IACR Transactions on Cryptographic Hardware and Embedded Systems
ISSN 2569-2925, Vol. 2022, No. 4, pp. 228–254. DOI:10.46586/tches.v2022.i4.228-254

Perceived Information Revisited

New Metrics to Evaluate Success Rate of Side-Channel Attacks

Akira Ito[1], Rei Ueno[2] and Naofumi Homma[2]

[1] NTT Social Informatics Laboratories, Nippon Telegraph and Telephone Corporation, 3–9–11 Midori-cho, Musashino-shi, Tokyo, 180-8535, Japan
akira.ito.as@hco.ntt.co.jp

[2] Tohoku University, 2–1–1 Katahira, Aoba-ku, Sendai-shi, Miyagi, 980-8577, Japan
rei.ueno.a8@tohoku.ac.jp, naofumi.homma.c8@tohoku.ac.jp

Abstract. In this study, we present new analytical metrics for evaluating the performance of side-channel attacks (SCAs) by revisiting the perceived information (PI), which is defined using cross-entropy (CE). PI represents the amount of information utilized by a probability distribution that determines a distinguishing rule in SCA. Our analysis partially solves an important open problem in the performance evaluation of deep-learning based SCAs (DL-SCAs) that the relationship between neural network (NN) model evaluation metrics (such as accuracy, loss, and recall) and guessing entropy (GE)/success rate (SR) is unclear. We first theoretically show that the conventional CE/PI is non-calibrated and insufficient for evaluating the SCA performance, as it contains uncertainty in terms of SR. More precisely, we show that an infinite number of probability distributions with different CE/PI can achieve an identical SR. With the above analysis result, we present a modification of CE/PI, named effective CE/PI (ECE/EPI), to eliminate the above uncertainty. The ECE/EPI can be easily calculated for a given probability distribution and dataset, which would be suitable for DL-SCA. Using the ECE/EPI, we can accurately evaluate the SR through the validation loss in the training phase, and can measure the generalization of the NN model in terms of SR in the attack phase. We then analyze and discuss the proposed metrics regarding their relationship to SR, conditions of successful attacks for a distinguishing rule with a probability distribution, a statistic/asymptotic aspect, and the order of key ranks in SCA. Finally, we validate the proposed metrics through experimental attacks on masked AES implementations using DL-SCA.

Keywords: Side-channel analysis · Deep learning · Optimal distinguisher · Success rate · Perceived information

1 Introduction

1.1 Background

Deep-learning based side-channel attack. Deep-learning based side-channel attacks (DL-SCAs) on cryptographic modules have been increasingly emerged in recent years [MHM14, CDP17, HHGG20, RWPP21, UXT+22]. DL-SCA is a profiling attack which consists of two phases: profiling and attack. In the profiling phase, an attacker obtains side-channel traces from profiling device(s) with similar leakage characteristics as the target device, then trains a neural network (NN) model representing the leakage characteristics. In the attack phase, the attacker utilizes the trained NN model to estimate the secret key from the target device's side-channel leakage. Compared with conventional profiling attacks, such as template attacks [CRR02], DL-SCA can achieve a higher attack performance (*e.g.*, key recovery capability) even on implementations with SCA countermeasures, such as masking

Received: 2022-04-15 Accepted: 2022-06-15 Published: 2022-08-31

and random delay. Thus, it has become necessary to develop an assessment methodology to evaluate DL-SCA threats because of the increasing number of cryptographic devices that are now operated in a scenario where an attacker can perform a profiling, such as Internet of Things applications. Actually, several literatures have showed/discussed the possibility and potential of profiling attacks in the real scenarios such as [OP11, DK18, WVdHG+20].

Performance evaluation of DL-SCA. Performance evaluation in DL-SCA has an important open problem. During the validation/test phase of a typical DL, the performance of an NN model has been evaluated through using metrics such as the accuracy, loss, and recall. In contrast, SCA performance has frequently been evaluated using the guessing entropy (GE) and success rate (SR) [SMY09]. However, the relationship between the evaluation metrics is unclear so far. In fact, it is reported that the DL metrics sometimes contradict the SCA performance [PHJ+19]. For example, in an extreme case, an NN model with an accuracy of 0% could succeed in the key recovery in the attack phase. In addition, in terms of cross-entropy (CE) loss, an overfitting model sometimes outperforms non-overfitting models. The mismatch between the metrics leads to a non-negligible computational cost for the empirical evaluation of the attack performance (*i.e.*, GE/SR) of a model, and makes it difficult to determine the timing of generalization and early stopping. Addressing the problem, we could easily evaluate the SR of a given NN model and determine its generalization of a model in terms of SR.

Perceived information. Perceived information (PI) was proposed in 2011 [RSVC+11]. Conceptually, PI represents the amount of information between secret intermediate variable and side-channel leakage exploited by a probability distribution that determines a distinguishing rule. Because mutual information between the secret intermediate variable and side-channel leakage is essential for SCA and can be used to evaluate the SR [GBTP08, DFS15, dCGRP19], PI might be useful for evaluating the SR given a probability distribution that determines a distinguishing rule (*e.g.*, a trained NN in DL-SCA). It was shown in [MDP20] that the PI of a probability distribution r was a lower-bound of mutual information and was defined using CE as $H(Z) - \mathrm{CE}(r)$, where $H(Z)$ is the entropy of secret intermediate variable Z and $\mathrm{CE}(r)$ is the CE between the true probability distribution p and r. Nevertheless, there was still a mismatch between the attack performance of a model and its PI, as the model could sometimes succeed in the key recovery in DL-SCA even when the CE loss was large, implying PI looked too small for attack success. In fact, it was mentioned in [BHM+19] and is confirmed also in this paper that PI-based SR evaluation sometimes underestimates the actual attack performance. This implies that there would still be a mismatch among the information-theoretic metrics (*i.e.*, PI), CE loss function, and SCA attack performance (*i.e.*, SR).

1.2 Our contribution

This study revisits the PI to analyze and discuss the above mismatch. We first review the Ito, Ueno, and Homma's theorem [IUH21], which states that a probability distribution with the minimum CE *sufficiently* but *not necessarily* provides an optimal distinguisher. We then derive the relationship between the theorem and the CE/PI as Proposition 3 and show that the existing definition of CE/PI is insufficient in SCA. Concretely, we prove that the existing PI contains an uncertainty in terms of SR; that is, an infinite number of probability distributions with different CE/PI yield distinct distinguishing rules with an identical SR. This shows that the conventional CE/PI definition is insufficient as a metric for evaluating the SCA performance.

With the above analysis result, we present a modification of CE/PI, named effective CE/PI (ECE/EPI) to eliminate the above uncertainty. ECE and EPI are (hypothetically)

respectively given by an infimum of CE and a supremum of PI of the probability distribution that can achieve an SR. The proposed metrics can be easily calculated for a given probability distribution and dataset as described in Section 4.4, which would be suitable for DL-SCA. The use of EPI makes it possible to perform a more accurate SR evaluation through the ECE during NN training in DL-SCA by a combination with an inequality developed by de Chérisey *et al.* [dCGRP19]. Note that an SR upper-bound is closely related to the lower-bound of the number of traces required for the attack success; and therefore, the bounds are used in a quantitative evaluation metric of SCA [SMY09]. The proposed metrics can also be used to measure the generalization of an NN model in terms of SR during the attack phase. We analyze and discuss the proposed metrics in terms of their relationship to SR, a statistic/asymptotic aspect, and conditions of successful attacks for a distinguishing rule with a probability distribution. In addition, we provide an analysis on the order of key ranks in SCA to show the suitability of ECE/EPI for SR evaluation. Finally, we validate the proposed metrics through an experimental attack on masked AES implementations using DL-SCA.

We suppose that the proposed approach would be especially helpful for evaluators and (white) attackers as it easily evaluates the attack performance of a model. This indicates that it would be useful for, for example, early stopping to maximize the SR and comparison of two (or more) models to determine which model is superior in DL-SCA. The experimental attack also validates this aspect. For example, in the experimental attack on masked hardware, the SR evaluation using the proposed metrics/method takes at most 0.53 seconds even using 100,000 test traces, whereas a common empirical SR evaluation requires far longer time, which may be in an order of minutes with 100,000 traces. Note that the computation time corresponds to one SR evaluation at an epoch; in practice, we should perform the computation for every epoch, which indicates that the usage of EPI would yield a significant reduction of computation time. Thus, EPI can also contribute to the SR evaluation in practical aspects, in addition to the theoretical contribution of this paper.

1.3 Paper organization

The remainder of this paper is organized as follows: Section 2 introduces the mathematical notation and reviews the previous studies on DL-SCA and PI. Section 3 derives the relation between PI and SR from a probability-theoretical perspective. Section 4 proposes, analyzes, and discusses a new information-theoretical metric named ECE/EPI. Section 5 demonstrates the validity of the proposed metric through experimental attacks on masked AES implementations. Finally, Section 6 concludes this study.

2 Preliminaries

2.1 Notation

A calligraphic letter (*e.g.*, $\mathcal{X}$) represents a set; an uppercase variable (*e.g.*, X) represents a random variable over the corresponding set (*i.e.*, $\mathcal{X}$ for X); and a lowercase variable (*e.g.*, x) is an element of the corresponding set, if it is defined otherwise. Pr denotes a probability measure. Throughout this paper, p denotes to the true density or mass function; q denotes the probability density or mass function represented by an NN[1]. For example, the true probability mass function of discrete random variables X, Y is $p_{X,Y}(x, y) = \Pr(X = x, Y = y)$. We may omit the subscripted random variables if the

[1]In this paper, the probability mass functions or density functions for any random variables exist by making appropriate assumptions because we will focus only on discrete, continuous random variables, or a mixture of them.

random variables of the probability distribution are obvious. For example, we may simply write $p(x, y)$ instead of $p_{X,Y}(x, y)$. In addition, we may write a conditional probability distribution represented by an NN with parameter θ by $q_\theta \triangleq q_{Z|\boldsymbol{X}}(\cdot \mid \cdot; \theta)$. The expectation is denoted by $\mathbb{E}$. For example, $\mathbb{E}_X f(X)$ denotes the expectation of $f(X)$, where $f : \mathcal{X} \to \mathbb{R}$ is a function. The conditional probability distribution is denoted by $p_{X|Y}(x \mid y) = p(x \mid y)$, and $\mathbb{E}[f(X, Y) \mid Y = y]$ denotes the expected value. Finally, log and ln denote the binary and natural logarithms.

Let $\boldsymbol{X}$ denote a random variable of the side-channel trace. Side-channel traces are represented as a multidimensional real vector $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^{m_t}$, where $m_t \in \mathbb{N}$ is the number of sample points. This study focuses on SCAs on block ciphers, particularly AES. Let n_k denote the bit length of the partial key, and let n_t denote the bit length of the partial plaintext and ciphertext. The secret intermediate value is denoted as $z = g(k, t) \in \mathcal{Z} = \{0, 1\}^{n_z}$, where g is a selection function[2], n_z denotes the bit length of z, $k \in \mathcal{K} = \{0, 1\}^{n_k}$ is a key, and $t \in \mathcal{T} = \{0, 1\}^{n_t}$ is public information such as plaintexts and ciphertexts. Their random variables are also defined in the aforementioned manner. Here, let K denote the random variable of the correct key, and k^* denote the correct key value. T and K are assumed to have uniform distributions. If we require to specify the key value for Z, we write $Z^{(k)} = g(k, T)$.

In this study, the conditional probability distribution between the secret intermediate variable Z and side-channel leakage $\boldsymbol{X}$ (*e.g.*, $p_{Z|\boldsymbol{X}}$, q_θ) plays an essential role. For simplicity, we assume that every conditional probability distribution $r_{Z|\boldsymbol{X}}$ satisfies $-\mathbb{E} \log r_{Z|\boldsymbol{X}}(Z \mid \boldsymbol{X}) < \infty$. This condition ensures that the cross entropy of every distribution exist. Let $\mathcal{R}$ be a set of all the conditional probabilities such that, for every $r_{Z|\boldsymbol{X}} \in \mathcal{R}$, $\forall z \in \mathcal{Z}, \boldsymbol{x} \in \mathcal{X}; r_{Z|\boldsymbol{X}}(z \mid \boldsymbol{x}) > 0$ holds, and $\forall z_1, z_2 \in \mathcal{Z}; z_1 \neq z_2 \Rightarrow r_{Z|\boldsymbol{X}}(z_1 \mid \boldsymbol{X}) \neq r_{Z|\boldsymbol{X}}(z_2 \mid \boldsymbol{X})$ holds almost surely. Because q_θ probably meets these two conditions in many cases, the conditional probability of model q_θ is contained in the set $\mathcal{R}$ in practice. The first condition is natural because the NN model cannot take a zero value if Softmax function is used as the activation function of its last layer. On the other hand, although there would exist parameters which do not satisfy the second condition, it would be highly unlikely that such parameters are selected during learning because of the randomness of the learning algorithms[3]. Note that the true distribution $p_{Z|\boldsymbol{X}}$ is necessarily not contained in the set $\mathcal{R}$. For example, there exist different z_1 and z_2 such that $\Pr(p(z_1 \mid \boldsymbol{X}) = p(z_2 \mid \boldsymbol{X})) > 0$ holds if the leakage model is Hamming weight, and there is no noise in the traces (*e.g.*, $(z_1, z_2) = (1, 2)$). Even in this case, we assume that $q_\theta \in \mathcal{R}$ holds because of the randomness of the learning algorithms.

For the sake of simplicity, we assume that the distribution of Z is independent of the key used. This assumption is closely related to the key-independence condition [IUH21], which states that the key can be fixed during the profiling. Many practical selection functions are proven to satisfy this condition. For example, a typical selection function for 16 bytes of software AES implementation (*i.e.*, $Z = \mathrm{Sbox}(K \oplus T)$ and its Hamming weight) satisfies this condition.

2.2 Overview of DL-SCA

The DL-SCA has two phases: profiling and attack. During the profiling phase, we train a model to approximate the conditional distribution as the device leakage characteristics. Let $\mathcal{S}_p = \{ (\boldsymbol{X}_i, Z_i) \mid 1 \leq i \leq m_{\mathrm{tr}} \}$ be a training dataset used in the profiling phase, $\boldsymbol{X}_i$ denotes the side-channel trace (*i.e.*, power consumption or electromagnetic radiation) of

[2]In this paper, we assume that g does not consist in a *leakage function* (*e.g.*, Hamming weight), as we focus on the probability distribution $r_{Z|\boldsymbol{X}}$.

[3]Formally, this can be stated as follows. Let $\mathcal{S}_p$ be a training dataset, and let $M : \mathcal{S}_p \mapsto \theta$ be a learning algorithm which is a randomized function. Note that a learned parameter $\hat{\theta} \xleftarrow{\$} M(\mathcal{S}_p)$ is regarded as a random variable. This paper assumes that $q_{\hat{\theta}} \in \mathcal{R}$ holds almost surely (*i.e.*, $\Pr(q_{\hat{\theta}} \in \mathcal{R}) = 1$))

the i-th observation, Z_i denotes the corresponding intermediate value, and $|\mathcal{S}_p| = m_{\mathrm{tr}}$ is the number of traces used in the profiling phase. We assume that $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_{m_{\mathrm{tr}}}$ and $Z_1, Z_2, \ldots, Z_{m_{\mathrm{tr}}}$ are independent and identically distributed (i.i.d) random variables, respectively. Let θ denote the NN model parameter. The goal of the profiling phase is to estimate the optimal model parameter $\hat{\theta}$ using the training dataset $\mathcal{S}_p$. This optimal parameter is usually given as the solution to the minimization problem of the CE loss function, defined as

$$\mathrm{CE}(q_\theta) = -\mathbb{E}_{Z,\boldsymbol{X}} \log q(Z \mid \boldsymbol{X}; \theta) = -\int \sum_z p_{Z,\boldsymbol{X}}(z, \boldsymbol{x}) \log q(z \mid \boldsymbol{x}; \theta)\, d\boldsymbol{x}, \tag{1}$$

where Z and $\boldsymbol{X}$ are the random variables of a label z and trace $\boldsymbol{x}$, respectively, and q_θ represents the conditional probability distribution represented by the NN with the parameter θ.

$\mathrm{CE}(q_\theta)$ in Equation (1) takes the minimum value if and only if $p = q_\theta$ [Bis06, GBC16]. Note that, depending on the hyperparameter and p, it is not generally guaranteed that there exists a model parameter such that $p = q_\theta$. We can obtain a model that approximates the true distribution p if we determine the optimal parameter $\hat{\theta}$ that makes $\mathrm{CE}(q_{\hat{\theta}})$ sufficiently small; however, we cannot calculate Equation (1) because it contains the integral and summation of the unknown probability distribution p. Therefore, in general, we approximate $\mathrm{CE}(q_\theta)$ using the training data $\mathcal{S}_p$ as follows:

$$\mathrm{CE}(q_\theta) \approx L(q_\theta) = -\frac{1}{m_{\mathrm{tr}}} \sum_{i=1}^{m_{\mathrm{tr}}} \log q(Z_i \mid \boldsymbol{X}_i; \theta). \tag{2}$$

The approximated CE in Equation (2) is called negative log-likelihood (NLL). The NLL is expected to converge in probability to $\mathrm{CE}(q_\theta)$ as $m_{\mathrm{tr}} \to \infty$ for fixed q_θ.

During the attack phase, we estimate the secret key k^* of the target device using the trained model. Let $\mathcal{S}_a = \{\, (\boldsymbol{X}_j, T_j) \mid 1 \le j \le m_{\mathrm{at}} \,\}$ be a dataset used during the attack phase, where $|\mathcal{S}_a| = m_{\mathrm{at}}$ is the number of traces, $\boldsymbol{X}_j$ is the side-channel trace at the j-th observation, and T_j is the corresponding plaintext or ciphertext. During the attack phase, we calculate the NLL for each hypothetical key candidate $k \in \mathcal{K}$ using the intermediate value $Z_j^{(k)}$ calculated from T_j as

$$L^{(k)}(q_{\hat{\theta}}) = -\frac{1}{m_{\mathrm{at}}} \sum_{j=1}^{m_{\mathrm{at}}} \log q(Z_j^{(k)} \mid \boldsymbol{X}_j; \hat{\theta}).$$

Following that, the correct key is estimated to be the key candidate with the smallest NLL value. This is equivalent to approximately computing and comparing

$$\mathrm{CE}^{(k)}(q_\theta) = -\mathbb{E} \log q(Z^{(k)} \mid \boldsymbol{X}; \hat{\theta}),$$

for each key candidate k.

In the following, for a simplified notation, we denote the number of traces for the attack phase by m, instead of m_{at}. As well, the number of traces for validation/test is also simply denoted by m as a validation/test corresponds to an attack phase.

Instead of CE, some loss functions have been presented to improve the learning cost and/or the attack performance of NN. In [ZZN+20], Zhang *et al.* presented the cross entropy ratio (CER), and showed that it is useful for improving the attack performance especially when the training and test datasets suffer from an imbalanced data problem, as also analyzed in [ISUH21]. In [ZBD+21], Zaid *et al.* presented the ranking loss (RkL), the usage of which can suppress the approximation error and can make the convergence faster. As investigated in [KWPP21], such loss functions dedicated to DL-SCA can yield a high attack performance, although a common CE can be a good option in most cases. Thus, it is worth investigating the loss function dedicated to DL-SCA.

2.3 SCA evaluation metrics

To evaluate the performance of (DL-)SCA, the SR and GE are commonly used as quantitative metrics during the attack phase. The SR and GE with m traces during the attack phase are represented as

$$\begin{aligned} \mathrm{SR}_m &= \Pr(\mathrm{rank}(k^*, m, q_\theta) = 1), \\ \mathrm{GE}_m &= \mathbb{E}\mathrm{rank}(k^*, m, q_\theta), \end{aligned}$$

respectively [SMY09]. In the case of DL-SCA, the rank of correct key is defined as

$$\mathrm{rank}(k^*, m, q_\theta) = 1 + \sum_{k \in \mathcal{K} \setminus \{k^*\}} \mathbb{1}_{\left\{L^{(k^*)}(q_\theta) \geq L^{(k)}(q_\theta)\right\}},$$

where $\mathbb{1}$ is the indicator function.

In [dCGRP19], de Chérisey *et al.* showed that the SR is upper-bounded using a conditional mutual information between the secret intermediate variable Z and side-channel leakage $\boldsymbol{X}$ given plaintext/ciphertext T, denoted by $I(Z^m; \boldsymbol{X}^m \mid T^m)$, where $Z^m = (Z_1, Z_2, \ldots, Z_m)$, $\boldsymbol{X}^m = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_m)$, and $T^m = (T_1, T_2, \ldots, T_m)$. In [IUH22], Ito, Ueno and Homma proved that $I(Z^m; \boldsymbol{X}^m \mid T^m) \leq mI(Z; \boldsymbol{X})$. According to de Chérisey *et al.* and Ito, Ueno, and Homma, SR is upper-bounded as

$$\xi(\mathrm{SR}_m) \leq mI(Z; \boldsymbol{X}), \tag{3}$$

where $\xi : [0, 1] \to \mathbb{R}_+$ denotes a function defined as

$$\xi(\mathrm{SR}_m) = H(K) - (1 - \mathrm{SR}_m) \log(2^{n_k} - 1) - H_2(\mathrm{SR}_m), \tag{4}$$

where $H(K)$ is the entropy of K (here, $H(K) = n_k$ if K is the uniform distribution on $\{0, 1\}^{n_k}$) and H_2 is the binary entropy function. Intuitively, $\xi(\mathrm{SR})$ represents the amount of information required for key recovery for a given SR. For example, if an attacker attempts key recovery with $\mathrm{SR}_m = 1$, the attacker requires n_k-bit information as represented by $\xi(1) = n_k$. In contrast, if the attacker has no advantage on the key estimation (that is, $\mathrm{SR}_m = 1/2^{n_k}$), the attacker requires zero bit information about the secret key as represented by $\xi(1/2^{n_k}) = 0$. Inequality (3) states that this amount of information is upper-bounded by mutual information. To achieve a desired SR, Inequality (3) states that the attacker requires to obtain $mI(Z; \boldsymbol{X})$ bit information through the observation of m side-channel traces.

Related to SR evaluation through the validation loss, Zaid *et al.* showed that RkL is an SR lower-bound [ZBD+21]. Although its usage has some advantages in DL-SCA (*e.g.*, a suppression of the approximation error and a faster convergence), RkL-based SR evaluation requires the computational cost as high as the conventional empirical evaluations. RkL can be evaluated only experimentally/empirically, but not analytically, because RkL is derived by approximating an indicator function in the GE as a binary loss function [IUH21]. This indicates that RkL-based SR evaluation includes the conventional empirical SR evaluation. For the assessment of DL-SCA performance, it is worth studying how to evaluate the SR through the validation loss with less costs.

2.4 Optimal distinguisher

SCA can be formulated using a distinguisher, which is denoted by a function $d\colon \mathcal{X}^m \times \mathcal{T}^m \to \mathcal{K}$. A distinguisher calculates the ranks of each key candidate using a score function from side-channel trace, and estimates the correct key as the candidate with the highest score. For example, correlation power analysis (CPA) uses Pearson's correlation coefficient as the score. DL-SCA uses the NLL as the score. An optimal distinguisher is a distinguisher that maximizes the SR. The optimal distinguisher is formally defined as follows:

Definition 1 (Optimal distinguisher [HRG14])**.** For attack traces $\boldsymbol{X}^m = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_m)$ and inputs $T^m = (T_1, T_2, \ldots, T_m)$, the success rate of a distinguisher $d : \mathcal{X}^m \times \mathcal{T}^m \to \mathcal{K}$ is defined as $\mathrm{SR}_m(d) = \Pr(K = d(\boldsymbol{X}^m, T^m))$. A distinguisher d_{opt} is called optimal if we have $\mathrm{SR}_m(d_{\mathrm{opt}}) = \sup_d \mathrm{SR}_m(d)$.

According to [HRG14], an optimal distinguisher d_{opt} is given by

$$\begin{aligned} d_{\mathrm{opt}}(\boldsymbol{X}^m, T^m) &= \arg\max_k \Pr(K = k \mid \boldsymbol{X}^m, T^m) \\ &= \arg\max_k \sum_j \log p_{\boldsymbol{X}|T,K}(\boldsymbol{X}_j \mid T_j, k). \end{aligned}$$

In [IUH21], Ito, Ueno, and Homma proved that d_{opt} has another equivalent form given by the true conditional probability distribution of the secret intermediate variable Z given a side-channel leakage $\boldsymbol{X}$ (denoted by $p_{Z|\boldsymbol{X}}$), which suits to the DL-SCA. This indicates that the CE minimization in DL-SCA makes sense to achieve an optimal attack, as the goal of DL is usually to imitate the true conditional probability distribution through the CE loss minimization. However, in [IUH21], Ito, Ueno, and Homma also proved that an infinite number of probability distributions with a non-minimum CE provide distinct optimal distinguishers. Their theorem states that the true conditional probability distribution (*i.e.*, a probability distribution with the minimum CE) *sufficiently* but *not necessarily* provides an optimal distinguisher. Using the theorem, they stated that a probability distribution with a relatively high CE does not always make the SR low in the attack phase of DL-SCA, motivating them to propose a loss function (named Probability Concentration Inequality (PCI) loss), which is used to directly maximize the SR. We review their theorem to reveal the relationship between PI and SR in Section 3.

2.5 Perceived Information

The concept of PI was initially presented by Renauld *et al.* [RSVC+11]. PI is considered as an amount of information utilized by a probability distribution (*e.g.*, NN output) providing a distinguishing rule. Let $J_r(Z; \boldsymbol{X})$ denote the PI of a probability distribution r between the secret intermediate variable Z and side-channel leakage $\boldsymbol{X}$. $J_r(Z; \boldsymbol{X})$ is defined as

$$J_r(Z; \boldsymbol{X}) = H(Z) - \mathrm{CE}(r) \approx H(Z) - L(r). \tag{5}$$

PI is a lower-bound of mutual information, that is, it holds $J_r(Z; \boldsymbol{X}) \leq I(Z; \boldsymbol{X})$ for any distribution $r_{Z|\boldsymbol{X}}$ [MDP20, BHM+19]. The equality holds if and only if $r_{Z|\boldsymbol{X}}$ is equivalent to the true probability distribution $p_{Z|\boldsymbol{X}}$. $J_r(Z; \boldsymbol{X})$ is expected to be always non-negative, as it represents an amount of information. However, the original PI can take a negative value as mentioned and shown in [BHM+19]. In this paper, Section 3.2 states one of its reasons why PI can be negative.

Let $\mathrm{SR}_m(r)$ denote the SR of an attack using m traces and a distinguisher with $r_{Z|\boldsymbol{X}}$. According to the intuitive meanings of PI and ξ in Equation (4), $\mathrm{SR}_m(r)$ is expected to be upper-bounded by

$$\xi(\mathrm{SR}_m(r)) \leq m J_r(Z; \boldsymbol{X}), \tag{6}$$

similarly to Inequality (3). However, in practice, some counterexamples are found (like the experiment in this paper). a probability distribution with so large CE (that makes PI too small for the attack success with regard to Inequality (6)) sometimes can succeed in the key recovery. This indicates that the existing PI does not adequately represent the amount of information that can be used with the SR inequality (3) as mutual information. In this study, we specify one of the reasons and present a modification of PI to address this issue.

3 Uncertainty of CE/PI for SR evaluation

3.1 Review of Ito–Ueno–Homma theorem [IUH21]

Theorem 1 (Ito, Ueno, and Homma [IUH21]). *Let $r_{Z|\boldsymbol{X}}$ be a conditional probability distribution of the secret intermediate value Z given side-channel leakage $\boldsymbol{X}$. $r_{Z|\boldsymbol{X}}$ yields an optimal distinguisher if $\mathrm{CE}(r_{Z|\boldsymbol{X}})$ is minimum (i.e., $r = p$). However, $\mathrm{CE}(r_{Z|\boldsymbol{X}})$ is not always minimum if $r_{Z|\boldsymbol{X}}$ yields an optimal distinguisher.*

Theorem 1 is proven by two propositions: one states that the true conditional probability distribution (*i.e.*, a conditional probability distribution with the minimum CE) *sufficiently* yields an optimal distinguisher, and the other states that its inverse is false; that is, a conditional probability distribution with the minimum CE does *not necessarily* yields an optimal distinguisher.

Proposition 1 (CE minimization is sufficient for optimal distinguisher [IUH21]). *Let $r_{Z|\boldsymbol{X}}$ be a conditional probability distribution of Z given $\boldsymbol{X}$, and let d_r be a distinguisher defined as*

$$d_r(\boldsymbol{X}^m, T^m) = \arg\max_k \sum_{j=1}^{m} \log r_{Z|\boldsymbol{X}}(Z_j^{(k)} \mid \boldsymbol{X}_j).$$

The distinguisher $d_r(\boldsymbol{X}^m, T^m)$ is optimal if $r_{Z|\boldsymbol{X}}$ is equivalent to the true probability distribution $p_{Z|\boldsymbol{X}}$; that is, $d_p(\boldsymbol{X}^m, T^m)$ is an optimal distinguisher.

Note that, for a probability distribution of trained $|\mathcal{Z}|$-classification NN $q_{\hat{\theta}}$, the distinguisher $d_{q_{\hat{\theta}}}$ is equivalent to

$$d_{q_{\hat{\theta}}}(\boldsymbol{X}^m, T^m) = \arg\max_k \sum_{j=1}^{m} \log q_{Z|\boldsymbol{X}}(Z_j^{(k)} \mid \boldsymbol{X}_j; \hat{\theta}).$$

which would be the reasons why, in DL-SCA, we train an NN to approximate the true probability distribution $p_{Z|\boldsymbol{X}}$ and utilize the NLL for the key estimation during the attack phase. In the following, we always consider the distinguishing rule defined in Proposition 1 for a given probability distribution.

Before introducing Proposition 2, we review Lemma 1 followed by Corollary 1, which are crucial to the proof of Proposition 2.

Lemma 1 (A conversion of probability distribution with order of key ranks preserved [IUH21]). *Let*

$$r'_{Z|\boldsymbol{X}}(z \mid \boldsymbol{x}) = \frac{r_{Z|\boldsymbol{X}}(z \mid \boldsymbol{x})^\beta}{\sum_{z'} r_{Z|\boldsymbol{X}}(z' \mid \boldsymbol{x})^\beta},$$

where $r_{Z|\boldsymbol{X}} : \mathcal{Z} \times \mathcal{X} \to (0, 1]$ is a probability distribution, and β is a positive real number. Then, for all $k \in \mathcal{K}$ and $m \in \mathbb{N}$, $\mathrm{rank}(k, m, r) = \mathrm{rank}(k, m, r')$ holds.

Corollary 1. *For a given probability distribution $r_{Z|\boldsymbol{X}}$ and $\mathcal{S}_a$, the success rate SR_m and guessing entropy GE_m are invariant to the above conversion of probability distribution with any β.*

Lemma 1 guarantees that the conversions from r to r' do not change the SCA performance (*i.e.*, SR and GE). Note that the conditional distribution $r_{Z|\boldsymbol{X}}$ must be a positive real-valued function to hold Lemma 1. NN models satisfy this condition because they usually use a Softmax as the activation function of the last layer. Lemma 1 implies that an infinite number of such conversions exist because β is any positive real number. Using Lemma 1, Ito, Ueno, and Homma proved Proposition 2.

Proposition 2 (CE minimization is not necessary for optimal distinguisher [IUH21])**.** *Let d be a distinguisher for the attack phase, defined as*

$$d_r(\boldsymbol{X}^m, T^m) = \arg\max_k \sum_{j=1}^{m} \log r_{Z|\boldsymbol{X}}(Z_j^{(k)} \mid \boldsymbol{X}_j),$$

where $r_{Z|\boldsymbol{X}} : \mathcal{Z} \times \mathcal{X} \to (0, 1]$ is a conditional probability distribution. Even when the distinguisher d is optimal, $\inf_{r''} \mathrm{CE}(r'') = \mathrm{CE}(r_{Z|\boldsymbol{X}})$ does not necessarily hold.

3.2 Relation between CE/PI and SR

We then show the uncertainty of CE/PI in terms of SR evaluation using Lemma 1. In this study, we focus on the conversion of probability distribution used in Lemma 1. We first define the conversion notation.

Definition 2. Let $r_{Z|\boldsymbol{X}}$ be a conditional probability distribution. For any positive real number β, define a conversion of $r_{Z|\boldsymbol{X}}$ as

$$\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}](z \mid \boldsymbol{x}) = \frac{r_{Z|\boldsymbol{X}}(z \mid \boldsymbol{x})^\beta}{\sum_{z'} r_{Z|\boldsymbol{X}}(z' \mid \boldsymbol{x})^\beta}.$$

The application of $\mathscr{H}_\beta$ to a probability distribution is equivalent to the usage of Softmax with temperature for the activation function of output layer of an NN model. In the DL community, such a Softmax with temperature is used for emphasizing a label with the highest probability if $\beta > 1$ or for placing importance relatively on labels with small probability if $0 < \beta < 1$. It is known that the accuracy of an NN model is invariant to the temperature [GPSW17], which obviously indicates that, for one trace attack (*i.e.*, $m = 1$), the rank order, SR, and GE are also invariant to the temperature. Proposition 2 generalizes this fact to more-than one traces attack; the temperature is generally meaningless for distinguishing rules in terms of attack performance with any (finite) number of traces. Meanwhile, CE and PI are dependant on β. To analyze the dependency, we derive the limits of CE and PI of $\mathscr{H}_\beta[r]$ as $\beta \searrow 0$ and $\beta \to \infty$. For the derivation, we introduce Lemma 2.

Lemma 2. *Let $r_{Z|\boldsymbol{X}} \in \mathcal{R}$ be a conditional probability distribution, and let β be a positive real number. $\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}](Z|\boldsymbol{X})$ converges almost surely to 2^{-n_z} (i.e., uniform distribution over $\mathcal{Z}$) as $\beta \searrow 0$ and $\mathbb{1}_{\{Z=\arg\max_{z'} r(z'|\boldsymbol{X})\}}$ (i.e., one-hot distribution) as $\beta \to \infty$, where $\mathbb{1}_{\{Z=\arg\max_{z'} r(z'|\boldsymbol{X})\}}$ is measurable.*

Proof. First, we derive the limit of $\beta \searrow 0$ as follows:

$$\lim_{\beta \searrow 0} \mathscr{H}_\beta[r_{Z|\boldsymbol{X}}](z \mid \boldsymbol{x}) = \lim_{\beta \searrow 0} \frac{r_{Z|\boldsymbol{X}}(z \mid \boldsymbol{x})^\beta}{\sum_z' r_{Z|\boldsymbol{X}}(z' \mid \boldsymbol{x})^\beta} = \frac{1}{|\mathcal{Z}|} = 2^{-n_z},$$

for every $z \in \mathcal{Z}$ and $\boldsymbol{x} \in \mathcal{X}$. Second, we derive the limit of $\beta \to \infty$. Let $(\Omega, \mathcal{F}, \Pr)$ be a probability space. From the assumption of $\mathcal{R}$, there exists a null set $\mathcal{N}$ such that

$$\mathcal{N} = \Omega \setminus \bigcap_{\substack{z_1, z_2 \in \mathcal{Z} \\ z_1 \neq z_2}} \{r_{Z|\boldsymbol{X}}(z_1|\boldsymbol{X}) \neq r_{Z|\boldsymbol{X}}(z_2|\boldsymbol{X})\}.$$

Let $\{\beta_i\}_{i=1}^{\infty}$ be any sequence such that $\beta_i \to \infty$. For every $(z, \boldsymbol{x}) \in (Z, \boldsymbol{X})(\Omega \setminus \mathcal{N})$,[4]

$$\begin{aligned}
\lim_{i\to\infty} \mathscr{H}[r_{Z|\boldsymbol{X}}]_{\beta_i}(z \mid \boldsymbol{x}) &= \lim_{i\to\infty} \frac{r_{Z|\boldsymbol{X}}(z \mid \boldsymbol{x})^{\beta_i}}{\sum_{z'} r_{Z|\boldsymbol{X}}(z' \mid \boldsymbol{x})^{\beta_i}} \\
&= \lim_{i\to\infty} \frac{\left(\frac{r_{Z|\boldsymbol{X}}(z|\boldsymbol{x})}{\max_{\bar{z}} r_{Z|\boldsymbol{X}}(\bar{z}|\boldsymbol{x})}\right)^{\beta_i}}{\sum_{z'} \left(\frac{r_{Z|\boldsymbol{X}}(z'|\boldsymbol{x})}{\max_{\bar{z}} r_{Z|\boldsymbol{X}}(\bar{z}|\boldsymbol{x})}\right)^{\beta_i}} \\
&= \begin{cases} 1 & \text{if } z = \arg\max_{z'} r_{Z|\boldsymbol{X}}(z' \mid \boldsymbol{x}) \\ 0 & \text{otherwise} \end{cases} \\
&= \mathbb{1}_{\{z=\arg\max_{z'} r_{Z|\boldsymbol{X}}(z'|\boldsymbol{x})\}}.
\end{aligned}$$

Therefore, $\mathbb{1}_{\{Z=\arg\max_{z'} r_{Z|\boldsymbol{X}}(z'|\boldsymbol{X})\}}$ is measurable, and it holds $\lim_{\beta\to\infty} \mathscr{H}_\beta[r_{Z|\boldsymbol{X}}](Z|\boldsymbol{X}) = \mathbb{1}_{\{Z=\arg\max_{z'} r_{Z|\boldsymbol{X}}(z'|\boldsymbol{X})\}}$ almost surely because $\Pr(\Omega \setminus \mathcal{N}) = 1$. □

We then introduce Proposition 3.

Proposition 3. *Let $r_{Z|\boldsymbol{X}} \in \mathcal{R}$ be a conditional probability distribution. Suppose that $\Pr\left(Z \neq \arg\max_{z'} r(z' \mid \boldsymbol{X})\right) > 0$. Then we have*

$$\mathrm{CE}(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}]) \to n_z \qquad \text{as} \quad \beta \searrow 0, \tag{7}$$

$$J_{\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}]}(Z; \boldsymbol{X}) \to 0 \qquad \text{as} \quad \beta \searrow 0, \tag{8}$$

$$\mathrm{CE}(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}]) \to \infty \qquad \text{as} \quad \beta \to \infty, \tag{9}$$

$$J_{\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}]}(Z; \boldsymbol{X}) \to -\infty \qquad \text{as} \quad \beta \to \infty. \tag{10}$$

Proof (Informal sketch). Intuitively[5], Limits (7) and (8) hold because $\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}]$ converges almost surely to a uniform distribution over $\mathcal{Z}$ and CE of the uniform distribution is equivalent to n_z. As well, Limits (7) and (8) hold because $\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}]$ converges almost surely to a one-hot distribution and $\lim_{x\searrow 0} \log x = -\infty$. See Appendix A for a formal proof. □

Proposition 3 states the CE/PI is dependant on β. In particular, we can make the CE arbitrarily large and the PI arbitrarily small by increasing β. A conditional probability distribution can be converted to other distributions with arbitrarily large CE and small PI, while the SR and GE of the distinguishing rule with such probability distributions are invariant to β. In an extreme case, according to Proposition 1, the true probability distribution $p_{Z|\boldsymbol{X}}$, which has the minimum CE/maximum PI, gives an optimal distinguisher (*i.e.*, achieves the theoretically maximum SR); but $\mathscr{H}_\beta[p_{Z|\boldsymbol{X}}]$ also provides an optimal distinguisher, although $J_{\mathscr{H}_\beta[p_{Z|\boldsymbol{X}}]}(Z; \boldsymbol{X})$ is smaller than zero for sufficiently large β. This statement also holds for any non-optimal probability distribution that can achieve a meaningful SR. Thus, CE and PI include an uncertainty in terms of SR; that is, for a given SR, the CE and PI of a probability distribution are not unique and the probability distribution can have an arbitrarily large CE/small PI. Moreover, PI can take any negative value although it is expected to intuitively represent an information amount. This reveals that the conventional CE/PI is non-calibrated and not always appropriate in terms of SR evaluation and is insufficient for the evaluation of the attack performance (with Inequality (6)).

[4] Note that $(Z, \boldsymbol{X})$ can also be regarded as a random variable.

[5] To prove Proposition 3 formally, we need a formal treatment of convergence through the measure theory, such as the interchange of lim and expectation in CE. This proof sketch describes just an intuition.

4 Proposed metrics: Effective CE/PI

4.1 Basic concept

In Section 3, we showed the uncertainty of CE/PI in terms of SR. To avoid such an uncertainty, we present a modification of CE/PI, named effective CE/PI (ECE/EPI). The ECE and EPI are defined as a CE lower-bound and PI upper-bound for a given probability distribution with regard to the conversion $\mathscr{H}_\beta$, respectively.

Definition 3 (Effective cross-entropy (ECE) and effective perceived information (EPI))**.** Let $r_{Z|\boldsymbol{X}} \in \mathcal{R}$ be a conditional probability distribution of secret intermediate variable Z given a side-channel leakage $\boldsymbol{X}$. ECE and EPI for $r_{Z|\boldsymbol{X}}$ are defined as

$$\begin{aligned} \mathrm{CE}^*(r_{Z|\boldsymbol{X}}) &= \inf_{\beta\in(0,\infty)} \mathrm{CE}(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}]) = \inf_{\beta\in(0,\infty)} -\mathbb{E}\log \mathscr{H}_\beta[r_{Z|\boldsymbol{X}}](Z \mid \boldsymbol{X}), \\ J_r^*(Z;\boldsymbol{X}) &= \sup_{\beta\in(0,\infty)} J_{\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}]}(Z;\boldsymbol{X}) = H(Z) - \mathrm{CE}^*(r_{Z|\boldsymbol{X}}), \end{aligned}$$

respectively. If Z follows a uniform distribution over $\{0,1\}^{n_z}$, then $H(Z) = n_z$.

Note that EPI is always non-negative as proven in Proposition 4.

Thus, we define the ECE/EPI of $r_{Z|\boldsymbol{X}}$ by the infimum of CE/supremum of PI of probability distributions in $\{\, \mathscr{H}_\beta[r_{Z|\boldsymbol{X}}] \mid \beta \in (0,\infty) \,\}$. In other words, given $r_{Z|\boldsymbol{X}}$, we can generate an infinite number of conditional probability distribution as $\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}]$ which has the same SR as $r_{Z|\boldsymbol{X}}$ with a different CE/PI. To uniquely determine CE/PI as ECE/EPI, we take the infimum of CE (or supremum of PI) among them. Thus, it is likely that ECE/EPI is given by a lower-bound of CE (or an upper-bound of PI) of probability distributions that can achieve an SR, and a probability distribution with the same ECE/EPI yields the same SR. This indicates that ECE/EPI is more appropriate for SR evaluation with regard to conversion $\mathscr{H}_\beta$ (See Section 4.3 for more detailed discussion).

Recall that PI is designed to represent the amount of information utilized by a conditional probability distribution between secret intermediate variable and side-channel leakage. Because ECE/EPI is defined as the infimum of CE/supremum of PI for a given probability distribution $r_{Z|\boldsymbol{X}}$, we expect that ECE/EPI can be used for the tightest and most accurate SR evaluation for $r_{Z|\boldsymbol{X}}$ with an SR inequality (3) by de Chérisey et al. [dCGRP19]. We expect that the following inequality holds.

Conjecture 1 (SR–EPI inequality)**.** *Let* $\mathrm{SR}_m(r)$ *denotes the success rate with a distinguishing rule in Proposition 1 using a conditional probability distribution* $r_{Z|\boldsymbol{X}}$ *when the number of attack traces is* m*. Then, we have*

$$\xi(\mathrm{SR}_m(r)) \le m J_r^*(Z;\boldsymbol{X}), \tag{11}$$

for the evaluation of SR upper-bound for a given probability distribution $r_{Z|\boldsymbol{X}}$*. To achieve* $\mathrm{SR}_m(r) = 1$*, Inequality* (11) *is also represented by*

$$\frac{n_k}{J_r^*(Z;\boldsymbol{X})} \le m. \tag{12}$$

As in Inequality (12), SR upper-bound conversely represents a lower-bound of the number of traces required for an attack success (*i.e.*, to achieve an SR). We demonstrate the validity, effectiveness, and tightness of the SR evaluation using EPI/ECE through experimental attacks in Section 5.

As well as PI, EPI is upper-bounded by the mutual information $I(Z;\boldsymbol{X})$. In addition, EPI is always non-negative, although the conventional PI can take any negative value. Proposition 4 describes the range of EPI.

Proposition 4 (Range of EPI)**.** *For any probability distribution $r_{Z|\boldsymbol{X}} \in \mathcal{R}$, we have*

$$0 \leq J_r^*(Z;\boldsymbol{X}) \leq I(Z;\boldsymbol{X}).$$

The equality $J_r^*(Z;\boldsymbol{X}) = I(Z;\boldsymbol{X})$ holds if and only if $r_{Z|\boldsymbol{X}}$ is equivalent to $\mathscr{H}_\beta[p_{Z|\boldsymbol{X}}]$ for some β, where $p_{Z|\boldsymbol{X}}$ denotes the true probability distribution.

Proof. Firstly, $0 \leq J_r^*(Z;\boldsymbol{X}) = H(Z) - \mathrm{CE}^*(r_{Z|\boldsymbol{X}})$ holds because $\sup_r \mathrm{CE}^*(r_{Z|\boldsymbol{X}}) = n_z$, which follows from Limit (7): $\mathrm{CE}(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}]) \to n_z$ as $\beta \searrow 0$.

Secondly, we prove $J_r^*(Z;\boldsymbol{X}) \leq I(Z;\boldsymbol{X})$. Recall that $\mathrm{CE}(r)$ takes the minimum if and only if $r = p$, which is followed by $\mathrm{CE}(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}]) \geq \mathrm{CE}(p_{Z|\boldsymbol{X}}) = H(Z \mid \boldsymbol{X})$ for any β, where $H(Z \mid \boldsymbol{X}) = -\mathbb{E}\log p_{Z|\boldsymbol{X}}(Z \mid \boldsymbol{X})$ is the conditional entropy of Z given $\boldsymbol{X}$. Therefore, we have

$$J_r^*(Z;\boldsymbol{X}) = H(Z) - \mathrm{CE}^*(r_{Z|\boldsymbol{X}}) \leq H(Z) - H(Z \mid \boldsymbol{X}) = I(Z;\boldsymbol{X}).$$

This inequality also states that the equality holds if and only if $\mathrm{CE}^*(r_{Z|\boldsymbol{X}}) = H(Z \mid \boldsymbol{X}) = \mathrm{CE}(p_{Z|\boldsymbol{X}})$. If $r_{Z|\boldsymbol{X}} = \mathscr{H}_\beta[p_{Z|\boldsymbol{X}}]$ for some β, then $\mathrm{CE}^*(r_{Z|\boldsymbol{X}}) = \inf_\beta \mathrm{CE}(\mathscr{H}_\beta[p_{Z|\boldsymbol{X}}]) = \mathrm{CE}(p_{Z|\boldsymbol{X}})$; otherwise, $\mathrm{CE}^*(r_{Z|\boldsymbol{X}}) > \mathrm{CE}(p_{Z|\boldsymbol{X}})$. □

Proposition 4 would validate the usage of Inequality (11) for the SR evaluation; that is, the SR–EPI inequality (11) does not overestimate the attack performance (*i.e.*, SR upper-bound and lower-bound of the number of traces required for attack success) than an optimal attack with $p_{Z|\boldsymbol{X}}$, as a larger $J_r(Z;\boldsymbol{X})$ implies a higher performance. Moreover, EPI provides a tighter and more accurate evaluation than the conventional PI (as it always holds $J_r(Z;\boldsymbol{X}) \leq J_r^*(Z;\boldsymbol{X})$ due to its definition), whereas PI is likely to underestimate the attack performance as discussed in Section 3. Major differences of Proposition 4 from the inequality $J_r(Z;\boldsymbol{X}) \leq I(Z;\boldsymbol{X})$ in [MDP20, BHM+19] are the equality condition and that EPI is guaranteed to be non-negative. EPI is consistent with its intuitive meanings, as EPI is maximized by all probability distributions that provide optimal distinguisher with regard to $\mathscr{H}_\beta$ and is always non-negative.

4.2 Relation between attack success and ECE/EPI

To discuss ECE/EPI in detail, we introduce Lemma 3, stating that $L(\mathscr{H}_\beta[r])$, which is an approximation of $\mathrm{CE}(\mathscr{H}_\beta[r])$, is a strictly convex function in terms of β.

Lemma 3. *Let $r_{Z|\boldsymbol{X}} \in \mathcal{R}$ be a conditional probability distribution, and let $\mathscr{H}_\beta$ be a conversion of probability distribution defined above. Let β be a positive real number. $L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}])$ is almost surely a strictly convex function in β and and $\mathrm{CE}(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}])$ is a strictly convex function in β.*

Proof. To handle NLL and CE simultaneously, we introduce the empirical distribution. Let $F_{Z,\boldsymbol{X}}$ be a true cumulative probability function, and let $\hat{F}_{Z,\boldsymbol{X}}^{(m)}$ be an empirical probability distribution for m samples. We can denote NLL and CE by

$$\begin{aligned} L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}]) &= -\mathbb{E}_{(Z,\boldsymbol{X})\sim\hat{F}_{Z,\boldsymbol{X}}^{(m)}} \log \frac{r_{Z|\boldsymbol{X}}(Z \mid \boldsymbol{X})^\beta}{\sum_{z'} r_{Z|\boldsymbol{X}}(z' \mid \boldsymbol{X})^\beta}, \\ \mathrm{CE}(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}]) &= -\mathbb{E}_{(Z,\boldsymbol{X})\sim F_{Z,\boldsymbol{X}}} \log \frac{r_{Z|\boldsymbol{X}}(Z \mid \boldsymbol{X})^\beta}{\sum_{z'} r_{Z|\boldsymbol{X}}(z' \mid \boldsymbol{X})^\beta}, \end{aligned} \tag{13}$$

respectively. Therefore, it is sufficient to consider the following equation:

$$-\mathbb{E}\log \frac{r_{Z|\boldsymbol{X}}(Z \mid \boldsymbol{X})^\beta}{\sum_{z'} r_{Z|\boldsymbol{X}}(z' \mid \boldsymbol{X})^\beta} = -\mathbb{E}\left[\beta \log r_{Z|\boldsymbol{X}}(Z \mid \boldsymbol{X}) - \log \sum_{z'} r_{Z|\boldsymbol{X}}(z' \mid \boldsymbol{X})^\beta\right]. \tag{14}$$

Recall that the sum of linear and convex/concave functions is a convex/concave function, and a concave function is the negative of a convex function and *vice versa*. In the expectation in Equation (14), the first term $\beta \log r_{Z|\boldsymbol{X}}(Z_j \mid \boldsymbol{X}_j)$ is linear in terms of β. We then consider the convexity of the second term $-\log \sum_{z'} r_{Z|\boldsymbol{X}}(z' \mid \boldsymbol{X})^\beta$. The second term can be rewritten as

$$\begin{aligned} -\log \sum_{z'} r_{Z|\boldsymbol{X}}(z' \mid \boldsymbol{X})^\beta &= -\log \sum_{z'} \exp(\beta \ln r(z' \mid \boldsymbol{X})) \\ &= -\log(e)\mathrm{LSE}(\beta \ln r(0 \mid \boldsymbol{X}), \dots, \beta \ln r(|\mathcal{Z}|-1 \mid \boldsymbol{X})), \quad (15) \end{aligned}$$

where LSE is a log-sum-exponential (LSE) function. Equation (15) is concave because it is well-known that a LSE function is convex. We then prove that it is *strictly* concave with probability 1. We first investigate the condition where a LSE function becomes strictly convex. Let $\boldsymbol{y} \in \mathbb{R}^n$ be an n-dimensional real vector. For a twice-differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, let $\nabla^2 f$ be the Hessian matrix of f, respectively. Let $\boldsymbol{v} = (\sum_i e^{y_i})^{-1}(e^{y_1}, e^{y_2}, \dots, e^{y_n})^T$. We then have

$$\nabla^2 \mathrm{LSE}(\boldsymbol{y}) = \mathrm{diag}(\boldsymbol{v}) - \boldsymbol{v}\boldsymbol{v}^T.$$

Note that $\mathbf{1}^T \nabla^2 \mathrm{LSE}(\boldsymbol{y})\mathbf{1} = 0$ because $\mathbf{1}^T \boldsymbol{v} = 1$, where $\mathbf{1} = (1, 1, \dots, 1)^T$ is an n-dimensional vector whose elements are 1. Since $\mathrm{rank}(\mathrm{diag}(\boldsymbol{v})) = n$ and $\mathrm{rank}(\boldsymbol{v}\boldsymbol{v}^T) = 1$, the rank of the Hessian matrix $\mathrm{rank}(\nabla^2 \mathrm{LSE}(\boldsymbol{y}))$ is equal to $n-1$. In other words, for any vector $\boldsymbol{u} \in \mathbb{R}^n$, $\boldsymbol{u}^T \nabla^2 f \boldsymbol{u} > 0$ if $\boldsymbol{u}$ is linearly independent of $\boldsymbol{v}$. Let f be a function $(0, \infty) \ni \beta \mapsto \mathrm{LSE}(\beta \ln r(0 \mid \boldsymbol{X}), \beta \ln r(1 \mid \boldsymbol{X}), \dots, \beta \ln r(|\mathcal{Z}|-1 \mid \boldsymbol{X}))$. For any $\beta_1, \beta_2 \in (0, \infty)$ and any $\lambda \in (0, 1)$, we have $f(\lambda\beta_1 + (1-\lambda)\beta_2) < \lambda f(\beta_1) + (1-\lambda) f(\beta_2)$ almost surely, because $(\ln r(0 \mid \boldsymbol{X}), \ln r(1 \mid \boldsymbol{X}), \dots, \ln r(|\mathcal{Z}|-1 \mid \boldsymbol{X}))^T$ is linearly independent of $\mathbf{1}$ almost surely due to $r \in \mathcal{R}$.

Thus, the summation $\beta \log r_{Z|\boldsymbol{X}}(Z \mid \boldsymbol{X}) - \log \sum_{z'} r_{Z|\boldsymbol{X}}(z' \mid \boldsymbol{X})^\beta$ is almost surely a strictly concave function. Because the expectation of a convex/concave function is a convex/concave function [BBV04, Section 3.2.1], Equation (14) is a strictly convex function. □

A strictly convex function has at most one stationary point and the function takes the unique minimum at the stationary point. There are two cases for a given r: there exists a minimum of $L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}])$ for $\beta > 0$ or not. If there exists $\min_\beta L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}])$ for $\beta > 0$, then $\mathrm{CE}^*(r_{Z|\boldsymbol{X}}) < H(Z)$ and $J_r^*(Z; \boldsymbol{X}) > 0$. This is because there exists some β such that $L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}]) < \lim_{\beta \searrow 0} L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}]) = n_k$ owing to the convexity. Therefore, in this case, we can conclude that the attack using such r would succeed for some number of traces m that satisfies the SR–EPI inequalities (11) and (12).

We then consider another case. If there does not exist $\min_\beta L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}])$, then $L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}])$ is monotonically increasing on the range of $\beta \in (0, \infty)$ owing to its convexity. Therefore, according to Proposition 3, $\inf_\beta L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}]) = n_z$ because $\lim_{\beta \searrow 0} L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}]) = n_z$ for any r, which is equivalent to $J_r^*(Z; \boldsymbol{X}) = 0$. Thus, such a conditional probability distribution exploits as little information about secret intermediate variable from side-channel leakage as a uniform distribution over $\mathcal{Z}$; and therefore, the attack using this conditional probability distribution would fail. Note that, if $J_r^*(Z; \boldsymbol{X}) = 0$, then $\xi(\mathrm{SR}_m(r)) = 0$ according to the SR–EPI inequality (11), which is followed by $\mathrm{SR}_m(r) = 1/2^{n_k}$ for any m. In contrast, the conventional PI cannot guarantee an attack failure even if $J_r(Z; \boldsymbol{X}) = 0$ as discussed in Section 3, whereas EPI would (relatively) correctly evaluate the attack performance with a probability distribution. Hence, $\inf_\beta L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}]) = \min_\beta L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}]) < n_z$ if there exists $\min_\beta L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}])$ for $\beta > 0$; otherwise, $\inf_\beta L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}]) = n_z$. This notation, which implies that EPI is always non-negative (as $J_r^*(Z; \boldsymbol{X}) = H(Z) - \mathrm{CE}^*(r_{Z|\boldsymbol{X}})$ where $\sup_r \mathrm{CE}^*[r_{Z|\boldsymbol{X}}] = n_z$), is consistent with the intuitive meanings of EPI, although the conventional PI defined in Equation (5) can be a negative value. In summary, the

condition that $\min_\beta L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}])$ exists for $\beta > 0$ would be sufficient for an attack success from the viewpoint of EPI, whereas our EPI-based SR evaluation method also indicates that a conditional probability distribution which does not satisfy this condition would fail to attack.

4.3 Suitability of ECE/EPI for SR evaluation

Using Lemma 1, we can prove that the order of ranks for each key candidate is invariant to the conversion $\mathscr{H}_\beta$ (as in Theorem 2); and thus, SR/GE is invariant to β. EPI/ECE is a metric to address this uncertainty. If there exist other conversions of conditional probability distribution which preserve the order of key ranks, EPI/ECE may not be able to accurately evaluate the SR using Inequalities (11) and (12). Fortunately, we can prove that there exists no such conversion except for $\mathscr{H}_\beta$, which states that ECE/EPI is appropriate for SR evaluation with regard to probability distribution conversion that preserves the order of key ranks. For the proof, we first define an equivalence relation, quotient set, and order to represent key ranks.

Definition 4 (Key rank order)**.** Let $\boldsymbol{a} = (a_1, a_2, \ldots, a_m)$ and $\boldsymbol{b} = (b_1, b_2, \ldots, b_n)$ be an element of $\mathbb{R}^m$. Define an equivalence relation $\sim$ as $\boldsymbol{a} \sim \boldsymbol{b} \Rightarrow \sum_j a_j = \sum_j b_j$. Denote a quotient set of $\boldsymbol{a}$ by $[\boldsymbol{a}] \in \mathbb{R}^m/\!\sim$. Define a strict total order $\lhd$ on $\mathbb{R}^m/\!\sim$ as $[\boldsymbol{a}] \lhd [\boldsymbol{b}] \Rightarrow \sum_j a_j < \sum_i b_j$.

This equivalence relation, quotient set, and order represent key rank, because key ranks are calculated as NLL, namely, a sum of negative log-probabilities (*i.e.*, real numbers) like $\sum_j a_j$. We then introduce Lemma 4.

Lemma 4. *Let $F\colon \mathbb{R}^m \to \mathbb{R}^m; (a_1, a_2, \ldots, a_m) \mapsto (f(a_1), f(a_2), \ldots, f(a_m))$ be a function defined using a function $f : \mathbb{R} \to \mathbb{R}$. Function F is assumed to be well-defined as a function from $\mathbb{R}^m/\!\sim$ to $\mathbb{R}^m/\!\sim$. F is order automorphic on $(\mathbb{R}^m/\!\sim, \lhd)$ if and only if f is given by a linear polynomial function $f(a) = \beta a + \gamma$, where β is a positive real number and γ is a real number.*

Proof. Let $S([\boldsymbol{a}]) = \sum_j a_j$. Because $\boldsymbol{a} \lhd \boldsymbol{b} \Leftrightarrow \sum_j a_j < \sum_j b_j$, it holds $F(\boldsymbol{a}) \lhd F(\boldsymbol{b}) \Leftrightarrow S(F(\boldsymbol{a})) < S(F(\boldsymbol{b}))$. Hence, if $S \circ F : \mathbb{R}^m/\!\sim \to \mathbb{R}$ is an order isomorphism from $(\mathbb{R}^m/\!\sim, \lhd)$ to $(\mathbb{R}, <)$, then F is an order automorphism. Therefore, we show that $S \circ F$ is an order isomorphism if and only if $f(a) = \beta a + \gamma$ ($\beta > 0$).

($\Leftarrow$) If $f(a) = \beta a + \gamma$, then $S \circ F([\boldsymbol{a}]) = \sum_j \beta a_j + m\gamma$ and $S \circ F([\boldsymbol{b}]) = \sum_j \beta b_j + m\gamma$. Hence, if $[\boldsymbol{a}] \lhd [\boldsymbol{b}]$, then $S \circ F([\boldsymbol{a}]) - S \circ F([\boldsymbol{b}]) = \beta \left(\sum_j a_j - \sum_j b_j \right) < 0$, which is followed by $S \circ F([\boldsymbol{a}]) < S \circ F([\boldsymbol{b}])$.

($\Rightarrow$) Order automorphism on $(\mathbb{R}, <)$ is always a strictly monotonically increasing function. According to the assumption that $S \circ F$ is order isomorphic, there exists a strictly monotonically increasing function g such that $g(\sum_j a_j) = (S \circ F)([\boldsymbol{a}]) = \sum_j f(a_j)$. Consider the functional equation $g(\sum_j a_j) = \sum_j f(a_j)$. Let h be a translation of f such that $h(0) = 0$, that is, $h(a) = f(a) - f(0)$. Let $e(a) = g(a) - mf(0)$. We then have $e(\sum_j a_j) = \sum_j h(a_j)$. In addition, it holds $e = h$ because $e(a_1) = e(a_1 + \sum_{j=2}^m 0) = h(a_1) + \sum_{j=2}^m h(0) = h(a_1)$. Here, h (and e) is a conditional solution of Cauchy's functional equation $h(a_1 + a_2 + \cdots + a_m) = h(a_1) + h(a_2) + \cdots + h(a_m)$, where the condition is that e is a monotone function. Therefore, $h(a)$ is given by $h(a) = \beta a$ for some positive real number β, and $\beta > 0$ because e is monotonically increasing. By letting $f(0) = \gamma$, we conclude $f(a) = \beta a + \gamma$. □

Using Lemma 4, we prove Theorem 2.

Theorem 2. *Let $\mathcal{S}_a$ be a trace dataset for the attack. Let $r_{Z|\boldsymbol{X}}, r'_{Z|\boldsymbol{X}} \in \mathcal{R}$ be conditional probability distributions. Then, for all $k \in \mathcal{K}$ and $m \in \mathbb{N}$, we have* $\mathrm{rank}(k, m, r) = \mathrm{rank}(k, m, r')$ *if and only if there exists a positive real number β such that* $r'_{Z|\boldsymbol{X}}(z \mid \boldsymbol{x}) = r_{Z|\boldsymbol{X}}(z \mid \boldsymbol{x})^\beta / \sum_{z'} r_{Z|\boldsymbol{X}}(z' \mid \boldsymbol{x})^\beta$.

Proof. It is obvious from Lemma 1 that the sufficient condition is true. We prove that the necessary condition is true; that is, if the order of ranks for each key candidate is identical for r and r', then $r' = \mathscr{H}_\beta[r]$ for some β. Because $L^{(k)}(r)$ is given by a sum of m real numbers (*i.e.*, negative log-output of conditional probability distribution), the ranks correspond to the strict total order $\lhd$ defined in Definition 4. According to the assumption that the order of ranks (namely, NLLs for key candidates) is preserved, Lemma 4 states that the conversion applicable to $-\log r(z_j^{(k)} \mid \boldsymbol{x}_j)$ is only $f(-\log r(z_j^{(k)} \mid \boldsymbol{x}_j)) = -\beta \log r(z_j^{(k)} \mid \boldsymbol{x}_j) + \gamma$ for any j. Therefore, it holds

$$-\log r'(z \mid \boldsymbol{x}) = -\beta \log r(z \mid \boldsymbol{x}) + \gamma,$$

which is followed by

$$r'(z \mid \boldsymbol{x}) = \gamma' r(z \mid \boldsymbol{x})^\beta,$$

where $\gamma' = 2^{-\gamma}$. Additionally, because r' is a probability distribution, $\sum_{z'} r'(z' \mid \boldsymbol{x}) = 1$; hence, $\sum_{z'} r'(z' \mid \boldsymbol{x}) = \gamma' \sum_{z'} r(z' \mid \boldsymbol{x})^\beta = 1$, which is followed by $\gamma' = 1/\sum_{z'} r(z' \mid \boldsymbol{x})^\beta$. Thus, we conclude $r'(z \mid \boldsymbol{x}) = r(z \mid \boldsymbol{x})^\beta / \sum_{z'} r(z' \mid \boldsymbol{x})^\beta$. □

Theorem 2 states that the proposed metrics are the most appropriate for SR and GE evaluations among any conversions of probability distribution that preserve the order of key ranks. However, if there is a conversion such that SR is preserved but the order of key ranks are not preserved, ECE and EPI are not guaranteed to be unique to an SR and be appropriate for SR evaluation. If there does not exist such a conversion, then ECE and EPI are truly unique in terms of SR evaluation. The analysis on the existence of such a conversion is an important future work.

4.4 Computation of ECE/EPI in practice

In this subsection, we describe how to evaluate ECE/EPI of a probability distribution using a given dataset. Recall that CE cannot be calculated directly in practice, and NLL is used for its approximation. This indicates that, to calculate $\inf_\beta \mathrm{CE}(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}])$, we need to approximate it as $\inf_\beta L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}])$. Let $\hat{\beta}_m$ be the optimal β value derived by an empirical calculation of $\inf_\beta L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}])$ using m traces (namely, $\hat{\beta}_m$ satisfies $L(\mathscr{H}_{\hat{\beta}_m}[r_{Z|\boldsymbol{X}}]) = \inf_\beta L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}])$).
Let β_0 be a parameter that satisfies $\mathrm{CE}(\mathscr{H}_{\beta_0}[r_{Z|\boldsymbol{X}}]) = \inf_\beta \mathrm{CE}(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}])$. For a better approximation of ECE through an empirical derivation of $L(\mathscr{H}_{\hat{\beta}_m}[r_{Z|\boldsymbol{X}}])$, $\hat{\beta}_m$ should converge to with β_0 in a sense. For this purpose, we prove Theorem 3, which states that empirically calculated $\hat{\beta}_m$ and $L(\mathscr{H}_{\hat{\beta}_m}[r_{Z|\boldsymbol{X}}])$ are "good" approximations of β_0 and $\mathrm{CE}(\mathscr{H}_{\beta_0}[r_{Z|\boldsymbol{X}}]) = \mathrm{CE}^*(r_{Z|\boldsymbol{X}})$ in terms of strong consistency, respectively.

Theorem 3. *Let $r_{Z|\boldsymbol{X}} \in \mathcal{R}$ be a conditional distribution. Assume that $-\mathbb{E}\sum_{z'} \ln r(z'|\boldsymbol{X}) < \infty$. Let $\mathcal{B} = [0, M_\beta]$ be the set of β, where M_β is a positive real number*[6]*. Let β_0 be a parameter that satisfies* $\mathrm{CE}(\mathscr{H}_{\beta_0}[r_{Z|\boldsymbol{X}}]) = \inf_{\beta\in\mathcal{B}} \mathrm{CE}(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}])$. *Suppose that $\beta_0 \in \mathcal{B}$. Let m be the number of traces that used for the empirical calculation of NLL $L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}])$, and let $\hat{\beta}_m \in \mathcal{B}$ be a random variable of a parameter such that $L(\mathscr{H}_{\hat{\beta}_m}[r]) = \inf_{\beta\in\mathcal{B}} L(\mathscr{H}_\beta[r])$. Then, we have $\hat{\beta}_m \xrightarrow{\text{a.s.}} \beta_0$ as $m \to \infty$, where $\xrightarrow{\text{a.s.}}$ denotes the almost sure convergence.*

[6] To preserve SR and GE, β must be in $(0, \infty)$. However, in this theorem, we assume the range of β contain 0 because the infimum of β can be 0. By changing the range of β, the infimum of β is equivalent to the minimum of β.

Proof. Theorem 3 is proven on the basis of [Fer96, Theorem 17]. See Appendix B. □

Theorem 3 means that if β is restricted to the closed set $[0, M_\beta]$ and $-\mathbb{E}\sum_{z'} \ln r(z'|\boldsymbol{X})$ is bounded, the maximum likelihood estimator[7] $\hat{\beta}_m$ converges almost surely to the true value β_0. According to Theorem 3, we can approximately evaluate the ECE for a given probability distribution $r_{Z|\boldsymbol{X}}$ (*i.e.*, $\inf_\beta \mathrm{CE}(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}])$) and the corresponding EPI using an empirical evaluation of $\min_\beta L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}])$. In other words, Theorem 3 validates the usage of $L(\mathscr{H}_{\hat{\beta}_m}[r_{Z|\boldsymbol{X}}])$ as an approximation of $\mathrm{CE}(\mathscr{H}_{\beta_0}[r_{Z|\boldsymbol{X}}])$ for a large number of traces.

Finally, we describe a concrete method for calculating $\hat{\beta}_m$ and $L(\mathscr{H}_{\hat{\beta}_m}[r_{Z|\boldsymbol{X}}])$ for given $r_{Z|\boldsymbol{X}}$ and dataset. Recall that $\hat{\beta}_m$ is a solution of $\frac{\partial}{\partial\beta} L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}]) = 0$ for $\beta > 0$ if it exists, as discussed in the proof of Theorem 3. Since $L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}])$ is convex as proved in Lemma 3, we can approximately derive $\hat{\beta}_m$ such that $L(\mathscr{H}_{\hat{\beta}_m}[r_{Z|\boldsymbol{X}}]) = \inf_\beta L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}])$ using the Newton–Raphson method. Owing to the convexity of $L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}])$, the solution of the Newton–Raphson method always converges to $\hat{\beta}_m$ that satisfies $\frac{\partial}{\partial\beta} L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}])\Big|_{\beta=\hat{\beta}_m} = 0$ if there exists $\min_\beta L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}])$ for $\beta > 0$; otherwise, let $\inf_\beta L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}]) = n_z$. Recall that the validation loss $L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}])$ (at base 2) is given by

$$L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}]) = -\frac{1}{m}\sum_{j=1}^{m}\left(\beta \log r_{Z|\boldsymbol{X}}(Z_j \mid \boldsymbol{X}_j) - \log \sum_{z'} r_{Z|\boldsymbol{X}}(z' \mid \boldsymbol{X}_j)^\beta\right). \tag{16}$$

For the use of Newton–Raphson method, the first and second partial derivatives of $L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}])$ in terms of β are given by[8]

$$\frac{\partial}{\partial\beta} L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}]) = -\frac{1}{m}\sum_{j=1}^{m}\left(\log r_{Z|\boldsymbol{X}}(Z_j \mid \boldsymbol{X}_j) - \frac{\psi^{(1)}_{r,\boldsymbol{X}_i,\beta}}{\ln(2)\psi^{(0)}_{r,\boldsymbol{X}_i,\beta}}\right),$$

$$\frac{\partial^2}{\partial\beta^2} L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}]) = -\frac{1}{m}\sum_{j=1}^{m}\frac{\left(\psi^{(1)}_{r,\boldsymbol{X}_i,\beta}\right)^2 - \psi^{(0)}_{r,\boldsymbol{X}_i,\beta}\psi^{(2)}_{r,\boldsymbol{X}_i,\beta}}{\ln(2)\left(\psi^{(0)}_{r,\boldsymbol{X}_i,\beta}\right)^2},$$

respectively, where $\psi^{(s)}_{r,\boldsymbol{X}_i,\beta}$ denotes the s-th partial derivative[9] of $\sum_{z'} r(z' \mid \boldsymbol{X}_j)^\beta$ in terms of β; that is,

$$\psi^{(s)}_{r,\boldsymbol{X}_i,\beta} = \frac{\partial^s}{\partial\beta^s}\sum_{z'} r_{Z|\boldsymbol{X}}(z' \mid \boldsymbol{X}_j)^\beta = \sum_{z'} r_{Z|\boldsymbol{X}}(z' \mid \boldsymbol{X}_j)^\beta (\ln r_{Z|\boldsymbol{X}}(z' \mid \boldsymbol{X}_j))^s.$$

Note again that the condition for the existence of $\min_\beta L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}])$ is that a solution of $\frac{\partial}{\partial\beta} L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}]) = 0$ exists for $\beta > 0$. If no solution exists for $\beta > 0$, then $J^*_r(Z; \boldsymbol{X}) = 0$.

5 Experimental validation using DL-SCA

5.1 Overview

In this section, we conduct two experiments of DL-SCAs to validate the effectiveness of EPI. In the first experiment, we investigate the relation between SR and EPI during training and show that EPI can be used to measure the generalization of the training

[7] Note that $\hat{\beta}_m$ is a maximum likelihood estimator due to its definition.

[8] If we calculate the NLL with the natural logarithm, the coefficient $\ln(2)$ is unnecessary and should be replaced with 1.

[9] We consider the zeroth derivative as the function itself.

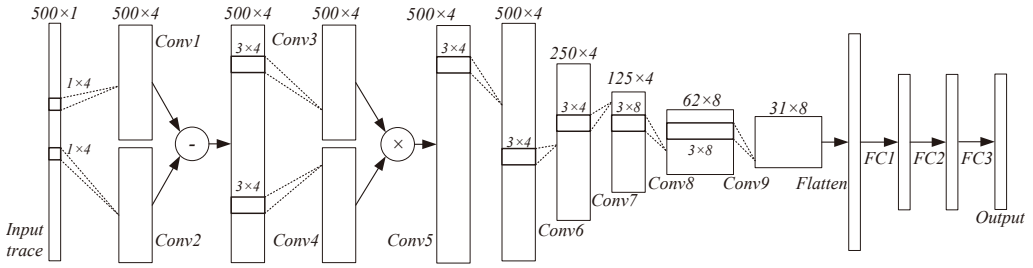


Figure 1: Our NN architecture used in experiment.

model in terms of SCAs. The second experiment demonstrates that EPI is also useful to compare the performances of several models. Our metrics enable us to select a model with good performance without the intensive calculation of SR.

5.2 Performance evaluation during training

5.2.1 Experimental setup

We demonstrate the validity of the proposed metrics through experimental attacks on masked AES software and hardware implementations. For the experiment, we employ a DL-SCA, considered as one of the best attacks with a distinguishing rule that directly utilizes a conditional probability distribution, which is the main focus of this study. The experiment also demonstrates that the proposed metrics can be used to measure the generalization of NN model in terms of SR during the attack phase.

As a masked software implementation, we employed the ASCAD dataset, which is one of the most common datasets to evaluate DL-SCA [BPS+20]. For the attack on ASCAD dataset, we employed an NN model presented by Zaid *et al.* [ZBHV19], which is a publicly available NN model developed for DL-SCA. For the training, we used a categorical CE as a loss function, used Adam as an optimizer, and set the learning rate to 0.001.

As a masked hardware implementation, we used an open-source masked AES hardware based on the threshold implementation (TI) [git21], which was presented in [UHA17]. We synthesized the masked AES hardware as it is (*i.e.*, without the hierarchy broken), implemented it on a Xilinx Kintex-7 FPGA on SAKURA-X board, and acquired its side-channel traces through an on-board co-axial connector. We used a Keysight DSOX6004A oscilloscope and set the sampling rate as 455 MSa/s. We used one million traces for the NN training with random secret keys and plaintexts. The target hardware is byte-serial implementation, which indicates that we should guess two consecutive key bytes to employ XOR based selection function in a practical attack. However, for the simplicity, we consider one byte known and attack on the other one byte in this experiment. Hence, the partial key length in the attack is $n_k = 8$ for both the AES software and hardware implementations in our experiment.

We attempted many NN architectures/hyperparameters to apply DL-SCA to the above TI-based AES hardware, and employed the most successful one for the experiment. In fact, we found that it was difficult to achieve a successful key recovery from the TI-based AES hardware using common NN models in DL-SCA, such as ASCAD and Zaid *et al.*'s models [BPS+20, ZBHV19]. Figure 1 illustrates the NN architecture finally used in our experiment. In the figure, $r \times c$ indicates the size of each feature map or kernel, where r is the length of each filter, and c is the number of channels. Table 1 summarizes our NN hyperparameters. We used CUDA 11.4 cuDNN 8.2.4 Tensorflow 2.6.0 for the training. We used the NLL as a loss function, and set the learning rate, batch size, and the number of epochs as 0.0001, 512, and 1,500, respectively.

Table 1: Hyperparameters of our model

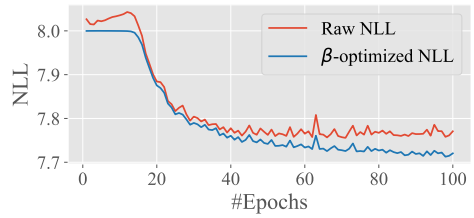
Name	Filter shape	Activation function	Batch normalization	Pooling layer
Conv1	$1 \times 1 \times 4$	SELU	No	-
Conv2	$4 \times 1 \times 4$	SELU	No	-
Conv3	$4 \times 3 \times 4$	tanh	No	-
Conv4	$4 \times 3 \times 4$	SELU	No	-
Conv5	$4 \times 3 \times 4$	SELU	Yes	-
Conv6	$4 \times 3 \times 4$	SELU	Yes	Avg pool (2)
Conv7	$4 \times 3 \times 8$	SELU	Yes	Avg pool (2)
Conv8	$8 \times 3 \times 8$	SELU	Yes	Avg pool (2)
Conv9	$8 \times 3 \times 8$	SELU	Yes	Avg pool (2)
FC1	248×20	SELU	-	-
FC2	20×20	SELU	-	-
FC3	20×256	Softmax	-	-

5.2.2 Experimental results

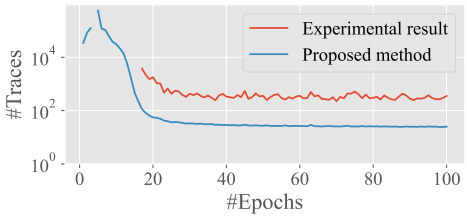
Let q_θ denote the probability distribution of the NN output with a parameter θ. Figure 2 and Figure 3 report the experimental results, where the horizontal axis is the number of epochs in the training. In Figure 2 and Figure 3(a), the red curve denotes raw NLL whereas the blue curve denotes β-optimized NLL loss (*i.e.*, $\inf_\beta L(\mathscr{H}_\beta[q_\theta])$ in Equation (13) or in its summation form (16)), which is an approximation of ECE loss. Figure 3(a') magnifies the blue curve of Figure 3(a) in its range. Figure 2(b) and Figure 3(b) denote the number of traces required for achieving $\mathrm{SR}_m = 0.9$, where the red curve is the empirical result and the blue curve is an estimation value using the proposed metrics with the SR–EPI inequality (11). Note here that the red curve in Figure 2(b) is missing in some epochs where the NN cannot achieve $\mathrm{SR}_{10,000} = 0.9$ because the ASCAD dataset contains only 10,000 test traces. To obtain the blue curves, for a given NN model, we computed $\inf_\beta L(\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}])$ with the test traces using the Newton–Raphson method described in Section 4.4 for an approximation of its ECE and EPI. The terminal condition of the Newton–Raphson method was set such that the difference of the values before and after an iteration is less than 0.001. For ASCAD dataset, it takes about 0.0378 and 14.1 seconds to calculate EPI and SR per one epoch, respectively. For masked AES hardware implementation, it takes about 0.531 and 145 seconds to calculate EPI and SR per one epoch, respectively. Figure 2 and Figure 3 denote the estimated β using the Newton–Raphson method. As we require to evaluate SR for many epochs, the usage of EPI yields a significant advantage in computational cost over the conventional empirical SR evaluation.

We confirm that the red and blue curves in Figure 2(b) and Figure 3(b) are similar in shape, which indicates that the proposed method can appropriately evaluate the lower-bound of the number of traces required for attack success (or SR upper-bound conversely) for a given probability distribution q_θ in an analytical manner using the SR–EPI inequality (11). In particular, we also confirm that the model at the number of epochs with a minimum value of β-optimized NLL loss (*i.e.*, 90 to 100 epochs for Figure 2 and around 720 epochs for Figure 3) achieves the highest attack performance (*i.e.*, achieves the attack success with the smallest number of traces). This implies that the β-optimized NLL loss, which is the approximation of the ECE loss, can also be used to measure the generalization of the NN model in terms of SR maximization in this experiment, and to determine the timing of early stopping.

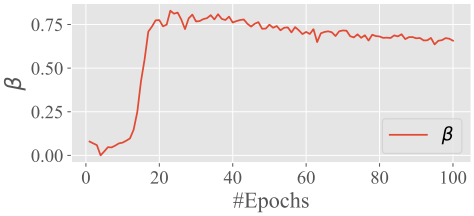
Furthermore, we confirm that the blue curve in Figure 2(a) and Figure 3(a) (and Figure 3(a')) does not exceed $n_k = 8$, as proven in Proposition 4. The attack did not succeed for epochs in the experiment if the β-optimized NLL was $n_k = 8$, which was consistent with the discussion in Section 4. In contrast, in Figure 2(a), the raw NLL is always greater than the β-optimized NLL, which is likely to result in an underestimation of attack performance. Note that PI-based SR estimation is not guaranteed to remain a



(a) Learning curve

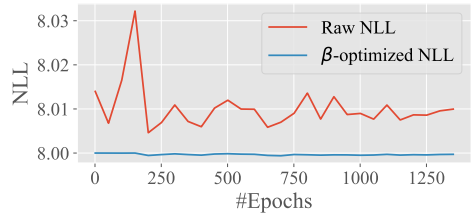


(b) SR evaluation result

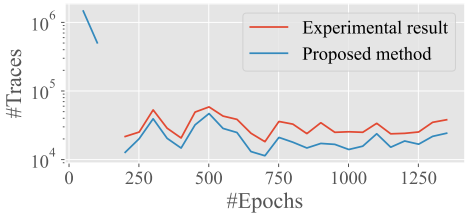


(c) β evaluation result

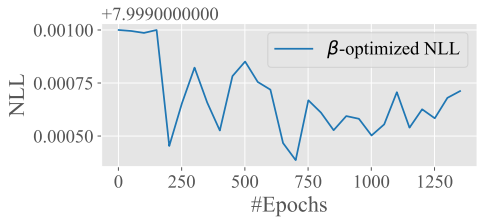
Figure 2: DL-SCA result on ASCAD dataset (*i.e.*, masked AES software implementation).



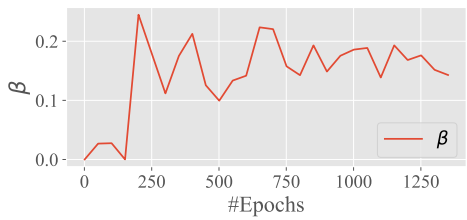
(a) Learning curve



(b) SR evaluation result



(a') Magnified blue curve of (a)



(c) β evaluation result

Figure 3: DL-SCA result on masked AES hardware implementation.

lower bound according to Proposition 3, although the EPI-based SR estimation provides a consistent lower bound of the true SR in accordance with Conjecture 1. The situation is more critical for Figure 3. The raw NLL was greater than $n_k = 8$ for the most parts of Figure 3(a); therefore, the corresponding conventional PI became smaller than zero, which implies that the SR–PI inequality (6) cannot be applicable or significantly underestimates the attack performance, although the attack was actually successful for most parts in the figures in our experiment. Thus, we can confirm the validity, effectiveness, and usefulness of the proposed method.

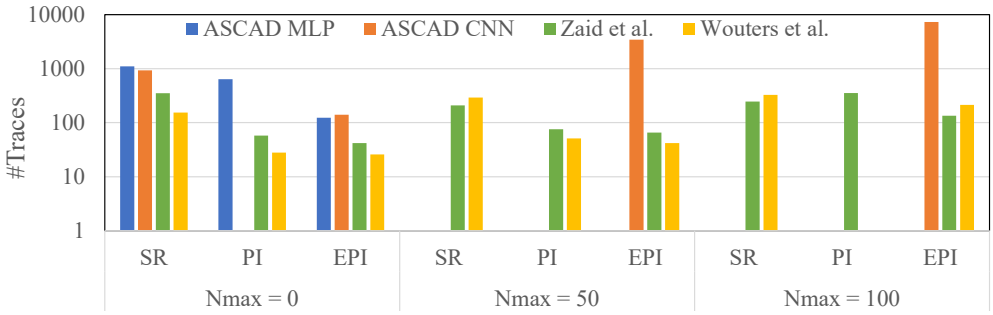


Figure 4: PI, EPI, and empirical SR evaluation results of four models.

5.3 Model comparison

In this subsection, we experimentally calculate SR, PI, and EPI for some models to confirm that our method can also be used as a performance metric for model selection through an experimental attack on the ASCAD dataset without and with desynchronization. We employed four pre-trained models [BPS+20, ZBHV19, WVdHG+20], which are publicly available, to compute and compare their SR, PI, and EPI. The literature [BPS+20], which proposes the ASCAD dataset, uses MLP based-NNs and CNN based-ones to attack the ASCAD dataset, and its authors release the parameters of the best ones in their GitHub repository[10]. For the experiment, we used the best MLP and CNN models from their repository. It is known in [WAGP20] that the performances of Zaid *et al.*'s model and Wouters *et al.*'s model depend significantly on the standardization of the input traces. Therefore, to enhance their performance, we employ "feature standardization" and "horizontal scaling between -1 to 1" for the ASCAD dataset with and without desynchronization, respectively. These model parameters are obtained from the GitHub repository released by Wouters *et al.*[11] Other experimental conditions are the same as that in Section 5.2.1.

Figure 4 reports the experimental results. In the figure, "ASCAD MLP" and "ASCAD CNN" correspond to the models proposed in [BPS+20]. Also, "Nmax" means the amount of desynchronization of the ASCAD dataset. The bars of the SR denotes the number of traces required for successful attacks with 90% probability. The bars of PI and EPI denote the estimated minimum number required for attack success with a 90% probability. The absence of the bar means that the number of required traces for successful attacks would be larger than 10,000.

First, when comparing the results of SR and PI, the number of required traces estimated by PI becomes larger than 10,000, even when attacks succeed with high probability. This would be because of the redundancy of CE/PI in terms of SR. Meanwhile, the figure shows that the proposed method never overestimates the number of required traces. In addition, the number of traces estimated by EPI is approximately proportional to that estimated by SR. Thus, we could compare the attack performances of models by the EPI-based method without the calculation of SR.

6 Conclusion

In this study, we revisited the perceived information (PI), and presented new metrics to evaluate the SCA performance using a conditional probability distribution. We first

[10] https://github.com/ANSSI-FR/ASCAD
[11] https://github.com/KULeuven-COSIC/TCHES20V3_CNN_SCA

showed that the conventional definitions of PI and cross-entropy (CE) had an uncertainty in terms of SR evaluation, and therefore, were non-calibrated and insufficient as metrics for evaluating the SCA performance (*i.e.*, SR). We then presented new metrics, named effective CE/PI (ECE/EPC), to remove the uncertainty. Using ECE/EPI, we can perform more accurate measurements of the SR upper-bound for a given probability distribution in an analytical manner using a PI-SR inequality. ECE/EPI is easily calculated from a given probability distribution for SCA and a dataset, which can be adopted in the context of DL-SCA. We experimentally validated the effectiveness of the proposed method through experimental DL-SCAs on masked AES software and hardware implementations. The experimental results validated our statement on the proposed method, and revealed that the proposed metrics could be used to measure the generalization of NN model in terms of SR maximization. In some ways, the proposed metrics could provide a solution on the open problems on DL-SCA: the relationship between a DL evaluation metric (*i.e.*, loss) and SCA evaluation metrics (*i.e.*, SR/GE) and the difficulty in measuring the generalization and determining the timing of early stopping through the loss value during training. In the future, we will conduct further validation of the proposed metrics using other datasets/implementations. It is also important to prove the unexistance/existance of probability distribution conversion that preserves SR but does not preserves the order of key ranks, to reveal whether ECE and EPI are truly unique to an SR and the most appropriate for the SR evaluation.

The side-channel trace dataset for our experiment on the masked AES hardware is available at https://github.com/ECSIS-lab/perceived_information_revisited.

Acknowledgment

We would like to thank Mr. Kenta Kojima for his technical coporation. We are grateful to Dr. Eleonora Cagli for the shepherding care. This research was supported by JST CREST Grant No. JPMJCR19K5 and the JSPS KAKENHI Grant No. 21H04867 and No. 20K19765, Japan.

References

[BBV04] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[BHM+19] Olivier Bronchain, Julien M. Hendrickx, Clément Massart, Alex Olshevsky, and François-Xavier Standeart. Leakage certification revisited: Bounding model errors in side-channel security evaluations. In *Advances in Cryptology—CRYPTO 2019*, volume 11692 of *Lecture Notes in Computer Science*, pages 713–737, 2019.

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[BPS+20] Ryad Benadjila, Emmanuel Prouff, Rémi Strullu, Eleonora Cagli, and Cécile Dumas. Deep learning for side-channel analysis and introduction to ASCAD database. *Journal of Cryptographic Engineering*, 10(2):163–188, 2020.

[CDP17] Eleonora Cagli, Cécile Dumas, and Emmanuel Prouff. Convolutional neural networks with data augmentation against jitter-based countermeasures. In *Cryptographic Hardware and Embedded Systems – CHES 2017*, volume 10529 of *Lecture Notes in Computer Science*, pages 45–68. Springer, 2017.

[CRR02] Suresh Chari, Josyula R. Rao, and Pankaj Rohatgi. Template attacks. In *International Workshop on Cryptographic Hardware and Embedded Systems*, LNCS, pages 13–28, 2002.

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006.

[dCGRP19] Eloi de Chérisey, Sylvain Guilley, Olivier Rioul, and Pablo Piantanida. Best information is most successful: Mutual information and success rate in side-channel analysis. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2019(2):49–79, 2019.

[DFS15] Alexandre Duc, Sebastian Faust, and François-Xavier Standaert. Making masking security proofs concrete: Or how to evaluate the security of any leakage device. In *Advances in Cryptology—EUROCRYPT 2015*, volume 9056 of *Lecture Notes in Computer Science*, pages 401–429. Springer, 2015.

[DK18] Daniel Dinu and Ilya Kizhvatov. EM analysis in the IoT context: Lessons learned from an attack on thread. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2018(1):73–97, 2018.

[Fer96] Thomas S Ferguson. *A course in large sample theory*. London: Chapman and Hall, 1996.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[GBTP08] Benedikt Gierlichs, Lejla Batina, Pim Tuyls, and Bart Preneel. Mutual information analysis: A generic side-channel distinguisher. In *International Conference on Cryptographic Hardware and Embedded Systems*, volume 5154 of *Lecture Notes in Computer Science*, 2008.

[git21] Curse_of_re-encryption. https://github.com/ECSIS-lab/curse_of_re-encryption, 2021.

[GPSW17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330, 2017.

[HHGG20] Benjamin Hettwer, Tobias Horn, Stefan Gehrer, and Tim Güneysu. Encoding power traces as images for efficient side-channel analysis. In *2020 IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*, pages 46–56, 2020.

[HRG14] Annelie Heuser, Olivier Rioul, and Sylvain Guilley. Good is not good enough: Deriving optimal distinguishers from communication theory. In *International Workshop on Cryptographic Hardware and Embedded Systems*, pages 55–74, 2014.

[ISUH21] Akira Ito, Kotaro Saito, Rei Ueno, and Naofumi Homma. Imbalanced data problems in deep learning-based side-channel attacks: Analysis and solution. *IEEE Transactions on Information Forensics and Security*, pages 1–1, 2021.

[IUH21] Akira Ito, Rei Ueno, and Naofumi Homma. 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.

[IUH22] Akira Ito, Rei Ueno, and Naofumi Homma. On the success rate of side-channel attacks on masked implementations: Information-theoretical bounds and their practical usage. Cryptology ePrint Archive, Report 2022/576, 2022. https://eprint.iacr.org/2022/576.

[KWPP21] Maikel Kerkhof, Lichao Wu, Guilherme Perin, and Stjepan Picek. No (good) loss no gain: Systematic evaluation of loss functions in deep learning-based side-channel analysis. Cryptology ePrint Archive, Paper 2021/1091, 2021. https://eprint.iacr.org/2021/1091.

[MDP20] Loïc Masure, Cécile Dumas, and Emmanuel Prouff. A comprehensive study of deep learning for side-channel analysis. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2020(1):348–375, 2020.

[MHM14] Zdenek Martinasek, Jan Hajny, and Lukas Malina. Optimization of power analysis using neural network. In Aurélien Francillon and Pankaj Rohatgi, editors, *Smart Card Research and Advanced Applications*, pages 94–107, Cham, 2014. Springer International Publishing.

[OP11] David Oswald and Christof Paar. Breaking Mifare DESFire MF3ICD40: Power analysis and templates in the real world. In *International Workshop on Cryptographic Hardware and Embedded Systems*, volume 6917 of *LNCS*, pages 207–222, 2011.

[PHJ+19] Stjepan Picek, Annelie Heuser, Alan Jovic, Shivam Bhasin, and Francesco Regazzoni. The Curse of Class Imbalance and Conflicting Metrics with Machine Learning for Side-channel Evaluations. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, (1):209–237, 2019.

[RSVC+11] Mathieu Renauld, François-Xavier Standeart, Nicolas Veyrat-Charvillon, Dina Kamel, and Denis Flandre. A formal study of power variability issues and side-channel attacks for nanoscale devices. In *Advances in Cryptology—Eurocrypt 2011*, volume 6632 of *Lecture Notes in Computer Science*, pages 109–128, 2011.

[RWPP21] Jorai Rijsdijk, Lichao Wu, Guilherme Perin, and Stjepan Picek. Reinforcement learning for hyperparameter tuning in deep learning-based side-channel analysis. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2021.

[SMY09] François-Xavier Standeart, Tal G. Malkin, and Moti Yung. A unified framework for the analysis of side-channel key recovery attacks. In *Advances in Cryptology—Eurocrypt 2009*, volume 5479 of *Lecture Notes in Computer Science*, pages 443–461, 2009.

[UHA17] Rei Ueno, Naofumi Homma, and Takafumi Aoki. Toward more efficient DPA-resistant AES hardware architecture based on threshold implementation. In *International Workshop on Constructive Side-Channel Analysis and Secure Design*, volume 10348 of *Lecture Notes in Computer Science*, pages 50–64, 2017.

[UXT+22] Rei Ueno, Keita Xagawa, Yutaro Tanaka, Akira Ito, Junko Takahashi, and Naofumi Homma. Curse of re-encryption: A generic power/EM analysis on post-quantum KEMs. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 1:296–322, 2022.

[WAGP20] Lennert Wouters, Victors Arribas, Benedikt Gierlichs, and Bart Praneel. Revisiting a methodology for efficient CNN architectures in profiling attacks. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2020(3):147–168, 2020.

[WVdHG+20] Lennert Wouters, Jan Van den Herrewegen, Flavio D. Garcia, David Oswald, Benedikt Gierlichs, and Bart Preneel. Dismantling DST80-based immobiliser systems. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2020(2):99–127, 2020.

[ZBD+21] Gabriel Zaid, Lilian Bossuet, François Dassance, Amaury Habrard, and Alexandre Venelli. Ranking loss: Maximizing the success rate in deep learning side-channel analysis. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, (1):25–55, 2021.

[ZBHV19] Gabriel Zaid, Lilian Bossuet, Amaury Habrard, and Alexandre Venelli. Methodology for Efficient CNN Architectures in Profiling Attacks. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2020(1):1–36, 2019.

[ZZN+20] Jiajia Zhang, Mengce Zheng, Jiehui Nan, Honggang Hu, and Nenghai Yu. 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*, 2020(3):73—-96, 2020.

Appendix A: Formal proof of Proposition 3

To prove Proposition 3, we introduce the following two lemmas.

Lemma 5 (Extension of Lebesgue's dominated convergence theorem)**.** *Let Λ be a subset of $\mathbb{R} \cup \{-\infty, \infty\}$, and let $b \in \overline{\Lambda}$ be a point of the closure of Λ denoted as $\overline{\Lambda}$. Let $\{X_\lambda\}_{\lambda \in \Lambda}$ be a family of random variable. Suppose that $\lim_{\lambda \to b} X_\lambda = X$ holds almost surely where X denotes a random variable, and there exists an integrable random variable Y such that, for all $\lambda \in \Lambda$, $|X_\lambda| \leq Y$ almost surely. We then have $\mathbb{E}X_\lambda \to \mathbb{E}X$ as $\lambda \to b$.*

Proof. Let $\{\lambda_i\}_{i=1}^{\infty} \subset \Lambda$ be any sequence converging to b. We have $\lim_{i \to \infty} X_{\lambda_i} = X$ almost surely. because $\lim_{\lambda \to b} X_\lambda = X$ almost surely. Therefore, from Lebesgue's dominated convergence theorem, we have $\lim_{i \to \infty} \mathbb{E}X_{\lambda_i} = \mathbb{E} \lim_{i \to \infty} X_{\lambda_i} = \mathbb{E}X$. Since this holds for any sequence $\{\lambda_i\}$, we have $\mathbb{E}X_\lambda \to \mathbb{E}X$. □

Lemma 6 (Extension of Fatou's lemma)**.** *Let Λ be a subset of $\mathbb{R} \cup \{-\infty, \infty\}$, and let $b \in \overline{\Lambda}$ be a point of the closure of Λ denoted as $\overline{\Lambda}$. Let $\{X_\lambda\}_{\lambda \in \Lambda}$ be a family of random variable, where $X_\lambda > 0$ holds almost surely for all $\lambda \in \Lambda$. If $\liminf_{\lambda \to b} X_\lambda$ is measurable, we have $\liminf_{\lambda \to b} \mathbb{E}X_\lambda \geq \mathbb{E} \liminf_{\lambda \to b} X_\lambda$.*

Proof. Let $\{\lambda_i\}_{i=1}^{\infty} \subset \Lambda$ be a sequence converging to b such that $\liminf_{i \to \infty} \mathbb{E}X_{\lambda_i} = \liminf_{\lambda \to b} \mathbb{E}X_\lambda$. Note that we have $\liminf_{\lambda \to b} X_\lambda \leq \liminf_{i \to \infty} X_{\lambda_i}$. Hence, from Fatou's lemma, we have

$$\mathbb{E} \liminf_{\lambda \to b} X_\lambda \leq \mathbb{E} \liminf_{i \to \infty} X_{\lambda_i} \leq \liminf_{i \to \infty} \mathbb{E}X_{\lambda_i} = \liminf_{\lambda \to b} \mathbb{E}X_\lambda.$$

□

We then prove Proposition 3.

Proof. We prove only Limits (7) and (9) because Limits (8) and (10) is trivially proved using them. Let $(\Omega, \mathcal{F}, \Pr)$ be a probability space. For a simplified notation, we denote $\mathscr{H}_\beta[r_{Z|\boldsymbol{X}}]$ by r_β in this proof.

We first prove Limit (7), which is represented as

$$\lim_{\beta \searrow 0} \mathrm{CE}(r_\beta) = \lim_{\beta \searrow 0} \mathbb{E} - \log r_\beta(Z \mid \boldsymbol{X}).$$

To use Lemma 2, we examine whether the limit $\lim_{\beta\searrow 0}$ and expectation $\mathbb{E}$ can be interchanged. Lemma 5 states that we can interchange them if there exists an integrable random variable Y such that $\sup_\beta - \log r_\beta(Z \mid \boldsymbol{X}) \leq Y$ holds almost surely. we first consider the range of β. Limit (7) can be rewritten as $\forall \epsilon > 0, \exists \delta_\epsilon > 0, \forall \beta > 0; (0 < \beta < \delta_\epsilon \Rightarrow |\mathrm{CE}(r_\beta) - n_z| < \epsilon)$. If there exists such δ_ϵ, this holds even when we replace δ_ϵ with $\min(\{1, \delta_\epsilon\})$. Therefore, without loss of generality, we assume that $\beta \in (0, 1)$. We then consider the supremum of $-\log r_\beta(Z \mid \boldsymbol{X})$. From the definition of r_β, we have

$$-\log r_\beta(Z \mid \boldsymbol{X}) = -\beta \log r(Z \mid \boldsymbol{X}) + \log \sum_{z'} r(z' \mid \boldsymbol{X})^\beta.$$

Here, $\sum_{z'} r(z' \mid \boldsymbol{X})^\beta$ is a monotonically decreasing function of β because it is a sum of decreasing functions $r(Z \mid \boldsymbol{X})^\beta$ (assuming that $r(Z \mid \boldsymbol{X}) \in (0, 1)$). Thus, $\sum_{z'} r(z' \mid \boldsymbol{X})^\beta < \sum_{z'} r(z' \mid \boldsymbol{X})^0 = 2^{n_z}$, which is followed by $\log \sum_{z'} r(z' \mid \boldsymbol{X})^\beta < n_z$. Therefore, it holds

$$-\log r_\beta(Z \mid \boldsymbol{X}) < -\beta \log r(Z \mid \boldsymbol{X}) + n_z \leq -\log r(Z \mid \boldsymbol{X}) + n_z.$$

Lemma 5 holds if we consider $Y = -\log r(Z \mid \boldsymbol{X}) + n_z$, where $-\log r(Z \mid \boldsymbol{X}) + n_z$ is an integrable random variable. Lemma 5, Lemma 2, and the continuous mapping theorem yield that

$$\lim_{\beta \searrow 0} \mathrm{CE}(r_\beta) = -\mathbb{E} \lim_{\beta \searrow 0} \log r_\beta(Z \mid \boldsymbol{X}) = n_z,$$

as required.

We then prove Limit (9). In this proof, we consider $\log 0$ as $\lim_{x\searrow 0} \log x = -\infty$ because we always consider the logarithm of positive real numbers. To use Lemma 6, we firstly show that

$$\limsup_{\beta \to \infty} \log r_\beta(Z \mid \boldsymbol{X}) = \lim_{\beta \to \infty} \log r_\beta(Z \mid \boldsymbol{X}) = \log \mathbb{1}_{\{Z = \arg\max_{z'} r(z' \mid \boldsymbol{X})\}}, \tag{17}$$

holds almost surely. From Lemma 2, there exists a null set $\mathcal{N}$ such that $r_\beta(z \mid \boldsymbol{x}) \to \mathbb{1}_{\{z = \arg\max_{z'} r_{Z|\boldsymbol{X}}(z' \mid \boldsymbol{x})\}}$ holds as $\beta \to \infty$, where $(z, \boldsymbol{x}) \in (Z, \boldsymbol{X})(\Omega \setminus \mathcal{N})$. For arbitrary $\omega \in \Omega \setminus \mathcal{N}$, let $\boldsymbol{x} = \boldsymbol{X}(\omega)$ and $z = Z(\omega)$. We divide the situation into two cases: (a) $z = \arg\max_{z'} r(z' \mid \boldsymbol{x})$ and (b) $z \neq \arg\max_{z'} r(z' \mid \boldsymbol{x})$. (a) If $z = \arg\max_{z'} r(z' \mid \boldsymbol{x})$, then $\lim_{\beta\to\infty} r_\beta(z \mid \boldsymbol{x}) = 1$ holds. In this case, we have $\lim_{\beta\to\infty} \log r_\beta(z \mid \boldsymbol{x}) = \log \lim_{\beta\to\infty} r_\beta(z \mid \boldsymbol{x}) = 0$ because the log function is continuous at a point of 1. (b) If $z \neq \arg\max_{z'} r(z' \mid \boldsymbol{x})$, then $\lim_{\beta\to\infty} r_\beta(z \mid \boldsymbol{x}) = 0$ holds. Note that $r_\beta(z \mid \boldsymbol{x})$ approaches 0 from the right side because $r_\beta(z \mid \boldsymbol{x}) > 0$. Therefore, we have $\lim_{\beta\to\infty} \log r_\beta(z \mid \boldsymbol{x}) = \log \lim_{\beta\to\infty} r_\beta(z \mid \boldsymbol{x}) = \log 0 = -\infty$. This discussion proves Equation (17). From Equation (17), we can easily confirm that $\liminf_{\beta\to\infty} - \log r_\beta(Z|\boldsymbol{X}) = -\limsup_{\beta\to\infty} \log r_\beta(Z|\boldsymbol{X}) = \log \mathbb{1}_{\{Z = \arg\max_{z'} r(z' \mid \boldsymbol{X})\}}$ is a measurable function. Thus, Lemma 6 yields that

$$\lim_{\beta \to \infty} \mathrm{CE}(r_\beta) = \lim_{\beta \to \infty} \mathbb{E} - \log r_\beta(Z \mid \boldsymbol{X}) \geq -\mathbb{E} \limsup_{\beta \to \infty} \log r_\beta(Z \mid \boldsymbol{X}). \tag{18}$$

From Inequality (18) and Equation (17), $\lim_{\beta\to\infty} \mathrm{CE}(r_\beta)$ can be bounded as follows:

$$\lim_{\beta \to \infty} \mathrm{CE}(r_\beta) \geq -\mathbb{E} \log \mathbb{1}_{\{Z = \arg\max_{z'} r(z' \mid \boldsymbol{X})\}}.$$

According to the assumption that $\Pr(Z \neq \arg\max_{z'} r(z' \mid \boldsymbol{X})) > 0$, we have

$$\begin{aligned}\Pr(Z \neq \arg\max_{z'} r(z' \mid \boldsymbol{X})) &= \Pr(\mathbb{1}_{\{Z=\arg\max_{z'} r(z'|\boldsymbol{X})\}} = 0) \\ &= \Pr(-\log \mathbb{1}_{\{Z=\arg\max_{z'} r(z'|\boldsymbol{X})\}} = \infty) > 0.\end{aligned}$$

We show that $\lim_{\beta\to\infty} \mathrm{CE}(r_\beta)$ is unbounded (*i.e.*, $\lim_{\beta\to\infty} \mathrm{CE}(r_\beta) = \infty$) by *reductio ad absurdum*. Suppose that it is bounded (*i.e.*, $\lim_{\beta\to\infty} \mathrm{CE}(r_\beta) < \infty$). Let an event $E = \left\{ -\log \mathbb{1}_{\{Z=\arg\max_{z'} r(z'|\boldsymbol{X})\}} = \infty \right\}$. Then,

$$\begin{aligned}-\mathbb{E} \log \mathbb{1}_{\{Z=\arg\max_{z'} r(z'|\boldsymbol{X})\}} &= \int_\Omega -\log \mathbb{1}_{\{Z(\omega)=\arg\max_{z'} r(z'|\boldsymbol{X}(\omega))\}} \Pr(d\omega) \\ &= \int_{\Omega\setminus E} -\log \mathbb{1}_{\{Z(\omega)=\arg\max_{z'} r(z'|\boldsymbol{X}(\omega))\}} \Pr(d\omega) \\ &\qquad + \int_E -\log \mathbb{1}_{\{Z(\omega)=\arg\max_{z'} r(z'|\boldsymbol{X}(\omega))\}} \Pr(d\omega) \\ &\geq \int_E -\log \mathbb{1}_{\{Z(\omega)=\arg\max_{z'} r(z'|\boldsymbol{X}(\omega))\}} \Pr(d\omega). \qquad (19)\end{aligned}$$

In Inequality (19), if the assumption of *reductio ad absurdum* is true (*i.e.*, $\lim_{\beta\to\infty} \mathrm{CE}(r_\beta) < \infty$), it should hold $\int_E -\log \mathbb{1}_{\{Z(\omega)=\arg\max_{z'} r(z'|\boldsymbol{X}(\omega))\}} \Pr(d\omega) < \infty$. For arbitrary $\omega \in E$, we should have $n < -\log \mathbb{1}_{\{Z(\omega)=\arg\max_{z'} r(z'|\boldsymbol{X}(\omega))\}}$ for every $n \in \mathbb{N}$ because it holds $-\log \mathbb{1}_{\{Z(\omega)=\arg\max_{z'} r(z'|\boldsymbol{X}(\omega))\}} = \infty$. Therefore, it should hold

$$n \Pr(E) = \int_E n \Pr(d\omega) < \int_E -\log \mathbb{1}_{\{Z(\omega)=\arg\max_{z'} r(z'|\boldsymbol{X}(\omega))\}} \Pr(d\omega) < \infty,$$

for arbitrary $n \in \mathbb{N}$. This should be followed by $\Pr(E) = 0$, which contradicts the assumption that $\Pr(E) > 0$. Thus, we conclude $\lim_{\beta\to\infty} \mathrm{CE}(r_\beta) = \infty$. □

Appendix B: Formal proof of Theorem 3

Before the proof, we introduce Lemma 7.

Lemma 7 ([Fer96, Theorem 16(b)]). *Let $X_1, X_2, \ldots$ be a sequence of i.i.d random variables with common distribution function. Let Θ is the set of parameters, and let $U(x,\theta)$ be a measurable function in x for all $\theta \in \Theta$. Assume that $\mathbb{E}U(X,\theta)$ exists, and $\mathbb{E}U(X,\theta)$ is finite for all $\theta \in \Theta$. Suppose that*

1. *Θ is compact,*
2. *$U(x,\theta)$ is upper semi-continuous in θ for all x,*
3. *there exists an integrable function $K(X)$ such that $U(x,\theta) < K(x)$ holds for all x and θ,*
4. *for all θ and sufficiently small $\rho > 0$, $\sup_{|\theta'-\theta|<\rho} U(x,\theta')$ is measurable in x.*

Then, we have

$$\Pr\left(\limsup_{m\to\infty} \sup_{\theta\in\Theta} \frac{1}{m} \sum_{i=1}^m U(X_i,\theta) \leq \sup_{\theta\in\Theta} \mathbb{E}U(X,\theta) \right) = 1.$$

Using Lemma 7, we prove Theorem 3.

Proof. This proof is on the basis of [Fer96, Theorem 17]. Let $U(\boldsymbol{x}, z, \beta) = \ln r_\beta(z \mid \boldsymbol{x}) - \ln r_{\beta_0}(z \mid \boldsymbol{x})$, where $r_\beta(z \mid \boldsymbol{x}) = \mathscr{H}_\beta(r)$. Note that, for all m,

$$\frac{1}{m}\sum_{i=1}^{m} U(\boldsymbol{X}_i, Z_i, \hat{\beta}_m) = \sup_{\beta\in\mathcal{B}} \frac{1}{m}\sum_{i=1}^{m} U(\boldsymbol{X}, Z, \beta).$$

To use Lemma 7, we confirm the conditions for the lemma. The first and second conditions are obviously satisfied. Function $U(\boldsymbol{x}, z, \beta)$ is continuous in β for all $\boldsymbol{x}$ and z, which implies that the forth condition is also satisfied because we have

$$\sup_{|\beta'-\beta|<\rho} U(\boldsymbol{x}, z, \beta') = \sup_{\beta'\in\mathcal{D}} U(\boldsymbol{x}, z, \beta'),$$

for any countable set $\mathcal{D}$ which is dense in $\{\, \beta' \mid |\beta' - \beta| < \rho \,\}$. We then examine the third condition. Fix $\beta \in \mathcal{B}$. For any $\boldsymbol{x}$ and z, we have

$$\begin{aligned} U(\boldsymbol{x}, z, \beta) &= \ln r_\beta(z \mid \boldsymbol{x}) - \ln r_{\beta_0}(z \mid \boldsymbol{x}) \\ &= (\beta - \beta_0)\ln r(z \mid \boldsymbol{x}) - \ln \sum_{z'} r^\beta(z' \mid \boldsymbol{x}) + \ln \sum_{z'} r^{\beta_0}(z' \mid \boldsymbol{x}) \\ &= (\beta - \beta_0)\ln r(z \mid \boldsymbol{x}) + \ln \frac{\sum_{z'} r^{\beta_0}(z' \mid \boldsymbol{x})}{\sum_{z'} r^\beta(z' \mid \boldsymbol{x})}. \end{aligned}$$

Using the log-sum inequality [CT06, Theorem 2.7.1], we have

$$\begin{aligned} U(\boldsymbol{x}, z, \beta) &\le (\beta - \beta_0)\ln r(z \mid \boldsymbol{x}) + \sum_{z'} \left(\frac{r^{\beta_0}(z' \mid \boldsymbol{x})}{\sum_{z'} r^{\beta_0}(z' \mid \boldsymbol{x})} \right) \ln \frac{r^{\beta_0}(z' \mid \boldsymbol{x})}{r^\beta(z' \mid \boldsymbol{x})} \\ &\le -M_\beta \ln r(z \mid \boldsymbol{x}) + (\beta_0 - \beta) \sum_{z'} \left(\frac{r^{\beta_0}(z' \mid \boldsymbol{x})}{\sum_{z'} r^{\beta_0}(z' \mid \boldsymbol{x})} \right) \ln r(z' \mid \boldsymbol{x}) \\ &< -M_\beta \ln r(z \mid \boldsymbol{x}) - M_\beta \sum_{z'} \ln r(z' \mid \boldsymbol{x}). \end{aligned}$$

Therefore, $K(\boldsymbol{x}, z) = -M_\beta(\ln r(z \mid \boldsymbol{x}) + \sum_{z'} \ln r(z' \mid \boldsymbol{x}))$ satisfies the third condition because, for all $\boldsymbol{x}$ and z, it holds $U(\boldsymbol{x}, z, \beta) \le K(\boldsymbol{x}, z)$ and $\mathbb{E}K(\boldsymbol{X}, Z) = -M_\beta(\mathbb{E}\ln r(Z \mid \boldsymbol{X}) + \mathbb{E}\sum_{z'} \ln r(z' \mid \boldsymbol{X})) < \infty$. Thus, the function $U(\boldsymbol{x}, z, \beta)$ satisfies the conditions of Lemma 7.

Let $\mathcal{V} = \{\, \beta \mid |\beta - \beta_0| \ge \rho \,\}$. Because $\mathcal{V}$ is compact, Lemma 7 yeilds that

$$\Pr\left(\limsup_{m\to\infty} \sup_{\beta\in\mathcal{V}} \frac{1}{m}\sum_{i=1}^{m} U(\boldsymbol{X}_i, Z_i, \beta) \le \sup_{\beta\in\mathcal{V}} \mathbb{E}U(\boldsymbol{X}, Z, \beta) \right) = 1. \tag{20}$$

Note that, for all $\beta \in \mathcal{V}$, it holds $\mathbb{E}U(\boldsymbol{X}, Z, \beta) < 0$ because $\mathbb{E}U(\boldsymbol{X}, Z, \beta) = -\ln(2)(\mathrm{CE}(r_\beta) - \mathrm{CE}(r_{\beta_0}))$ and $\mathrm{CE}(r_\beta)$ is strictly convex in β, which takes the minimum value only at β_0. From Equation (20), with probability 1, there exists $N \in \mathbb{N}$ such that, for all $m > N$,

$$\sup_{\beta\in\mathcal{V}} \frac{1}{m}\sum_{i=1}^{m} U(X_i, Z_i, \beta) \le \sup_{\beta\in\mathcal{V}} \mathbb{E}U(X, Z, \beta) < 0.$$

However, we also have

$$\frac{1}{m}\sum_{i=1}^{m} U(X_i, Z_i, \hat{\beta}_m) = \sup_{\beta\in\mathcal{B}} \frac{1}{m}\sum_{i=1}^{m} U(X_i, Z_i, \beta) \ge 0.$$

Thus, for all $m > N$, $\hat{\beta}_m \notin \mathcal{V}$ and $|\hat{\beta}_m - \beta_0| < \rho$ hold almost surely. Since ρ is arbitrary, we conclude that $\hat{\beta}_m \xrightarrow{\text{a.s.}} \beta_0$ as $m \to \infty$.

□