utilizing their results to design masked cryptographic modules in practice. Among the four approaches, the mutual information based approach is probably the most promising for non-masked implementation because of its generality and practicality, but it still has been unknown to extend it to masked implementations with any order in an analytical manner. Addressing the aforementioned issues, this paper presents information-theoretical SR bounds for masked implementation in any order.

1.4 Paper organization

The rest of this paper is organized as follows: Section 2 introduces the mathematical notations. Section 3 illustrates the communication channel model of side-channel attacks on masked implementation. Section 4 derives the SR upper-bounds for a given number of traces through information-theoretical analysis. Section 5 gives a proof and analysis about the security of masking schemes. Section 6 presents the profiling method for estimating conditional probability $p_{S_i|L_i}$ from unprotected (*i.e.*, non-masked) implementation using a DL technique. Section 7 validates our theorems and bounds through an experimental simulation of attack on AES. Finally, Section 8 concludes this paper. In addition, Appendices introduce the Walsh–Hadamard transform (WHT) and the inequalities used in our analysis.

2 MATHEMATICAL NOTATIONS

In this paper, a probability distribution function is referred to as a probability distribution. Throughout this paper, a calligraphic character denotes a set, an upper-case italic character denotes a random variable, and a lower-case character denotes its element, if it is not explicitly defined otherwise. For example, a random variable on a value in a measurable set $\mathcal{X}$ is denoted by X, and the element of $\mathcal{X}$ is denoted by x. Let Pr denote a probability measure. For example, the probability of a discrete random variable X taking x is written as $\Pr(X = x)$. Let p denote the probability density or mass function. For example, a joint probability of two random variables can be written as $p_{X,Y}(x, y)$.[3] For two random variables X and Y, the conditional probability distribution of $Y = y$ given $X = x$ is defined as $p_{Y|X}(y \mid x) = p_{Y,X}(y, x)/p_X(x)$ if $p_X(x) \neq 0$; otherwise, $p_{Y|X}(y \mid x) = 0$. Let $\mathbb{E}$ denote the expectation operator, and $\mathbb{E}[Y \mid X]$ denote the conditional expectation of Y given X. The entropy of random variable X is defined as $H(X) = -\mathbb{E} \log p_X(X)$. The conditional entropy of Y given X is defined as $H(Y \mid X) = -\mathbb{E} \log p_{Y|X}(Y \mid X)$. The mutual information between X and Y is defined as $I(X; Y) = H(X) - H(X \mid Y)$. We also assume that the logarithms of all the probability distributions (*e.g.*,

[2] Note that the target circuit includes logic circuits evaluated on software; therefore, their results are not limited to hardware implementation.

[3] Such a probability density or mass function does not always exist for any probability distribution. However, for simplicity, we assume that all probability distributions in this study have a probability density or mass function. This assumption allows us to describe most parts of this paper with probability density or mass function without probability measure nor distribution.

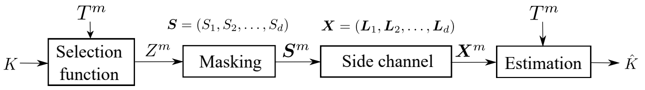

T^m
$\boldsymbol{S} = (S_1, S_2, \ldots, S_d)$
$\boldsymbol{X} = (\boldsymbol{L}_1, \boldsymbol{L}_2, \ldots, \boldsymbol{L}_d)$
T^m
K
Selection function
Z^m
Masking
$\boldsymbol{S}^m$
Side channel
$\boldsymbol{X}^m$
Estimation
$\hat{K}$

Figure 2: Attack model on masked implementation as communication channel.

$\log p_X(X), \log p_{X,Y}(X, Y), \log p_{X|Y}(X \mid Y))$ used in this study are essentially integrable, and therefore, their entropies basically exist. Operator $\oplus$ denotes addition over $\mathbb{F}_2^n$ or an XOR operation. We may represent an element of $\mathbb{F}_2^n$ like an integer (*e.g.*, $(1111)_2 \in \mathbb{F}_2^4$ may be represented as $2^4 - 1$). In this study, log and ln denote binary and natural logarithms, respectively. Finally, for a function $f : \mathbb{F}_2^n \to \mathbb{R}$, its Walsh–Hadamard transform (WHT), which is a discrete Fourier transform (DFT) over $\mathbb{F}_2 = \mathbb{Z}/2\mathbb{Z}$, is denoted by $\hat{f} : \mathbb{F}_2^n \to \mathbb{R}$ (See Appendix A).

3 COMMUNICATION CHANNEL MODEL TO SIDE-CHANNEL ATTACK ON MASKED IMPLEMENTATION

In this section, we formulate side-channel attacks on masked implementations using a communication channel model given as an extension of the non-masked model in [13], which is utilized for deriving our SR upper-bounds in the following section. Figure 2 illustrates the extended communication channel model, where the symbols are given as follows:

- $m \in \mathbb{N}$ is the number of traces used for the attack.
- $d \in \mathbb{N}$ is the number of masking shares.
- K and $\hat{K}$ are random variables for an n-bit secret key and estimated key on a key space $\mathcal{K} = \{0, 1\}^n$, respectively (for example, $n = 8$ for AES).
- $T^m = (T_1, T_2, \ldots, T_m)$ is an m-dimensional random vector for plaintexts or ciphertexts and consists of m random variables $T_1, T_2, \ldots, T_m$, each of which denotes an n-bit partial plaintext/ciphertext (*i.e.*, an element of $\{0, 1\}^n$). In this study, $T_1, T_2, \ldots, T_m$ are assumed to be independent and uniformly distributed.
- $Z^m = (Z_1, Z_2, \ldots, Z_m)$ is an m-dimensional random vector for secret intermediate values. For each $j \in \{1, 2, \ldots, m\}$, Z_j is an n-bit secret value calculated from K and T_j by using a selection function ϕ. The selection function ϕ is assumed to be bijective in terms of T_j for fixed K, as in [13]. For example, $Z_j = \phi(K, T_j) = \mathrm{Sbox}(K \oplus T_j)$ for AES. Here, Z_j is independent and uniformly distributed because T_j is assumed to be the same.
- $\boldsymbol{S}^m = (\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots, \boldsymbol{S}_m)$ is a $(d \times m)$-dimensional random vector for shares of Z^m (this random vector forms a 2D array). For each $j \in \{1, 2, \ldots, m\}$, $\boldsymbol{S}_j$ is a d-dimensional random vector consisting of d random variables for shares $S_1, S_2, \ldots, S_i, \ldots, S_d$. Note that, for Z_j and its shares $\boldsymbol{S}_j = (S_1, S_2, \ldots, S_d)$, $Z_j = S_1 \oplus S_2 \oplus \cdots \oplus S_d$ always holds. Shares $S_1, S_2, \ldots, S_d$ are assumed to be identically and uniformly distributed, which corresponds to the uniformity property commonly required for masking schemes [3, 26, 39, 43]. Note that d-share masking corresponds to *at most* $(d-1)$-th order masking.
- $\boldsymbol{X}^m = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_m)$ is a $(d \times m \times \ell)$-dimensional random vector for side-channel traces (this random vector forms a 3D array), where ℓ is the number of sample points for one leakage (*i.e.*, partial trace corresponding to one share). For each $j \in \{1, 2, \ldots, m\}$, $\boldsymbol{X}_j$ is a $(d \times \ell)$-dimensional random vector of leakages $\boldsymbol{L}_1, \boldsymbol{L}_2, \ldots, \boldsymbol{L}_d$ corresponding to shares $S_1, S_2, \ldots, S_d$, respectively. Leakage $\boldsymbol{L}_i$ ($i \in \{1, 2, \ldots, d\}$) is an ℓ-dimensional random vector on $\mathbb{R}^\ell$. It is assumed that $\boldsymbol{L}_i$ is dependent only on the corresponding share S_i, and is independent of other leakage/shares. In addition, $\boldsymbol{L}_1, \boldsymbol{L}_2, \ldots, \boldsymbol{L}_d$ are assumed to be independent, but not necessarily identically distributed.

The communication channel in Figure 2 corresponds to a Markov chain $K \leftrightarrow Z^m \leftrightarrow \boldsymbol{S}^m \leftrightarrow \boldsymbol{X}^m \leftrightarrow \hat{K}$.

In the following, for simplicity, the subscripts $j \in \{1, 2, \ldots, m\}$ are omitted as it does not need specifying owing to the assumption of independent and identically distributed (i.i.d.). For example, the expectation of $Z_1, Z_2, \ldots, Z_m$ can be represented using that of an independent copy of Z.

This communication channel shows that the secret variable Z is decomposed into d shares as $S_1, S_2, \ldots, S_d$, and that the attacker obtains information about each share S_i from the side-channel leakage (*i.e.*, partial trace) $\boldsymbol{L}_i$. In addition, the attacker is supposed to obtain information about secret variable Z from the entire side-channel trace $\boldsymbol{X}$ that contains leakages on all shares $\boldsymbol{L}_1, \boldsymbol{L}_2, \ldots, \boldsymbol{L}_d$.

Relation to probing model. The attack model as the communication channel is reduced to a d-th order probing model [3, 16, 19, 43]. We assume that a d-th-order probing attacker directly obtains the information of the d shares instead of the corresponding side-channel leakages. In other words, the probing attacker in our model obtains $H(S_i)$-bit information of each share as the maximum value of $I(S_i; \boldsymbol{L}_i)$ (note that $I(S_i; \boldsymbol{L}_i) = H(S_i) - H(S_i \mid \boldsymbol{L}_i) \leq H(S_i)$). Such a probing attacker is far stronger than the real side-channel attacker who observes $\boldsymbol{L}_i$ but does not directly observe S_i, because $H(S_i) \geq I(S_i; \boldsymbol{L}_i)$ and $I(Z; \boldsymbol{S}) \geq I(Z; \boldsymbol{X})$, according to the data processing inequality for the Markov chain $Z \leftrightarrow \boldsymbol{S} \leftrightarrow \boldsymbol{X}$. This indicates that the d-th order probing model attacker always has greater advantages than any other attacker in our model and is equivalent to the strongest attacker in this model. Thus, the proposed attack model is reduced to the probing model. This indicates that the results using Figure 2 can be applied to provably secure masking schemes in the probing model; that is, using the proposed theorems, we can evaluate the SR upper-bound of the d-th order attack on a provably secure $(d-1)$-th order masked implementation in the probing model. Note here that the d-th order probing model mentioned in this paper is not completely equivalent to the original Ishai–Sahai–Wagner (ISW) probing model [30]; however, such a model is common to discuss masked implementations in several previous studies and is closely related to the ISW probing model.

Independence condition. In our communication channel, we assume that each leakage $\boldsymbol{L}_i$ depends only on the corresponding share S_i. This implies that $\boldsymbol{L}_1, \boldsymbol{L}_2, \ldots, \boldsymbol{L}_d$ are i.i.d because the shares $S_1, S_2, \ldots, S_d$ are i.i.d due to the requirement of masking schemes (many masking schemes have been developed with this assumption). This assumption is essential for the proof of Lemma 4.4 and the following theorems. However, several previous studies [1, 14, 15, 22, 24, 38, 42, 47, 48] have pointed out that, in actual masked implementations, each leakage may depend on multiple shares because of physical defaults and/or micro-architectural features such as glitches, transitions, couplings, and interaction. For example, if a leakage $\boldsymbol{L}_1$ depends on the two shares S_1 and S_2, an attacker can estimate the values of these shares through the leakage $\boldsymbol{L}_1$. In other words, S_1 and S_2 are conditionally dependent given $\boldsymbol{L}_1$ (*i.e.*, $S_1 \not\perp\!\!\!\perp S_2 \mid \boldsymbol{L}_1$). In this case, the substantial number of masking shares (security order) will be smaller than the expected one. Therefore, in practice, consideration may need to be given to some security-order reductions, even if the masking scheme is provably secure. Alternatively, masking should be carefully implemented to eliminate such an interaction. A tool that automatically modifies masked software to eliminate leakages due to such an architectural interaction, named Rosita [47, 48], would be useful for realizing practical and secure implementation with the independence condition and would work well with our theorems and profiling method.

Remark 3.1. The independence assumption used in this study makes our analysis simple and general (as used in many previous studies). Actually, the cross-share dependency would be very specific and unique to device/implementation and should analyzed for a given device/implementation. Our analysis and theorems are enough practical for a simple and general evaluation which do not target a specific device, which may be followed by more complex analyses and specific case studies to derive more accurate evaluation for a given device/implementation without the independence condition.

4 INFORMATION-THEORETICAL SR UPPER-BOUNDS

4.1 Overview

In this section, we derive information-theoretical SR upper-bounds from the communication channel model extended above, namely, Theorems 1 and 2. Theorem 4.5 bounds SR with the conditional probability of S_i given $\boldsymbol{L}_i$ (*i.e.*, $p_{S_i \mid \boldsymbol{L}_i}$). Theorem 4.7 bounds SR with the mutual information between S_i and $\boldsymbol{L}_i$ (*i.e.*, $I(S_i; \boldsymbol{L}_i)$). For obtaining the theorems, we first prove Lemma 4.1: SR is upper-bounded by $I(Z^m; \boldsymbol{X}^m \mid T^m)$, owing to the communication channel model. We then prove Lemma 4.2: $I(Z^m; \boldsymbol{X}^m \mid T^m)$ is upper-bounded by $I(Z; \boldsymbol{X})$, which allows for a simple experimental evaluation and makes the following analyses easy. From Lemma 4.1 and Lemma 4.2, we derive Proposition 4.3: an analytical relationship between $I(Z; \boldsymbol{X})$ and the SR of attack on masked implementation. We thirdly prove Lemma 4.4: $I(Z; \boldsymbol{X})$ is upper-bounded using the WHT of $p_{S_i \mid \boldsymbol{L}_i}$. We derive Theorem 4.5 from Lemma 4.4 and Proposition 4.3. Finally, we prove Lemma 4.6: $I(Z; \boldsymbol{X})$ is upper-bounded by the product of $I(S_1 \mid \boldsymbol{L}_1), I(S_2 \mid \boldsymbol{L}_2), \ldots,$ and $I(S_d \mid \boldsymbol{L}_d)$ using Lemma 4.4. We derive Theorem 4.7 from Theorem 4.5 and Lemma 4.6.

4.2 Relation between SR and mutual information

Lemma 4.1. *Let* SR $= \Pr(K = \hat{K})$ *be the success rate of side-channel attacks. In the communication channel shown in Figure 2, the success rate* SR *is upper-bounded using mutual information as follows:*

$$\xi(\mathrm{SR}) \leq I(Z^m; \boldsymbol{X}^m \mid T^m),$$

where $\xi(r)$ is a measurable function on a closed interval $[0, 1]$, defined as

$$\xi(r) = H(K) - (1 - r)\log_2(2^n - 1) - H_2(r),$$

and $H_2(r) = -r\log(r) - (1 - r)\log(1 - r)$ is the binary entropy function.

Proof. We omit this proof because Lemma 4.1 is proven in the manner almost same as in [13], despite the difference between our and de Chérisey *et al.*'s communication channels. □

As proven in Lemma 5.1, the function ξ in Lemma 4.1 is non-negative, is minimized as $\xi(2^{-n}) = 0$, and is maximized as $\xi(1) = n$. Intuitively, ξ converts the probability (*i.e.*, SR) to the entropy of the recovered secret key (*i.e.*, the number of recovered bits of the secret key). For an n-bit secret key, if the attacker has no information about the key, then SR $= \Pr(K = \hat{K}) = 2^{-n}$; that is, the key recovery is equivalent to a completely random guess from 2^n candidates. This indicates that the attacker recovers a zero bit as $\xi(2^{-n}) = 0$. In contrast, if the attacker has all the key information, then SR = 1. This indicates that the attacker recovers n bits, as $\xi(1) = n$. Thus, ξ derives the entropy (*i.e.*, number of key bits to be recovered) to achieve a given SR. Lemma 4.1 states that this entropy is upper-bounded by $I(Z^m; \boldsymbol{X}^n \mid T^m)$. If $I(Z^m; \boldsymbol{X}^m \mid T^m) = 0$ (*i.e.*, the attacker obtains no information from the side-channel traces), then $\xi(\mathrm{SR}) \leq 0$, followed by SR $= 2^{-n}$. However, it is quite difficult to derive or compute the conditional mutual information $I(Z^m; \boldsymbol{X}^n \mid T^m) = \mathbb{E}\log p(Z^m, \boldsymbol{X}^n \mid T^m)/(p(Z^m \mid T^m)p(\boldsymbol{X}^m \mid T^m))$ analytically, because it contains a multiple integral for the expectation. To simplify the analysis and computation, we introduce and utilize Lemma 4.2.

Lemma 4.2. *In the communication channel in Figure 2, mutual information $I(Z^m; \boldsymbol{X}^m \mid T^m)$ and $I(Z; \boldsymbol{X})$ satisfy $I(Z^m; \boldsymbol{X}^m \mid T^m) \leq mI(Z; \boldsymbol{X})$.*

Proof. Because mutual information is decomposed into entropies, $I(Z^m; \boldsymbol{X}^m \mid T^m)$ is upper-bounded as follows:

$$\begin{aligned} I(Z^m; \boldsymbol{X}^m \mid T^m) &\overset{(a)}{=} H(\boldsymbol{X}^m \mid T^m) - H(\boldsymbol{X}^m \mid Z^m, T^m) \\ &\overset{(b)}{=} H(\boldsymbol{X}^m \mid T^m) - H(\boldsymbol{X}^m \mid Z^m) \\ &= H(\boldsymbol{X}^m) - H(\boldsymbol{X}^m \mid Z^m) \\ &\qquad - (H(\boldsymbol{X}^m) - H(\boldsymbol{X}^m \mid T^m)) \\ &\overset{(c)}{=} I(Z^m; \boldsymbol{X}^m) - I(T^m; \boldsymbol{X}^m) \\ &\overset{(d)}{\leq} mI(Z; \boldsymbol{X}), \end{aligned}$$

where the equalities (a) and (c) follow from the definition of mutual information; the equality (b) holds because $(T^m, \boldsymbol{X}^m)$ is conditionally independent given Z^m; and the inequality (d) holds because $I(Z^m; \boldsymbol{X}^m) = mI(Z; \boldsymbol{X})$ and $I(T^m; \boldsymbol{X}^m) \geq 0$. □

Lemma 4.2 indicates that the mutual information with m traces $I(Z^m; \boldsymbol{X}^m \mid T^m)$ is upper-bounded by a multiple of mutual information of one trace $I(Z;\boldsymbol{X})$ by m. The evaluation of $I(Z;\boldsymbol{X})$ is much easier than the computation of $I(Z^m;\boldsymbol{X}^m \mid T^m)$. According to Lemma 4.1 and Lemma 4.2, we obtain Proposition 4.3.

Proposition 4.3. *The success rate* SR *is upper-bounded using mutual information $I(Z,\boldsymbol{X})$ as follows:*

$$\xi(\mathrm{SR}) \leq mI(Z;\boldsymbol{X}).$$

Proof. It is obvious from Lemma 4.1 and Lemma 4.2. □

Proposition 4.3 states that SR is upper-bounded by $mI(Z;\boldsymbol{X})$ in the proposed communication channel model in Figure 2.

4.3 SR upper-bound by conditional probability distribution

We introduce Lemma 4.4 that derives the relation between $I(Z;\boldsymbol{X})$ and the conditional probability distribution $p_{S_i|\boldsymbol{L}_i}$.

Lemma 4.4. *Consider the communication channel in Figure 2. Mutual information $I(Z;\boldsymbol{X})$ is upper-bounded using the conditional probability distribution $p_{S_i|\boldsymbol{L}_i}$ as follows:*

$$I(Z;\boldsymbol{X}) \leq \log\left(\sum_w \prod_{i=1}^d \mathbb{E}\hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i)^2\right),$$

where $\hat{p}_{S_i|\boldsymbol{L}_i}$ is the WHT of $p_{S_i|\boldsymbol{L}_i}$.

Proof. According to the definition of mutual information and Jensen's inequality (8), we have

$$\begin{aligned} I(Z;\boldsymbol{X}) &= H(Z) - H(Z \mid \boldsymbol{X}) \\ &= \mathbb{E}\log p_{Z|\boldsymbol{X}}(Z \mid \boldsymbol{X}) + H(Z) \\ &\leq \log \mathbb{E} p_{Z|\boldsymbol{X}}(Z \mid \boldsymbol{X}) + n \\ &= \log \mathbb{E}\left[\sum_z p_{Z|\boldsymbol{X}}(z \mid \boldsymbol{X})^2\right] + n. \qquad (1) \end{aligned}$$

Here, $\mathbb{E}p_{Z|\boldsymbol{X}}(Z \mid \boldsymbol{X}) = \mathbb{E}\left[\sum_z p_{Z|\boldsymbol{X}}(z \mid \boldsymbol{X})^2\right]$ holds because

$$\begin{aligned} \mathbb{E}p_{Z|\boldsymbol{X}}(Z \mid \boldsymbol{X}) &= \int \sum_z p_{Z,\boldsymbol{X}}(z,\boldsymbol{x}) p_{Z|\boldsymbol{X}}(z \mid \boldsymbol{x})\, d\boldsymbol{x} \\ &= \int p_{\boldsymbol{X}}(\boldsymbol{x}) \sum_z p_{Z|\boldsymbol{X}}(z \mid \boldsymbol{x})^2 \, d\boldsymbol{x} \\ &= \mathbb{E}\left[\sum_z p_{Z|\boldsymbol{X}}(z \mid \boldsymbol{X})^2\right]. \end{aligned}$$

The conditional probability $p_{Z|\boldsymbol{X}}(z \mid \boldsymbol{X})$ is represented by

$$\begin{aligned} p_{Z|\boldsymbol{X}}(z \mid \boldsymbol{X}) &\overset{(e)}{=} \sum_{s_1,\ldots,s_d} p_{Z,S_1,\ldots,S_d|\boldsymbol{L}_1,\ldots,\boldsymbol{L}_d}(z,s_1,\ldots,s_d \mid \boldsymbol{L}_1,\ldots,\boldsymbol{L}_d) \\ &\overset{(f)}{=} \sum_{s_1,\ldots,s_d} p_{Z|S_1,\ldots,S_d}(z \mid s_1,\ldots,s_d) \prod_{i=1}^d p_{S_i|\boldsymbol{L}_i}(s_i \mid \boldsymbol{L}_i) \\ &\overset{(g)}{=} \sum_{s_1 \oplus s_2 \oplus \cdots \oplus s_d = z} \prod_{i=1}^d p_{S_i|\boldsymbol{L}_i}(s_i \mid \boldsymbol{L}_i), \qquad (2) \end{aligned}$$

where the equality (*e*) follows from the marginalization in terms of $(S_1, S_2, \ldots, S_d)$ and $\boldsymbol{X} = (\boldsymbol{L}_1, \ldots, \boldsymbol{L}_d)$; the equality (*f*) follows from the Markov chain $Z \leftrightarrow \boldsymbol{S} \leftrightarrow \boldsymbol{L}$ and the independence of $(S_1,\boldsymbol{L}_1),\ldots,(S_d,\boldsymbol{L}_d)$; and the equality (*g*) follows from the relation between the intermediate value and its shares:

$$p_{Z|S_1,\ldots,S_d}(z \mid s_1,\ldots,s_d) = \begin{cases} 1 & \text{if } z = s_1 \oplus s_2 \oplus \cdots \oplus s_d, \\ 0 & \text{otherwise.} \end{cases}$$

Equation (2) indicates that the conditional probability $p_{Z|\boldsymbol{X}}(\cdot \mid \boldsymbol{X})$ is represented by the XOR convolution of $p_{S_i|\boldsymbol{L}_i}(\cdot \mid \boldsymbol{L}_i)$. Here, we define the WHT of $p_{Z|\boldsymbol{X}}(\cdot \mid \boldsymbol{X})$ by

$$\hat{p}_{Z|\boldsymbol{X}}(w \mid \boldsymbol{X}) = \sum_{z \in \mathbb{F}_2^n} p_{Z|\boldsymbol{X}}(z \mid \boldsymbol{X})(-1)^{\langle w \cdot z \rangle},$$

and the WHT of $p_{S_i|\boldsymbol{L}_i}(\cdot \mid \boldsymbol{L}_i)$ by

$$\hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i) = \sum_{s \in \mathbb{F}_2^n} p_{S_i|\boldsymbol{L}_i}(s \mid \boldsymbol{L}_i)(-1)^{\langle w \cdot s \rangle}.$$

Using these WHTs, Equation (2), and Parseval's identity (7), we have

$$\begin{aligned} \sum_z p_{Z|\boldsymbol{X}}(z \mid \boldsymbol{X})^2 &= 2^{-n} \sum_w \hat{p}_{Z|\boldsymbol{X}}(w \mid \boldsymbol{X})^2 \\ &= 2^{-n} \sum_w \prod_{i=1}^d \hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i)^2. \qquad (3) \end{aligned}$$

Thus, according to Inequality (1), and Equation (3), we have

$$\begin{aligned} I(Z;\boldsymbol{X}) &\leq \log 2^{-n} \mathbb{E} \sum_w \prod_{i=1}^d \hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i)^2 + n \\ &\leq \log \sum_w \prod_{i=1}^d \mathbb{E}\hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i)^2. \end{aligned}$$

□

Lemma 4.4 states that $I(Z;\boldsymbol{X})$ is upper-bounded by the WHT of conditional probability distribution $\hat{p}_{S_i|\boldsymbol{L}_i}$, which is followed by Theorem 4.5.

Theorem 4.5 (SR upper-bound by conditional probability distribution). *Let d and m be the number of masking shares and traces used in an attack, respectively. For each $i \in \{1,2,\ldots,d\}$, let $p_{S_i|\boldsymbol{L}_i}$ be the conditional probability distribution of share S_i given its leakage $\boldsymbol{L}_i$. The success rate* SR *is upper-bounded as*

$$\xi(\mathrm{SR}) \leq m \log \sum_w \prod_{i=1}^d \mathbb{E}\hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i)^2.$$

Proof. It is apparent from Lemma 4.4 and Proposition 4.3. □

Theorem 4.5 can be used for a tight evaluation of SR and number of attack traces with a combination of a DL technique, as proposed in Section 6.

4.4 SR upper-bound by mutual information

We derive an upper-bound of the WHT of $p_{S_i|L_i}$ by $I(S_i; \boldsymbol{L}_i)$ to reveal the relationship between $I(Z; \boldsymbol{X})$ and $I(S_i; \boldsymbol{L}_i)$ with Theorem 4.5.

Lemma 4.6. *Consider the communication channel in Figure 2. For any $i \in \{1, 2, \dots, d\}$, the WHT of the conditional probability distribution $p_{S_i|\boldsymbol{L}_i}$ is upper-bounded using mutual information $I(S_i; \boldsymbol{L}_i)$ as follows:*

$$\begin{cases} \mathbb{E}\hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i)^2 = 1 & (w = 0), \\ \mathbb{E}\hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i)^2 \le 2\ln(2) I(S_i \mid \boldsymbol{L}_i) & \text{(otherwise)}, \end{cases}$$

where ln *denotes the natural logarithm.*

Proof. According to the definition of WHT, we have

$$\hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i) = \sum_{s \in \mathcal{S}} p_{S_i|\boldsymbol{L}_i}(s \mid \boldsymbol{L}_i)(-1)^{\langle w \cdot s \rangle},$$

where $\langle w \cdot s \rangle$ denotes the sum of the products of w and s modulo 2, when we consider them as $\mathbb{F}_2^n$ elements ($\langle w \cdot s \rangle$ is a function from $\mathbb{F}_2^n \times \mathbb{F}_2^n$ to $\{0, 1\}$). Owing to the law of total probability, for $w = 0$, it holds

$$\hat{p}_{S_i|\boldsymbol{L}_i}(0 \mid \boldsymbol{L}_i) = \sum_{s \in \mathcal{S}} p_{S_i|\boldsymbol{L}_i}(s \mid \boldsymbol{L}_i) = 1.$$

Consider the case that $w \neq 0$. Let $\mathbb{1}$ denote the indicator function. The WHT is decomposed as

$$\hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i) = \sum_{s \in \mathcal{S}} p_{S_i|\boldsymbol{L}_i}(s \mid \boldsymbol{L}_i)\mathbb{1}_{\{\langle w \cdot s \rangle = 0\}} - \sum_{s \in \mathcal{S}} p_{S_i|\boldsymbol{L}_i}(s \mid \boldsymbol{L}_i)\mathbb{1}_{\{\langle w \cdot s \rangle = 1\}}.$$

Define a random variable $Y_i^{(w)} = \langle w \cdot S_i \rangle$. The above equation is equivalent to

$$\begin{aligned} \hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i) &= \mathbb{E}\left[\mathbb{1}_{\{\langle w \cdot S_i \rangle = 0\}} \mid \boldsymbol{L}_i\right] - \mathbb{E}\left[\mathbb{1}_{\{\langle w \cdot S_i \rangle = 1\}} \mid \boldsymbol{L}_i\right] \\ &= p_{Y_i^{(w)}|\boldsymbol{L}_i}(0 \mid \boldsymbol{L}_i) - p_{Y_i^{(w)}|\boldsymbol{L}_i}(1 \mid \boldsymbol{L}_i). \end{aligned}$$

Note that, for any $w \neq 0$, it holds that $p_{Y_i^{(w)}}(0) = p_{Y_i^{(w)}}(1) = 1/2$ because the number of candidates of s satisfying $\langle w \cdot s \rangle = 1$ is equivalent to half of $|\mathcal{S}|$. Taking the absolute value of $\hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i)$, we have

$$\begin{aligned} \left|\hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i)\right| &= \left|p_{Y_i^{(w)}|\boldsymbol{L}_i}(0 \mid \boldsymbol{L}_i) - p_{Y_i^{(w)}|\boldsymbol{L}_i}(1 \mid \boldsymbol{L}_i)\right| \\ &\overset{(h)}{=} \left|\left|p_{Y_i^{(w)}|\boldsymbol{L}_i}(0 \mid \boldsymbol{L}_i) - 1/2\right| + \left|p_{Y_i^{(w)}|\boldsymbol{L}_i}(1 \mid \boldsymbol{L}_i) - 1/2)\right|\right| \\ &= \sum_{y \in \{0,1\}} \left|p_{Y_i^{(w)}|\boldsymbol{L}_i}(y \mid \boldsymbol{L}_i) - p_{Y_i^{(w)}}(y)\right| \\ &\le \sqrt{2\ln(2) D_{\mathrm{KL}}(p_{Y_i^{(w)}|\boldsymbol{L}_i} \parallel p_{Y_i^{(w)}})}, \end{aligned}$$

where D_{KL} denotes the Kullback–Leibler (KL) divergence, and we here use Pinsker's inequality (9) to bound $\hat{p}_{S_i|\boldsymbol{L}_i}$ using the KL divergence. The equality (h) holds because $p_{Y_i^{(w)}|\boldsymbol{L}_i}(0 \mid \boldsymbol{L}_i) + p_{Y_i^{(w)}|\boldsymbol{L}_i}(1 \mid \boldsymbol{L}_i) = 1$ and $0 \le p_{Y_i^{(w)}|\boldsymbol{L}_i}(y \mid \boldsymbol{L}_i) \le 1$. Then, squaring both sides and taking expectation, we have

$$\begin{aligned} \mathbb{E}\hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i)^2 &\le 2\ln(2) D_{\mathrm{KL}}(p_{Y_i^{(w)},\boldsymbol{L}_i} \parallel p_{Y_i^{(w)}} p_{\boldsymbol{L}_i}) \\ &= 2\ln(2) I(Y_i^{(w)}; \boldsymbol{L}_i). \end{aligned}$$

As $Y_i^{(w)}$ is considered a function of Z, it holds that $I(Y_i^{(w)}; \boldsymbol{L}_i) \le I(S_i; \boldsymbol{L}_i)$ according to the data processing inequality. Therefore, we have

$$\mathbb{E}\hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i)^2 \le 2\ln(2) I(S_i; \boldsymbol{L}_i),$$

for $w \neq 0$. This completes the proof. □

From Theorem 4.5 and Lemma 4.6, we prove Theorem 4.7.

Theorem 4.7 (SR upper-bound by mutual information). *Let d and m be the number of masking shares and number of traces in an attack, respectively. For each $i \in \{1, 2, \dots, d\}$, let $I(S_i; \boldsymbol{L}_i)$ denote the mutual information between shares S_i and leakages $\boldsymbol{L}_i$. The success rate* SR *is upper-bounded as*

$$\xi(\mathrm{SR}) \le m \log\left((2^n - 1)(2\ln(2))^d \prod_{i=1}^{d} I(S_i; \boldsymbol{L}_i) + 1\right).$$

Proof. According to Theorem 4.5 and Lemma 4.6, we conclude

$$\begin{aligned} \xi(\mathrm{SR}) &\le m \log\left(\sum_{w} \prod_{i=1}^{d} \mathbb{E}\hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i)^2\right) \\ &\le m \log\left(\sum_{w \neq 0} \prod_{i=1}^{d} 2\ln(2) I(S_i; \boldsymbol{L}_i) + 1\right) \\ &= m \log\left((2^n - 1)(2\ln(2))^d \prod_{i=1}^{d} I(S_i; \boldsymbol{L}_i) + 1\right). \end{aligned}$$

□

5 SECURITY PROOF OF MASKED IMPLEMENTATIONS

5.1 Overview

In this section, we prove that SR decreases exponentially by d and that SR converges to $1/2^n$ as $d \longrightarrow \infty$ using Theorem 4.5. We then describe the security of masked implementation using the lemmas and theorems from the viewpoint of the convergence condition of $\mathrm{SR} \longrightarrow 1/2^n$ as $d \longrightarrow \infty$.

5.2 SR convergence

Theorem 4.5 and Theorem 4.7 state that there is a relation between SR and the number of masking shares d. In [20], Duc *et al.* showed that the SR on masked implementations decrease exponentially by d if $I(S_i; \boldsymbol{L}_i)$ is significantly small (*i.e.*, $I(S_i; \boldsymbol{L}_i) \le 2^{-2n+1}$). We prove a similar but stronger result using Theorem 4.5, that holds if $I(S_i; \boldsymbol{L}_i) < 1/(2\ln(2)) \approx 0.72$, which is a much more relaxed condition than $I(S_i; \boldsymbol{L}_i) \le 2^{-2n+1}$ for any $n \in \mathbb{N}$. For the proof, we introduce Lemma 5.1.

Lemma 5.1. *Let SR_d be the success rate of a side-channel attack on a masked implementation with d shares. Let $n \in \mathbb{N}$ be the bit length.*

Let $\epsilon > 0$ be a real number, such that $\forall d \in \mathbb{N}$, $\sup_i \max_{w \neq 0} \mathbb{E}\hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i)^2 < \epsilon$. If $\epsilon < 1$, it holds $\mathrm{SR}_d - 1/2^n = O(\epsilon^{d/2})$ $(d \to \infty)$.

Proof. Define a continuous function $\xi\colon [0,1] \to [0,\infty)$ as $\xi(r) = n - (1-r)\log(2^n - 1) - H_2(r)$. Function ξ is of class C^2 on the open interval $(0,1)$ (namely, there exists the second derivative $\frac{\partial^2 \xi}{\partial r^2}$ on $(0,1)$ and the second derivative is continuous). We prove Lemma 5.1 using Theorem 4.7 and a relation between $\xi(r)$ and $r - 1/2^n$ derived from Taylor's theorem. To derive the above relation, we first show that ξ is strictly convex with a global minimum of $\xi(1/2^n) = 0$.

Show that function ξ is strictly convex with a minimum of 0 at $r_0 = 1/2^n$. Recall that $(1-r)\log(2^n - 1)$ is convex because it is a linear function of r and $-H_2(r)$ is strictly convex. Hence, the function $\xi(r) = n - (1-r)\log(2^n-1) - H_2(r)$, which is the sum of the above (strictly) convex functions and constant coefficient n, is also strictly convex. If a strictly convex function has an extremum, then the extremum is always a unique global minimum. Let r_0 be the stationary point of ξ that minimizes ξ globally. Point r_0 is given by a solution of the following equation:

$$\frac{\partial \xi(r_0)}{\partial r} = \log(r_0) - \log(1 - r_0) + \log(2^n - 1) = 0.$$

Thus, $r_0 = 1/2^n$, and $\xi(r_0) = 0$ is the *global* minimum owing to the convexity.

Derive a relation between $\xi(r)$ and $r - 1/2^n$ using Taylor's theorem. Let d_0 be a positive integer that satisfies

$$\min\{\xi(0), \xi(1)\} > m\log(e)(2^n - 1)\epsilon^{d_0}.$$

For any integer $d \geq d_0$, define a half-open interval $\mathcal{I}_d = \{\, j \in [0,\infty) \mid j < m\log(e)(2^n-1)\epsilon^d \,\}$. Let an interval $\mathcal{U}_d = \xi^{-1}(\mathcal{I}_d)$. According to their definition, it always holds $\mathcal{I}_d \subset \mathcal{I}_{d_0}$, which is followed by $\mathcal{U}_d \subset \mathcal{U}_{d_0}$. Here, $\mathcal{U}_d$ always includes r_0 and is an open interval because ξ is convex. In addition, ξ is of class C^2 on $\mathcal{U}_d$ because $\mathcal{U}_d \subset (0,1)$. Hence, if we consider the second order Taylor expansion of ξ at the point $r_0 \in \mathcal{U}_d$, according to Taylor's theorem, there exists a real number $c \in \mathcal{U}_d \subset \mathcal{U}_{d_0}$ that satisfies

$$\begin{aligned}\xi(r) &= \xi(r_0) + (r - r_0)\frac{\partial \xi(r_0)}{\partial r} + \frac{(r-r_0)^2}{2}\frac{\partial^2 \xi(c)}{\partial r^2} \\ &= \frac{(r-r_0)^2}{2}\frac{\partial^2 \xi(c)}{\partial r^2}.\end{aligned}$$

Recall that ξ is strictly convex, indicating that the range of $\frac{\partial^2 \xi}{\partial r^2}$ is positive and bounded below. Thus, $\xi(r)$ is lower-bounded by

$$\xi(r) \geq \frac{(r-r_0)^2}{2} \inf_{r' \in \mathcal{U}_{d_0}} \frac{\partial^2 \xi(r')}{\partial r^2}. \tag{4}$$

Main part of proof. Let SR_d be the success rate when the number of masking shares is d. Let $\epsilon = 2\ln(2)\sup_i I(S_i; \boldsymbol{L}_i)$. According to Theorem 4.5, Lemma 4.6, and $\ln(1+x) \leq x$, for $d > d_0$, it holds

$$\begin{aligned}\xi(\mathrm{SR}_d) &< m\log\left(\sum_{w \neq 0}\prod_{i=1}^{d} \mathbb{E}\hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i)^2 + 1\right) \\ &< m\log\left((2^n - 1)\epsilon^d + 1\right) \\ &< m\log(e)(2^n-1)\epsilon^d.\end{aligned} \tag{5}$$

Let ξ_c denote a coefficient of right-hand side of Inequality (4); that is, $\xi_c = \inf_{r' \in \mathcal{U}_{d_0}} \frac{\partial^2 \xi(r')}{\partial r^2}$ Note that ξ_c is independent of r and d. Because $\mathrm{SR}_d \in \mathcal{U}_d$, applying Inequality (4) to the left-hand side of Inequality (5), we have

$$\frac{(\mathrm{SR}_d - 1/2^n)^2}{2}\xi_c < m\log(e)(2^n - 1)\epsilon^d,$$

which is followed by

$$\left|\mathrm{SR}_d - 1/2^n\right| < \sqrt{\frac{2m\log(e)(2^n-1)}{\xi_c}}\,\epsilon^{d/2}.$$

As $\sqrt{2m\log(e)(2^n-1)/\xi_c}$ is a constant coefficient that is independent of d, we conclude that $\mathrm{SR}_d - 1/2^n = O(\epsilon^{d/2})$ $(d \to \infty)$. □

From Lemma 5.1, we provide a proof that confirms the security of masking schemes under a relaxed assumption compared to [19].

Theorem 5.2. *If $\forall d \in \mathbb{N}, 2\ln(2)\sup_i I(S_i; \boldsymbol{L}_i) < \epsilon < 1$, then $\mathrm{SR}_d - 1/2^n = O(\epsilon^{d/2})$ $(d \to \infty)$.*

Proof. Lemma 4.6 states that $\forall w \neq 0, \hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i)^2 \leq 2\ln(2)I(S_i \mid \boldsymbol{L}_i)$, and thus $\hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i)^2 < \epsilon$ holds. From Lemma 5.1 and $\mathbb{E}\hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i)^2 < \epsilon$, we can prove $\mathrm{SR}_d - 1/2^n = O(\epsilon^{d/2})$ $(d \to \infty)$. □

In Theorem 5.2, the condition that $\epsilon < 1$ corresponds to that $\sup_i I(S_i; \boldsymbol{L}_i) < 1/(2\ln(2)) \approx 0.72$. Thus, if $I(S_i; \boldsymbol{L}_i) < 1/(2\ln(2))$, the number of traces required for attack success increases exponentially with d. Theorem 5.2 also states that, if $I(S_i; \boldsymbol{L}_i) < 1/(2\ln(2))$, the success rate $\mathrm{SR} = \Pr(\hat{K} = K)$ converges to $1/2^n$ when $d \to \infty$, which indicates that we can make the attacker's advantage arbitrary smaller by increasing d.

Remark 5.1. Theorem 5.2 shows an *asymptotic* SR decay, which implies that SR may not decay exponentially by d if d is small and/or ϵ is large. In other words, Theorem 5.2 also states that it would be difficult to protect the cryptographic implementation by masking with practical d if ϵ is very large (*i.e.*, the noise level is very low), as experimentally demonstrated in [7]. However, the exponential SR decay for small d (in this study, $d = 1$, 2, and 3) when $\sup_i I(S_i; \boldsymbol{L}_i) < 1/(2\ln(2))$ is confirmed through a numerical experiment in Section 7, as the number of traces required for an attack success increases exponentially by even small d in the some experimental condition. The detailed analysis on the condition for an exponential SR decay would be a future work, although our novelty and theoretical contribution also include that we show that masking scheme is asymptotically secure for large d with a relaxed condition compared to the previous study.

5.3 Conditions for security through masking scheme

This subsection discusses the convergence condition of $\mathrm{SR} \to 1/2^n$ as $d \to \infty$, using Lemma 5.1 and Theorem 5.2.

First, we discuss what happens when mutual information $I(S_i; \boldsymbol{L}_i)$ is significantly large from the viewpoint of Theorem 5.2. Consider an extreme case (the strongest attacker in our model) in which $I(S_i; \boldsymbol{L}_i) = H(S_i)$ (*i.e.*, $I(S_i; \boldsymbol{L}_i)$ is maximized) for each i. This attacker always succeeds recovering the secret key K independently of d, because the attacker can exactly know Z from $S_1, S_2, \ldots, S_d$ (note that such an attacker is equivalent to a d-th order probing attacker discussed in Section 3). In this case, $I(Z; \boldsymbol{X})$ does not decrease with increasing d. In fact, a masked implementation can be secure[4] only if $\sup_i I(S_i; \boldsymbol{L}_i) < \gamma$ for some positive real number γ, which guarantees that $I(Z; \boldsymbol{X})$ decreases with an increase in d. If a device has a side-channel leakage amplitude $I(S_i; \boldsymbol{L}_i)$ greater than the threshold γ, no masking scheme can protect the implementation on the device (*i.e.*, the attacker can succeed with a trivial number of traces), even when d is large.

Duc *et al.*'s result shows that $\gamma \geq 2^{-2n+1}$, which indicates that their result is valid only if the SNR of the side-channel measurement is significantly small. For example, in the case of AES (*i.e.*, $n = 8$), the mutual information should be $I(S_i; \boldsymbol{L}_i) \leq 2^{-15} \approx 3.05 \times 10^{-5}$, which is too small to evaluate the side-channel resistance. Note again that the value of $I(S_i; \boldsymbol{L}_i)$ corresponds to the side-channel leakage amplitude or approximately represents the SNR of the measurement. In this study, we proved that $\gamma \geq 1/(2\ln(2)) \approx 0.72$ independently of n. Thus, Theorem 5.2 is a much stronger and more generalized result than Duc *et al.*'s result. However, its tightness is still unclear (although it is far tighter than the existing bound). The derivation of exact value of γ is an open problem.

Then, we consider the security of masking schemes from the viewpoints of Lemma 5.1, the meanings of WHT, and the random variable $Y_i^{(w)} = \mathbb{1}_{\{\langle w \cdot S_i \rangle = 1\}}$ defined in the proof of Lemma 4.6, instead of the value of $I(S_i; \boldsymbol{L}_i)$. In a $2^n \times 2^n$ WHT matrix, a column (corresponding to a value of w) contains 2^{n-1} number of 1 and 2^{n-1} number of -1 coefficients, except for the case where $w = 0$. Therefore, given $w \neq 0$, the values that a share can take are divided into two sets of $\{ s \in \mathcal{S} \mid \langle w \cdot s \rangle = 0 \}$ and $\{ s \in \mathcal{S} \mid \langle w \cdot s \rangle = 1 \}$, corresponding to the coefficients of 1 and -1, respectively. For $w \neq 0$, the value of WHT, namely $\left|\hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i)\right| = \left|\Pr(Y_i^{(w)} = 0 \mid \boldsymbol{L}_i) - \Pr(Y_i^{(w)} = 1 \mid \boldsymbol{L}_i)\right|$, represents a confidence value that attacker can distinguish which S_i belongs to $\{ s \in \mathcal{S} \mid \langle w \cdot s \rangle = 0 \}$ or $\{ s \in \mathcal{S} \mid \langle w \cdot s \rangle = 1 \}$ from a given leakage. We call this value distinguish advantage in this paper. Lemma 4.4 and Lemma 4.6 state that the success rate $\xi(\mathrm{SR})$ (precisely, the mutual information $I(Z; \boldsymbol{X})$) is upper-bounded by the maximum distinguish advantage as

$$\xi(\mathrm{SR}) \leq m \log \left(1 + \sum_{w \neq 0} \prod_{i=1}^{d} \mathbb{E} \left| p_{Y_i^{(w)}|\boldsymbol{L}_i}(0 \mid \boldsymbol{L}_i) - p_{Y_i^{(w)}|\boldsymbol{L}_i}(1 \mid \boldsymbol{L}_i) \right|^2 \right).$$

For example, if the attacker cannot estimate which s_i belongs to $\{ s \in \mathcal{S} \mid \langle w \cdot s \rangle = 0 \}$ or $\{ s \in \mathcal{S} \mid \langle w \cdot s \rangle = 1 \}$ for any $w \neq 0$ from $\boldsymbol{L}_i$, the distinguish advantage is zero as $\hat{p}_{S_i|\boldsymbol{L}_i} = |1/2 - 1/2| = 0$. In this case, it holds $\xi(\mathrm{SR}) = m \log 1 = 0$, which implies that the attacker has no advantage in key recovery (*i.e.*, $\mathrm{SR} = 1/2^n$). By contrast, if the attacker can completely estimate/distinguish it for a given $w = v \neq 0$, the distinguish advantage for v is maximized as $\hat{p}_{S_i|\boldsymbol{L}_i}(v \mid \boldsymbol{L}_i) = 1$. Recall that $\xi(\mathrm{SR}) \leq \log \sum_w \prod_i \mathbb{E}\hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i)^2$. If a complete distinguish is possible for $w = v \neq 0$ (*i.e.*, $\hat{p}_{S_i|\boldsymbol{L}_i}(v \mid \boldsymbol{L}_i) = 1$), then $\log \sum_w \prod_i \mathbb{E}\hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i)^2 \geq \log 2 = 1$ for any d. This indicates that such an attack would succeed in key recovery with only n traces according to Theorem 4.5 for any d. In addition, the convergence rate of SR $\longrightarrow 1/2^n$ as $d \longrightarrow \infty$ (in Theorem 5.2) mainly depends on $\max_{w \neq 0} \mathbb{E}\hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i)^2$. This is because, if $\boldsymbol{L}_1, \boldsymbol{L}_2, \ldots,$ and $\boldsymbol{L}_d$ are i.i.d, $\sum_w \prod_{i=1}^{d} \mathbb{E}\hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i)^2$ in Theorem 4.5 can be bounded as

$$\sum_{w} \prod_{i=1}^{d} \mathbb{E}\hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i)^2 = 1 + \sum_{w \neq 0} \left[\mathbb{E}\hat{p}_{S|\boldsymbol{L}}(w \mid \boldsymbol{L})^2\right]^d$$
$$\leq 1 + (2^n - 1) \left[\max_{w \neq 0} \mathbb{E}\hat{p}_{S|\boldsymbol{L}}(w \mid \boldsymbol{L})^2\right]^d. \quad (6)$$

Here, we omit the subscript i and replace $\prod_i$ with the d-th power owing to the assumption that $\boldsymbol{L}_1, \boldsymbol{L}_2, \ldots,$ and $\boldsymbol{L}_d$ are i.i.d (see Section 6 for the detail). This formally represents a simplified upper-bound of the attacker's advantage in distinguishing between $Y_i^{(w)} = 1$ or $Y_i^{(w)} = 0$ (*i.e.*, $S_i \in \{ s \in \mathcal{S} \mid \langle w \cdot s \rangle = 1 \}$ or $S_i \in \{ s \in \mathcal{S} \mid \langle w \cdot s \rangle = 0 \}$). In addition, if $\max_{w \neq 0} \mathbb{E}\hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i)^2 = 1$, the SR does not converge to $1/2^n$ and the attack is successful with a trivial number of traces. Therefore, in order for a masked implementation to be secure, it must be non-trivially difficult to distinguish which $S_i \in \{ s \in \mathcal{S} \mid \langle w \cdot s \rangle = 1 \}$ or $S_i \in \{ s \in \mathcal{S} \mid \langle w \cdot s \rangle = 0 \}$ is true for any w from the leakage $\boldsymbol{L}_i$. In this sense, side-channel traces must contain noise to guarantee security against side-channel attacks, as well as the discussion about γ in the above.

5.4 Security of masking scheme with Hamming weight leakage

We discussed that masking countermeasures are meaningless when the share values leak as they are, and there is only trivial noise. This subsection details how noise plays an important role even when information of shares are leaked as its Hamming weight (HW). Let us consider an attacker who can obtain the HW of S_i from $\boldsymbol{L}_i$, as many attacks on software implementation. Let $H_i = \mathrm{HW}(S_i)$. The attacker has to obtain H_i from $\boldsymbol{L}_i$ to estimate S_i. We then focus on the value of w which maximizes the distinguish advantage of the above HW-based attacker. In the absence of noise, we can easily confirm that the masking countermeasures do not make sense if w exists such that $\mathcal{H}_0 \cap \mathcal{H}_1 = \emptyset$, where $\mathcal{H}_0 = \{ \mathrm{HW}(s) \mid \langle w \cdot s \rangle = 0, s \in \mathcal{S} \})$ and $\mathcal{H}_1 = \{ \mathrm{HW}(s) \mid \langle w \cdot s \rangle = 1, s \in \mathcal{S} \})$. In fact, such w exists; for example, $w = 2^n - 1$ because $\mathcal{H}_0 = \{ \mathrm{HW}(s) \mid \langle (2^n - 1) \cdot s \rangle = 0, s \in \mathcal{S} \} = \{ h \mid h = 0 \mod 2, h \in \mathcal{H} \}$ and $\mathcal{H}_1 = \{ \mathrm{HW}(s) \mid \langle (2^n - 1) \cdot s \rangle = 1, s \in \mathcal{S} \} = \{ h \mid h = 1 \mod 2, h \in \mathcal{H} \}$, where $\mathcal{H}$ is the set of HWs (*e.g.*, $\mathcal{H} = \{0, 1, \ldots, 8\}$ when $n = 8$). Its intuitive meaning can be explained through the following proposition:

Proposition 5.3. *Let s_1 and s_2 be variables over $\mathbb{F}_2^n$. It holds that*

(1) *If* $\mathrm{HW}(s_1)$ *and* $\mathrm{HW}(s_2)$ *are even,* $\mathrm{HW}(s_1 \oplus s_2)$ *is even.*
(2) *If* $\mathrm{HW}(s_1)$ *and* $\mathrm{HW}(s_2)$ *are odd,* $\mathrm{HW}(s_1 \oplus s_2)$ *is even.*
(3) *If* $\mathrm{HW}(s_1)$ *is even and* $\mathrm{HW}(s_2)$ *is odd,* $\mathrm{HW}(s_1 \oplus s_2)$ *is odd.*

Proof. Note that $\mathrm{HW}(s_1 \oplus s_2) = \mathrm{HW}(s_1) + \mathrm{HW}(s_2) - 2\mathrm{HW}(s_1 \wedge s_2)$. If $\mathrm{HW}(s_1)$ and $\mathrm{HW}(s_2)$ are even, then there exist k_1 and $k_2 \in \mathbb{N}$ such that $\mathrm{HW}(s_1) = 2k_1$ and $\mathrm{HW}(s_2) = 2k_2$. Thus, $\mathrm{HW}(s_1 \oplus s_2)$ is even, because $\mathrm{HW}(s_1 \oplus s_2) = 2(k_1 + k_2) - 2\mathrm{HW}(s_1 \wedge s_2)$.

[4] Here, we mean that a masked implementation is secure if, by the increase in d, SR decreases and the number of traces for attack success increases exponentially. This also indicates that, for a secure masked implementation, SR $-$ $1/2^n$ is *negligible* in terms of d.

If $\mathrm{HW}(s_1)$ and $\mathrm{HW}(s_2)$ are odd, then there exist k_1 and k_2 such that $\mathrm{HW}(s_1) = 2k_1 + 1$ and $\mathrm{HW}(s_2) = 2k_2 + 1$. Thus, $\mathrm{HW}(s_1 \oplus s_2)$ is even, because $\mathrm{HW}(s_1 \oplus s_2) = 2(k_1 + k_2) + 2 - 2\mathrm{HW}(s_1 \wedge s_2)$.

If $\mathrm{HW}(s_1)$ is even and $\mathrm{HW}(s_2)$ is odd, then there exist k_1 and k_2 such that $\mathrm{HW}(s_1) = 2k_1$ and $\mathrm{HW}(s_2) = 2k_2 + 1$. Thus, $\mathrm{HW}(s_1 \oplus s_2)$ is odd, because $\mathrm{HW}(s_1 \oplus s_2) = 2(k_1 + k_2) + 1 - 2\mathrm{HW}(s_1 \wedge s_2)$. □

According to Proposition 5.3, the parity of the HW of the secret intermediate value (*i.e.*, $\mathrm{HW}(Z)$) can be calculated if the parity of each $\mathrm{HW}(S_i)$ is known. The parity of $\mathrm{HW}(Z)$, which is one-bit information about Z, enables the attacker to halve the number of candidates for secret key values. If the traces do not contain any noise, the attack will succeed in only n traces on average, no matter how much d increases. Thus, an HW-based attacker can easily exploit the distinguish advantage as $\max_{w \neq 0} \mathbb{E}\hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i)^2$. If the attacker can always obtain the correct HW (*i.e.*, $\max_{w \neq 0} \mathbb{E}\hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i)^2 = 1$), no masking scheme can protect the device even with a very large d. In addition, according to Inequality (6) (which represents a simplified upper-bound of d-th order attacker's distinguish advantage), the convergence rate of $\mathrm{SR}_d \longrightarrow 1/2^n$ as $d \longrightarrow \infty$ mainly depends on

$$\begin{aligned}\max_{w \neq 0} \mathbb{E}\hat{p}_{S|\boldsymbol{L}}(w \mid \boldsymbol{L})^2 &= \mathbb{E}\hat{p}_{S|\boldsymbol{L}}(2^n - 1 \mid \boldsymbol{L})^2 \\ &= \mathbb{E}\Big(\Pr(\mathrm{HW}(S) \text{ is even} \mid \boldsymbol{L}) - \Pr(\mathrm{HW}(S) \text{ is odd} \mid \boldsymbol{L})\Big)^2,\end{aligned}$$

for the HW-based attacker.

6 TIGHT SUCCESS RATE EVALUATION IN PRACTICE

In this section, it is assumed that $\boldsymbol{L}_1, \boldsymbol{L}_2, \ldots, \boldsymbol{L}_d$ are i.i.d, which implies that there exists $p_{S|\boldsymbol{L}}$ and $I(S; \boldsymbol{L})$ such that $p_{S|\boldsymbol{L}} = p_{S_i|\boldsymbol{L}_i}$ and $I(S; \boldsymbol{L}) = I(S_i; \boldsymbol{L}_i)$ for any i, respectively. In other words, we omit the subscript i and use an independent copy of these random variables/vectors owing to the i.i.d assumption, as well as subscript j in Section 3. This implies that the leakage characteristics/amplitude is identical for all shares, and implies that $I(S_i; \boldsymbol{L}_i) = I(S_{i'}; \boldsymbol{L}_{i'})$ and $p_{S_i|\boldsymbol{L}_i} = p_{S_{i'}|\boldsymbol{L}_{i'}}$ for any i and i'. This assumption is likely to hold for certain applications. For example, because masked software may process each share in a serial manner on an identical device and instructions, the leakage characteristics/amplitude is likely to be identical for all shares. This assumption enables us to evaluate the SR of attacks on masked implementations from an experiment (*i.e.*, profiling) using a non-protected implementation on an identical device.[5] Note that $I(S; \boldsymbol{L})$ is equivalent to $I(Z; \boldsymbol{X})$ if $d = 1$ (*i.e.*, without masking).

To evaluate the SR using Theorem 4.5, we required to know the conditional probability distribution $p_{S|\boldsymbol{L}}$ (*i.e.*, a leakage characteristics), whereas Theorem 4.7 can be used if the mutual information $I(S; \boldsymbol{L})$ (*i.e.*, leakage amplitude or SNR) is known. For a practical and tight SR evaluation, it would be better to use the $p_{S|\boldsymbol{L}}$ based bound of Theorem 4.5, as it is tigher than the $I(S; \boldsymbol{L})$ based bound of Theorem 4.7. However, there have been few studies on the estimation of the conditional probability distribution $p_{S|\boldsymbol{L}}$ for Theorem 4.5. This is because estimating the conditional probability distribution is difficult owing to the curse of dimensionality. Therefore, for the practical and tight SR evaluation, which would be useful for design flow of cryptographic modules, we propose a profiling method to estimate $p_{S|\boldsymbol{L}}$ using a DL technique. In contrast, note that several studies have been devoted to estimate mutual information $I(S; \boldsymbol{L})$ [13, 25, 37], which indicates that, given a device, we can evaluate the SR with Theorem 4.7 by estimating $I(S; \boldsymbol{L})$ using these previous studies through an experiment to profile a non-masked implementation.

Theorem 4.5 states that SR is upper-bounded using the expected WHT of conditional probability $p_{S_i|\boldsymbol{L}_i}$ as

$$\xi(\mathrm{SR}) \leq m \log \left(\sum_{w} \prod_{i=1}^{d} \mathbb{E}\hat{p}_{S_i|\boldsymbol{L}_i}(w \mid \boldsymbol{L}_i)^2 \right).$$

If the above assumption that $p_{S|\boldsymbol{L}} = p_{S_i|\boldsymbol{L}_i}$ holds for any i, this inequality can be written as

$$\xi(\mathrm{SR}) \leq m \log \sum_{w} \left[\mathbb{E}\hat{p}_{S|\boldsymbol{L}}(w \mid \boldsymbol{L})^2 \right]^d,$$

where $\hat{p}_{S|\boldsymbol{L}}$ is the WHT of $p_{S|\boldsymbol{L}}$. This indicates that, if we can estimate $p_{S|\boldsymbol{L}}$ precisely, we can approximately evaluate the SR upper-bound tightly.

We employ a neural network (NN) to approximate the conditional probability distribution $p_{S|\boldsymbol{L}}$. Let $q(s \mid \boldsymbol{L}; \theta)$ be the conditional probability distribution represented by NN, where θ is a model parameter. In addressing a multiclass classification problem with DL, the negative log likelihood (NLL, also known as categorical cross-entropy) is commonly used as a loss function. NLL is known to be asymptotically equivalent to cross-entropy defined as

$$\mathrm{CE}(p, q) = -\mathbb{E} \log q(S \mid \boldsymbol{L}; \theta) = -\int \sum_{s} p(s, \boldsymbol{\ell}) \log q(s \mid \boldsymbol{\ell}; \theta) d\boldsymbol{\ell}.$$

The cross-entropy $\mathrm{CE}(p, q)$ takes the global minimum if and only if $p = q$, which indicates that p can be imitated/approximated using the conditional probability distribution $q_{S|\boldsymbol{L}}$ estimated using DL with the NLL as a loss function. Let $\hat{\theta}$ be the trained model parameter. The SR is approximately upper-bounded by

$$\xi(\mathrm{SR}) \lessapprox m \log \sum_{w} \left[\mathbb{E}\hat{q}_{S|\boldsymbol{L}}(w \mid \boldsymbol{L}; \hat{\theta})^2 \right]^d.$$

Remark 6.1. Specifically, we used the DL technique to estimate $I(Z; \boldsymbol{X})$ according to Lemma 4.4. This estimation is far more precise than using Lemma 4.6. In other words, a precise estimation of $I(Z; \boldsymbol{X})$ (and SR) can be achieved using the DL-based estimation of $p_{S|\boldsymbol{L}}$ instead of Lemma 4.6. Therefore, Theorem 4.7 is important especially in theory and reveals the properties and conditions required for a secure masked implementation, whereas the proposed SR estimation method with Theorem 4.5 and DL-based approximation is useful especially in practice and builds a bridge between theory and practice.

[5] This means that we assume that $p_{S|\boldsymbol{L}}$ and $I(S; \boldsymbol{L})$ are identical for the two implementations if they are on the same device and are in a similar style. For example, in the case of software implementation, we assume that the SR of attacks on bit-slicing masked software implementation can be evaluated by profiling the corresponding instruction in a non-masked bit-slicing software implementation on the same device. As well, the SR of an attack on a table-lookup-based masked implementation can be evaluated by profiling the corresponding instruction in a table-lookup-based non-masked implementation.

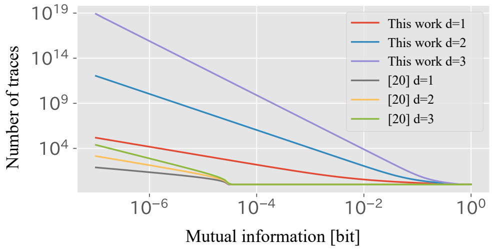


Figure 3: Number of required traces in attacking AES for achieving SR = 0.99 **when** $d = 1, 2,$ **and** 3**.**

7 EXPERIMENTAL VALIDATION

7.1 Evaluation using only mutual information

We first evaluate our bound of Theorem 4.7 in comparison with Duc *et al.*'s bound [20] for a simple validation from a theoretical viewpoint. Figure 3 shows the number of traces in attacking AES (*i.e.*, $n = 8$) for achieving SR = 0.99 evaluated only from $I(S; \boldsymbol{L})$ using Theorem 4.7. For comparison, Figure 3 shows the conventional bound in [20]. Here, previous studies [13, 37] were not evaluated because their methods are inapplicable to masked implementations. Given a value of mutual information $I(S; \boldsymbol{L})$ (*i.e.*, leakage amplitude), each curve denotes the lower-bound of the number of traces to achieve SR = 0.99. For a given d, the evaluation result is more precise if the result is a larger number of traces, as they are lower-bounds. From Figure 3, we can confirm that our bound is much tighter than the conventional bound. For example, when $d = 3$, our results indicates that the attack success with 99% probability requires at least approximately 10^{16} traces, whereas the conventional bound states that attack success requires at least approximately 10^3 traces. In addition, with regard to Remark 5.1, we confirm that the number of traces required for an attack success increases by even small $d = 1$, 2, and 3 in this experiment (note that the vertical axis is in the logarithmic scale). Moreover, as mentioned previously, the conventional bound is valid only if $I(S; \boldsymbol{L}) < 2^{-15} \approx 3.05 \times 10^{-5}$. Therefore, their result is trivial when $I(S; \boldsymbol{L}) \geq 2^{-15}$, that is, it only states that at least one trace is required for attack success. By contrast, our bound is meaningful for a wider range of $I(S; \boldsymbol{L})$. Thus, we confirm the effectiveness of our theorems.

7.2 Evaluation using simulated traces

We further validate the effectiveness and practicality of Theorem 4.5 with a DL technique and Theorem 4.7 through experimental simulations of attacks on AES. We target the first-round S-box output $Z = \mathrm{Sbox}(k^* \oplus T)$, where k^* is an eight-bit secret key, and Z is masked as $Z = S_1 \oplus S_2 \oplus \cdots \oplus S_d$ if $d > 1$ (note that $d = 1$ corresponds to a non-masked implementation). We generate the corresponding leakage for each share such that the leakage is given by the Hamming weight (HW) of S_i with Gaussian noise, that is, for each $i \in \{1, 2, \ldots, d\}$, $\boldsymbol{L}_i = \mathrm{HW}(S_i) + N_i$, where N_i is Gaussian noise. According to the Shannon–Hartley theorem, the upper-bound of mutual information of a channel with Gaussian noise is given by $\log(1 + \mathrm{SNR})/2$. Therefore, the SR upper-bound/number of traces can be evaluated for different values of mutual information $I(S; \boldsymbol{L})$ by changing the variance of the Gaussian noise. For the evaluation of Theorem 4.5 with a DL approximation, we employ a multilayer perceptron (MLP) that consists of four layers with output dimensions of 128, 256, 128, and 256 in the order of input to output layers. The output layer has a softmax function as an activation function, and the other layers have an ELU function. We employ Adam as the optimizer and NLL as the loss function. We use five million simulated traces for each training and test. In addition, we adopt a Python open-source library NPEET [53] to estimate mutual information $I(S; \boldsymbol{L})$ (note that the estimation is different from the upper-bound by the Shannon–Hartley theorem). We use 10 million simulated traces for the estimation.

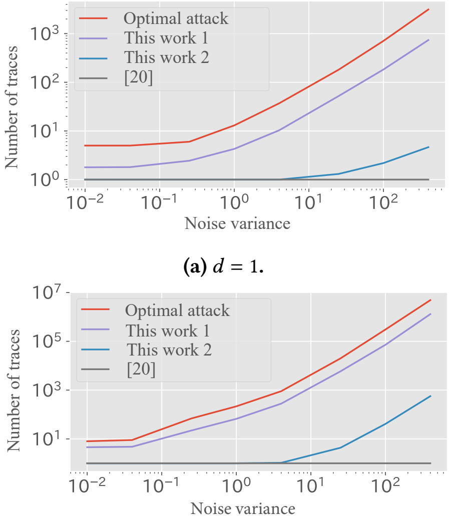
We also perform a template attack on the simulated traces for the ground truth. Since the distribution of traces is known, the template in the attack is an optimal distinguisher that theoretically maximizes SR and minimizes the number of required traces, as proven in [8, 29, 32]. Figure 4a and Figure 4b show the numbers of traces in attacking AES to achieve SR = 0.80 when $d = 1$ and 2, respectively. These figures include the evaluation results of the template attack, the bound of Theorem 4.5 with a DL-based approximation of $p_{S|\boldsymbol{L}}$ (denoted by "This work 1"), the bound of Theorem 4.7 with $I(S; \boldsymbol{L})$ (denoted by "This work 2"), and the conventional bound with $I(S; \boldsymbol{L})$ in [20]. A bound is tighter if the result is closer to that of a template attack, and such a tight bound is useful for practical evaluation. From these figures, we confirm that "This work 1" is far tighter than the other bounds, including "This work 2." This is because "This work 1" evaluates the bound by exploiting much information (*i.e.*, conditional probability distribution) rather than the leakage amplitude or SNR (*i.e.*, mutual information value). In contrast, as in Lemma 5.1, "This work 2" (*i.e.*, Theorem 4.7) is meaningful only if $I(S; \boldsymbol{L}) < 1/(2\ln(2)) \approx 0.72$;[6] otherwise, it shows that at least one trace is required for an attack success. Note that the conventional bound shows that at least one trace is always required for attack success, as it never holds $I(S; \boldsymbol{L}) < 2^{-2n+1} \approx 3.05 \times 10^{-5}$ in this experiment. Thus, we confirm the effectiveness and practicality of the proposed theorems and evaluation method method. In particular, the evaluation of "This work 1" is more precise, tight, and practical than other bounds, thanks to the proposed DL-based approximation of the conditional probability distribution (*i.e.*, profiling the leakage characteristics).

8 CONCLUSION

8.1 Summary

This study derived information-theoretical bounds of SR in attacking masked implementations. The derived bounds are used to evaluate the security of masked implementations from mutual information or conditional probability distribution and can be estimated by profiling a non-masked implementation on a target device. Based on a numerical evaluation and an experimental simulation, we confirmed the effectiveness, tightness, and practicality of the proposed bounds and profiling method. In addition, this paper also provided a proof for the concrete security of masking schemes, discussed

[6] Note that the convergence condition of $I(S; \boldsymbol{L}) < 1/(2\ln(2))$ (*i.e.*, $1 < \epsilon$) comes from the proof of Lemma 4.6, but is related to neither Theorem 4.5 nor Lemma 4.4. This is one of reasons why the bound of Theorem 4.5 is tighter than that of Theorem 4.7.



(a) $d = 1$.

(b) $d = 2$.

Figure 4: Evaluation results of number of traces in attacking masked AES implementation with $\mathrm{SR} = 0.80$.

some important aspects of masked implementation from the viewpoint of the convergence of SR $\to 1/2^n$ as $d \to \infty$, and showed that SNR and the distinguish advantage play an essential role in attacking/protecting masked cryptographic devices. A more concrete and quantitative analysis/evaluation of the relationship between convergence and $I(S; \boldsymbol{L})$ or $\max_{w \neq 0} \mathbb{E}\hat{p}_{S|\boldsymbol{L}}(w \mid \boldsymbol{L})^2$ remains an open problem that is important from both practical and theoretical viewpoints.

Our theorems/method for cryptographic module designs. The proposed theorems can be used to determine the number of masking shares d (or mask order) for a given device, which would be useful for practical system design. For example, the implementer first determines the maximum number of encryptions/decryptions using an identical key, namely, the number of traces that can be acquired by a side-channel attacker. This can be determined based on the application and mode of operation. Adoption of leakage-resilient authenticated encryption (*e.g.*, Ascon-AEAD, ISAP, TEDT, *etc.* [6, 17, 18]) would contribute to an increase of this number. Then, he/she determines d using the proposed theorems and method, such that the number of traces required for attack success should be greater than the above number. Rosita would be useful for secure and practical cryptographic software design with the independence condition [47, 48]. Finally, he/she implements a masked cipher and verifies whether the $(d-1)$-th order masking is correctly implemented, for example, using a formal/symbolic verification tool such as maskVerif and SILVER [2, 33] and/or the test vector leakage assessment (TVLA) [46].

8.2 Future works

Extension to non-Boolean masking. In this study, we focused on Boolean masking as the most major scheme. On one hand, Theorem 4.5 and the proposed DL-based SR evaluation method would be extended and applicable to other *additive* masking schemes such as arithmetic masking, by utilizing a DFT over the field/ring instead of WHT. On the other hand, Theorem 4.7 cannot be extended to the other masking schemes because it is unknown how to apply our proof strategy with Pinsker's inequality to non-Boolean masking schemes. In addition, it has been discussed that, given a noise level, the SCA resistance of some non-Boolean masking schemes would be higher than Boolean masking [21], which indicates that the threshold γ for secure masking condition may be different between Boolean and non-Boolean masking schemes. Further analyses for non-Boolean masking would be a future work.

On independence condition and real device evaluation. The experimental simulation results in this study confirmed the effectiveness and practicality of proposed theorems/method for a simple but general case (with the independence condition) that does not target a specific device. However, a real device may have such a dependency. Actually, the independence assumption may (partially) hold for some software implementations that process each share in a serial manner, although it would be not valid for some implementations due to, for example, micro-architectural features. We emphasize that a tool named Rosita addresses this problem: it would be useful to realize a secure and practical masked implementation that satisfies the independence condition [47, 48]. In contrast, this assumption may not hold for many masked hardware implementations, as they usually process all shares in parallel in one clock cycle, and glitch and coupling effects may cause dependency between shares [1, 14, 36]. It is important to investigate the effect of the dependency on our theorems/method and how to treat the dependency appropriately. Also, developing a profiling method for share-parallel implementations to use the proposed theorems would be subject to future research work.

Investigation of NN architecture/hyperparameter for proposed method. Although the effectiveness and practicality of proposed DL-based SR estimation are validated through the experiments, as important future works, we plan to investigate the impact of the NN approximation error on the resulting SR evaluation and develop an efficient NN architecture for our purpose. In addition, for a further validation, it would be also important to analyze the effect of NN approximation error on the proposed method.

APPENDIX A: WALSH–HADAMARD TRANSFORM (WHT)

Walsh–Hadamard transform (WHT) is a discrete Fourier transform (DFT) over $\mathbb{F}_2 = \mathbb{Z}/2\mathbb{Z}$. WHT of a function $f : \mathbb{F}_2^n \to \mathbb{R}$ is a function $\hat{f} : \mathbb{F}_2^n \to \mathbb{R}$ and is defined as

$$\hat{f}(w) = \sum_{s \in \mathbb{F}_2^n} f(s)(-1)^{\langle w \cdot s \rangle},$$

where $\langle w \cdot s \rangle$ denotes the sum of the products of each element modulo 2 (defined as a function from $\mathbb{F}_2^n \times \mathbb{F}_2^n$ to $\{0, 1\}$), like an inner product. Inverse WHT is written by

$$f(w) = \frac{1}{2^n} \sum_{s \in \mathbb{F}_2^n} \hat{f}(s)(-1)^{\langle w \cdot s \rangle}.$$

A convolution over $\mathbb{F}_2^n$ can be decomposed into a product in the WHT representation. Let f_1 and f_2 be functions from $\mathbb{F}_2^n$ to $\mathbb{R}$, and let f be their convolution:

$$f(s) = \sum_{s_1 \oplus s_2 = s} f_1(s_1) f_2(s_2) = (f_1 * f_2)(s).$$

This is equivalently represented using their respective WHTs, $\hat{f}(w)$, $\hat{f}_1(w)$, and $\hat{f}_2(w)$, as

$$\hat{f}(w) = \hat{f}_1(w) \hat{f}_2(w).$$

In addition, Parseval's identity holds for WHT; that is,

$$\sum_s f(s)^2 = 2^{-n} \sum_w \hat{f}(w)^2 = 2^{-n} \sum_w \hat{f}_1(w)^2 \hat{f}_2(w)^2. \qquad (7)$$

These WHT properties also hold for convolution of more-than two functions.

APPENDIX B: INEQUALITIES USED IN THIS PAPER

This appendix introduces the inequalities used in the study. See [11] for more details.

Theorem 8.1 (Data processing inequality). *Let A, B, and C be random variables that form a Markov chain $A \leftrightarrow B \leftrightarrow C$. Mutual information $I(A; B)$ and $I(A; C)$ always satisfy*

$$I(A; B) \geq I(A; C),$$

which is called a data processing inequality.

Theorem 8.2 (Jensen's inequality). *Let X be a random variable and let f be a function convex on the range of X. Jensen's inequality in the probability theory states that, if there exist finite $\mathbb{E}X$ and $\mathbb{E}f(X)$, then*

$$\mathbb{E}f(X) \geq f(\mathbb{E}X). \qquad (8)$$

Theorem 8.3 (Pinsker's inequality [23, 40]). *Let P and Q be probability distributions and let $D_{\mathrm{KL}}(P \parallel Q)$ be the Kullback–Leibler divergence to base 2 between P and Q. Pinsker's inequality states that*

$$\delta(P, Q) \leq \sqrt{\frac{\ln(2) D_{\mathrm{KL}}(P \parallel Q)}{2}},$$

where $\delta(P, Q)$ denotes the total variation distance (or statistical distance) between P and Q. If P and Q are discrete and have the probability mass functions p and q, respectively, Pinsker's inequality has the following alternative form:

$$\sum_a |p(a) - q(a)| \leq \sqrt{2 \ln(2) D_{\mathrm{KL}}(P \parallel Q)}. \qquad (9)$$

REFERENCES

[1] Josep Balasch, Benedikt Gierlichs, Vincent Grosso, Oscar Reparaz, and François-Xavier Standaert. 2014. On the Cost of Lazy Engineering for Masked Software Implementations. In *International Conference on Smart Card Research and Advanced Applications (Lecture Notes in Computer Science)*, Vol. 8968. 64–81.

[2] Gilles Barthe, Sonia Balaïd, Gaëtan Cassiers, Pierre-Alan Fouque, Benjamin Grégoire, and François-Xavier Standaert. 2019. maskVerif: Automated Verification of Higher-Order Masking in Presence of Physical Defaults. In *European Symposium on Research in Computer Security (ESORICS) (Lecture Notes in Computer Science)*, Vol. 11735. Springer, 300–318.

[3] Gilles Barthe, Sonia Balaïd, Fraonçois Dupressoir, Pierre-Alan Fouque, Benjamin Grégoire, Pierre-Yves Strub, and Rébecca Zucchini. 2016. Strong Non-Interference and Type-Directed Higher-Order Masking. In *ACM SIGSAC Conference on Computer Communications Security*. 116–129.

[4] Gilles Barthe, François Dupressoir, Sebastian Faust, Benjamin Grégoire, François-Xavier Standaert, and Pierre-Yves Strub. 2017. Parallel Implementations of Masking Schemes and the Bounded Moment Leakage Model. In *Advances in Cryptology—Eurocrypt 2017 (Lecture Notes in Computer Science)*, Vol. 10210. 535–556.

[5] Gilles Barthe, Marc Gourjon, Benjamin Grégoire, Maximilian Orlt, Clara Paglialonga, and Lars Porth. 2021. Masking in Fine-Grained Leakage Models: Construction, Implementation and Verification. *IACR Transactions on Cryptographic Hardware and Embedded Systems* 2021, 2 (2021), 189–228.

[6] Francesco Berti, Chun Guo, Olivier Pereira, Thomas Peters, and François-Xavier Standeart. 2020. TEDT, a Leakage-Resist AEAD Mode for High Physical Security Applications. *IACR Transactions on Cryptographic Hardware and Embedded Systems (TCHES)* 1 (2020), 256–320.

[7] Olivier Bronchain and François-Xavier Standeart. 2021. Breaking Masked Implementations with Many Shares on 32-bit Software Platforms: or When the Security Order Does Not Matter. *IACR Transactions on Cryptographic Hardware and Embedded Systems* 3 (2021), 202–234.

[8] Nicolas Bruneau, Sylvain Guilley, Annelie Heuser, and Olivier Rioul. 2014. Mask will Fall Off: Higher Order Optimal Distinguishers. In *Advances in Cryptology—ASIACRYPT 2020 (Lecture Notes in Computer Seience)*, Vol. 8874. 344–365.

[9] Suresh Chari, Charanjit S. Jutla, Josyula R. Rao, and Pankaj Rohatgi. 1999. Towards Sound Approaches to Counteract Power-Analysis Attacks. In *Advances in Cryptology — CRYPTO' 99*, Michael Wiener (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 398–412.

[10] Wei Cheng, Yi Liu, Sylvain Guilley, and Olivier Rioul. 2021. Attacking Masked Cryptographic Implementations: Information-Theoretic Bounds. arXiv:2105.07436. (2021). https://arxiv.org/abs/2105.07436.

[11] Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory, 2nd Edition*. Wiley-Interscience, 605 Third Avenue New York, NY, United States.

[12] Joan Daemen. 2017. Changing of the Guards: A Simple and Efficient Method for Achieving Uniformity in Threshold Sharing. In *International Conference on Cryptographic Hardware and Embedded Systems (Lecture Notes in Computer Science)*, Vol. 10529.

[13] Eloi de Chérisey, Sylvain Guilly, Olivier Rioul, and Pablo Piantanida. 2019. Best Information is Most Successful: Mutual Information and Success Rate in Side-Channel Analysis. *IACR Transactions on Cryptographic Hardware and Embedded Systems* 2019, 2 (2019), 49–79.

[14] Thomas de Cnudde, Begül Bilgin, Benedikt Gierlichs, Ventzislav Nikov, Svetla Nikova, and Vincent Rijmen. 2017. Does Coupling Affect the Security of Masked Implementations?. In *International Workshop on Constructive Side-Channel Analysis and Secure Design (Lecture Notes in Computer Science)*, Vol. 10348. 1–18.

[15] Thomas De Cnudde, Maik Ender, and Amir Moradi. 2018. Hardware Masking, Revisited. *IACR Transactions on Cryptographic Hardware and Embedded Systems* 2 (2018), 123–148.

[16] Lauren De Meyer, Begül Bilgin, and Reparaz Oscar. 2019. Consolidating Security Notions in Hardware Masking. *IACR Transactions on Cryptographic Hardware and Embedded Systems* 2019, 3 (2019), 119–147.

[17] Christoph Dobraunig, Maria Eichlseder, Stefan Mangard, Florian Mendel, and Thomas Unterluggauer. 2017. ISAP—Towards Side-Channel Secure Authenticated Encryption. *IACR Transactions on Symmetric Cryptology* 1 (2017), 80–105.

[18] Christoph Dobraunig, Maria Eichlseder, Florian Mendel, and Martin Schläffer. 2021. Ascon: Lightweight Authentication & Hashing. (2021). https://ascon.iaik.tugraz.at/index.html.

[19] Alexandre Duc, Stefan Dziembowski, and Sebastian Faust. 2014. Unifying Leakage Models: From Probing Attacks to Noisy Leakage. In *Advances in Cryptology—EUROCRYPT 2014 (Lecture Notes in Computer Science)*, Vol. 8441. Springer, 423–440.

[20] Alexandre Duc, Sebastian Faust, and François-Xavier Standaert. 2015. Making Masking Security Proofs Concrete: Or How to Evaluate the Security of Any Leakage Device. In *Advances in Cryptology—EUROCRYPT 2015 (Lecture Notes in Computer Science)*, Vol. 9056. Springer, 401–429.

[21] Stefan Dziembowski, Sebastian Faust, and Maciej Skórski. 2015. Optimal Amplification of Noisy Leakages. In *Theory of Cryptography Conference (TCC) (Lecture Notes in Computer Science)*, Vol. 9563. 291–318.

[22] Sebastian Faust, Vincent Grosso, Santos Merino Del Pozo, Clara Paglialonga, and François-Xavier Standaert. 2018. Composable Masking Schemes in the Presence of Physical Defaults & the Robust Probing Model. *IACR Transactions on Cryptographic Hardware and Embedded Systems* 3 (2018), 89–120.

[23] Alexei A Fedotov, Peter Harremoës, and Flemming Topsoe. 2003. Refinements of Pinsker's inequality. *IEEE Transactions on Information Theory* 49, 6 (2003), 1491–1498. https://doi.org/10.1109/TIT.2003.811927

[24] Si Gao, Ben Marshall, Dan Page, and Elisabeth Oswald. 2020. Share-slicing: Friend or Foe? *IACR Transactions on Cryptographic Hardware and Embedded Systems* 2020, 1 (2020), 152–174.

[25] Benedikt Gierlichs, Lejla Batina, Pim Tuyls, and Bart Preneel. 2008. Mutual Information Analysis: A Generic Side-Channel Distinguisher. In *International Conference on Cryptographic Hardware and Embedded Systems (Lecture Notes in*

Computer Science), Vol. 5154.

[26] Hannes Gross and Stefan Mangard. 2017. Recosiling $d + 1$ Masking in Hardware and Software. In *International Conference on Cryptographic Hardware and Embedded Systems (Lecture Notes in Computer Science)*, Vol. 10529. Springer.

[27] Hannes Gross, Stefan Mangard, and Thomas Korak. 2016. Domain-Oriented Masking: Compact Masked Hardware Implementations with Arbitrary Protection Order. In *ACM Workshop on Theory of Implementation Security*. 3.

[28] Sylvain Guilley, Annelie Heuser, and Olivier Rioul. 2015. A Key to Success. In *Progress in Cryptology – INDOCRYPT 2015 (Lecture Notes in Computer Science)*, Alex Biryukov and Vipul Goyal (Eds.). Springer International Publishing, 270–290.

[29] Annelie Heuser, Olivier Rioul, and Sylvain Guilley. 2014. Good Is Not Good Enough: Deriving Optimal Distinguishers from Communication Theory. In *International Workshop on Cryptographic Hardware and Embedded Systems (Lecture Notes in Computer Science)*, Vol. 8731. 55–74.

[30] Yuval Ishai, Amit Sahai, and David Wagner. 2003. Private Circuits: Securing Hardware againgst Probing Attacks. In *Advances in Cryptology—CRYPTO 2003 (Lecture Notes in Computer Science)*, Vol. 2729. Springer, 461–481.

[31] Akira Ito, Kotaro Saito, Rei Ueno, and Naofumi Homma. 2021. Imbalanced Data Problems in Deep Learning-Based Side-Channel Attacks: Analysis and Solution. *IEEE Transactions on Forensics and Security* (2021). DOI: 10.1109/TIFS.2021.3092050.

[32] Akira Ito, Rei Ueno, and Naofumi Homma. 2021. Toward Optimal Deep-Learning Based Side-Channel Attacks: Probability Concentration Inequality Loss and Its Usage. Cryptology ePrint Archive, Report 2021/1216. (2021). https://ia.cr/2021/1216.

[33] David Knichel, Pascal Sasdrich, and Amir Moradi. 2020. SILVER—Statistical Independence and Leakage Verification. In *Advances in Cryptology—ASIACRYPT 2020 (Lecture Notes in Computer Seience)*, Vol. 12491. 787–816.

[34] Victor Lomné, Emmanuel Prouff, Matthieu Rivain, Thomas Roche, and Adrian Thillard. 2014. How to Estimate the Success Rate of Higher-Order Side-Channel Attacks. In *International Workshop on Cryptographic Hardware and Embedded Systems (Lecture Notes in Computer Science)*, Vol. 8731. 35–54.

[35] Stefan Mangard. 2004. Hardware Countermeasures against DPA – A Statistical Analysis of Their Effectiveness. In *Topics in Cryptology – CT-RSA 2004 (Lecture Notes in Computer Science)*, Tatsuaki Okamoto (Ed.). Springer, Berlin, Heidelberg, 222–235.

[36] Stefan Mangrad, Norbert Pramstaller, and Elisabeth Oswald. 2005. Successfully Attacking Masked AES Hardware Implementations. In *Workshop on Cryptographic Hardware and Embedded Systems (Lecture Notes in Computer Science)*, Vol. 3659. Springer, 157–171.

[37] Loïc Masure, Cécile Dumas, and Emmanuel Prouff. 2020. A Comprehensive Study of Deep Learning for Side-Channel Analysis. *IACR Transactions on Cryptographic Hardware and Embedded Systems (TCHES)* 1 (2020), 348–375.

[38] Nicolai Müller, David Knichel, Pascal Sasdrich, and Amir Moradi. 2022. Transitional Leakage in Theory and Practice: Unveiling Security Flaws in Masked Circuits. *IACR Transactions on Cryptographic Hardware and Embedded Systems* 2 (2022), 266–288.

[39] Svetla Nikova, Vincent Rijmen, and Martin Schläffer. 2011. Secure Hardware Implementation of Nonlinear Functions in the Presence of Glitches. *Journal of Cryptology* 24, 2 (2011), 292–321.

[40] Mark S Pinsker and Amiel Feinstein. 1964. *Information and information stability of random variables and processes*. Vol. 12. Holden-Day San Francisco.

[41] Emmanuel Prouff and Matthieu Rivain. 2013. Masking against Side-Channel Attacks: A Formal Security Proof. In *Advances in Cryptology – EUROCRYPT 2013*, Thomas Johansson and Phong Q. Nguyen (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 142–159.

[42] Mathieu Renauld, François-Xavier Standeart, Nicolas Veyrat-Charvillon, Dina Kamel, and Denis Flandre. 2011. A Formal Study of Power Variability Issues and Side-Channel Attacks for Nanoscale Devices. In *Advances in Cryptology—Eurocrypt 2011 (Lecture Notes in Computer Science)*, Vol. 6632. 109–128.

[43] Oscar Reparaz, Begül Bilgin, Svelta Nikova, Benedikt Gierlichs, and Ingrid Verbauwhede. 2015. Consolidating Masking Schemes. In *Advances in Cryptology—CRYPTO 2015 (Lecture Notes in Computer Science)*, Vol. 9215. Springer, 764–783.

[44] Matthieu Rivain. 2008. On the Exact Success Rate of Side Channel Analysis in the Gaussian Model. In *Selected Areas in Cryptography—SAC 2008*. 165–183.

[45] M. Rivain. 2009. Differential Fault Analysis on DES Middle Rounds. *Cryptographic Hardware and Embedded Systems* 5747, 457–469.

[46] Tobias Schneider and Amir Moradi. 2015. Leakage Assesment Methodology—A Clear Roadmap for Side-Channel Evaluations. In *Workshop on Cryptographic Hardware and Embedded Systems (Lecture Notes in Computer Science)*, Vol. 9293. Springer, 495–513.

[47] A. Madura Shelton, Łukasz Chmielewski, Niels Samwel, Markus Wagner, Lejla Batina, and Yuval Yarom. 2021. Rosita++: Automatic Higher-Order Leakage Elimination from Cryptographic Code. In *ACM SIGSAC Conference on Computer Communications Security*. 685–699.

[48] A. Madura Shelton, Niels Samwel, Lejla Batina, Francesco Regazzoni, Markus Wagner, and Yuval Yarom. 2021. Rosita: Towards Automatic Elimination of Power-Analysis Leakage in Ciphers. In *Network and Distributed System Security Symposium, NDSS 2021*.

[49] François-Xavier Standeart, Tal G. Malkin, and Moti Yung. 2009. A Unified Framework for the Analysis of Side-Channel Key Recovery Attacks. In *Advances in Cryptology—Eurocrypt 2009 (Lecture Notes in Computer Science)*, Vol. 5479. 443–461.

[50] Adrian Thillard, Emmanuel Prouff, and Thomas Roche. 2013. Success through Confidence: Evaluating the Effectiveness of a Side-Channel Attack. In *Cryptographic Hardware and Embedded Systems – CHES 2013 (Lecture Notes in Computer Science)*, Guido Bertoni and Jean-Sébastien Coron (Eds.), Vol. 8086. Springer, Heidelberg, 21–36. https://doi.org/10.1007/978-3-642-40349-1_2

[51] Rei Ueno, Naofumi Homma, Sumio Morioka, and Takafumi Aoki. 2017. Automatic Generation of Formally-Proven Tamper-Resistant Galois-Field Multipliers Based on Generalized Masking Scheme. In *Design, Automation and Test in Europe Conference and Exhibition (DATE)*. IEEE, 978–983.

[52] Rei Ueno, Naofumi Homma, Sumio Morioka, and Takafumi Aoki. 2021. A Systematic Design Methodology of Formally Proven Side-Channel-Resistant Cryptographic Hardware. *IEEE Design & Test* 38, 3 (2021), 84–92.

[53] Greg Ver Steeg. 2021. Non-parametric Entropy Estimation Toolbox (NPEET). (2021). https://github.com/gregversteeg/NPEET.

[54] Roman Vershynin. 2018. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press.

[55] Gabriel Zaid, Lilian Bossuet, François Dassance, Amaury Habrard, and Alexandre Venelli. 2021. Ranking Loss: Maximizing the Success Rate in Deep Learning Side-Channel Analysis. *IACR Transactions on Cryptographic Hardware and Embedded Systems (TCHES)* 1 (2021), 25–55.

[56] Hailong Zhang and Wei Yang. 2021. Theoretical Estimation on the Success Rate of Asymptotic Higher Order Optimal Distinguisher. *Comput. J.* 64, 8 (2021), 1277–1292.

[57] Jiajia Zhang, Mengce Zheng, Jiehui Nan, Honggang Hu, and Nenghai Yu. 2020. A Novel Evaluation Metric for Deep Learning-Based Side Channel Analysis and Its Extended Application to Imbalanced Data. *IACR Transactions on Cryptographic Hardware and Embedded Systems (TCHES)* 1 (2020), 73–96.

[58] Ziyue Zhang, A. Adam Ding, and Yunsi Fei. 2020. A Fast and Accurate Guessing Entropy Estimation Algorithm for Full-key Recovery. *IACR Transactions on Cryptographic Hardware and Embedded Systems* (March 2020), 26–48.