Formally Bounding the Side-Channel Leakage in Unknown-Message Attacks

Michael Backes
MPI-SWS and
Saarland University
backes@cs.uni-sb.de

Boris Köpf
MPI-SWS
bkoepf@mpi-sws.mpg.de

Abstract

We present a novel information measure that captures the quantity of secret information that an unknown-message attacker can extract from a system in a given number of side-channel measurements. We provide an algorithm to compute this measure, and we use it to analyze hardware implementations of cryptographic algorithms with respect to their vulnerabilities against power and timing attacks. In particular, we show that message-blinding – the common countermeasure against timing attacks – does not constitute a provably secure protection against timing attacks, but that it only reduces the rate at which information is leaked in multiple measurements. Finally, we compare information measures corresponding to different kinds of side-channel attackers and show that they form a strict hierarchy.

1 Introduction

Side-channel attacks against cryptographic algorithms aim at breaking cryptography by exploiting information that is revealed by the algorithm's physical execution. Characteristics such as running time [15, 4, 24], power consumption [16], and electromagnetic radiation [14, 26] have all been exploited to recover secret keys from implementations of different cryptographic algorithms. Side-channel attacks are now so effective that they pose a real threat to the security of devices which can be subjected to different kinds of measurements. This threat is not covered by traditional notions of cryptographic security, and models for reasoning about the resistance against such attacks are only now emerging [22, 29, 17].

Two quantities characterize the attacker's effort for successfully mounting a side-channel attack and recovering a secret key from a given system. The first is the computational power needed to recover the key from the information that is revealed through the side-channel. The second is the number of measurements

needed to gather sufficient side-channel information for this task. To prove that a system is resistant to side-channel attacks, one must ensure that the overall effort for a successful attack is out of the range of realistic attackers.

The attacker's computational power is typically not the limiting factor in practice, as a large number of documented attacks show [4, 6, 8, 16, 24]. Hence, the security of a system often entirely depends on the amount of secret information that an attacker can gather in his side-channel measurements. Note that the number of measurements may be bounded – for example, by the number of times the system re-uses a session key – and must be considered when reasoning about a system's vulnerability to side-channel attacks.

A model to express the revealed information as a function of the number of side-channel measurements has recently been proposed, and it has been applied to characterize the resistance of cryptographic algorithms against side-channel attacks [17]. The model captures attackers that can interact with the system by adaptively choosing the messages that the system decrypts (or encrypts).

However, many documented side-channel attacks are *unknown-message*, i.e., the attacker cannot see or control the messages that are decrypted (or encrypted) by the system, and they typically involve multiple side-channel measurements. Such attacks in particular comprise differential power attacks [16] as well as timing attacks against public-key cryptosystems that are run with state-of-the-art countermeasures such as message blinding. Quantifying the information that a system reveals in such an attack was an open problem prior to this work.

1.1 Our Contributions

We propose a novel measure for quantifying the resistance of systems against unknown-message side-channel attacks. This measure Λ captures the quantity of secret information that a system reveals as a function of the number of side-channel measurements. Moreover, we provide an explicit formula for Λ when the number of measurements tends to infinity, corresponding to the maximum amount of secret information that is eventually leaked.

In order to apply our measure to realistic settings, we provide algorithms for computing Λ for finite and infinite numbers of measurements, respectively. We subsequently use these algorithms to formally analyze the resistance of non-trivial hardware implementations of cryptographic algorithms to side-channel attacks: First, we show that an AES SBox falls prey to a power analysis in that the key is fully determined by a sufficiently large number of measurements. Second, we show that finite-field exponentiation as used in, e.g., the generalized ElGamal decryption algorithm, falls prey to timing attacks in that the key is fully determined by a sufficiently large numbers of measurements. We use this result to show that message-blinding, which aims at protecting against timing attacks by decoupling the running time of the exponentiation algorithm from the secret, does not constitute a suitable technique in general to protect against chosen-message timing attacks: we can show that, for the analyzed exponentiation algorithm, message-blinding only reduces the rate at which information about the secret is revealed, and that the entire key information is still even-

tually leaked. This yields the first formal assessment of the (un-)suitability of message-blinding to counter timing attacks.

We conclude by putting our novel measure Λ into perspective with information measures for different kinds of attacks. The result is a formal hierarchy of side-channel attackers that is ordered in terms of the information they can extract from a system. We distinguish unknown-message attacks, in which the attacker does not even know the messages (as is typically the case in power attacks), known-message attacks, in which the attacker knows but cannot influence the messages, and chosen-message attacks, in which the attacker can adaptively choose the messages (as is typically the case in timing attacks). As expected, more comprehensive attackers are capable of extracting more information in a given number of measurements. Moreover, we show that this inclusion is strict for certain side-channels. We believe that clarifying on the different attack scenarios will provide guidance on which measure to pick for which application scenario.

1.2 Related Work

While there has been substantial work in information-flow security on detecting or quantifying information leaks, there are no results for quantifying the information leakage in unknown-message attacks. Lowe [18] quantifies information flow in a possibilistic process algebra by counting the number of distinguishable behaviors. Clarkson et al. [10] develop a model for reasoning about an adaptive attacker's beliefs about the secret, which may be right or wrong. The information measure proposed by Clark et al. [9] is closest to ours, however, it is not applicable to side-channel attacks as it does not capture multiple computations with the same key. The measures proposed in [17] provide tools for reasoning about the information leakage in side-channel attacks, but only for stronger, chosen-message, attackers.

There is a large body of work on side-channel cryptanalysis, in particular on attacks and countermeasures. However, models and theoretical bounds on what side-channel attackers only started to emerge. Chari et al. [7] are the first to present methods for proving hardware implementations secure. They propose a generic countermeasure for power attacks and prove that it resists a given number of side-channel measurements. Micali et al. [22] propose physically observable cryptography, a mathematical model that aims at providing provably secure cryptography on hardware that is only partially shielded. Their model has been been specialized in a line of work by Standaert et al. [29, 19, 25]. In [25], the success rate of a side-channel attacker with access to multiple measurements is analyzed. The analysis is based on leakage functions that express what information the side-channel reveals, e.g., the input's Hamming weight. Our approach tackles the problem on a different level of abstraction: starting from a model of the physical characteristics of the system, e.g. its power or time consumption, we capture the information that the corresponding side-channel reveals. It is subject of future work to investigate whether our approach can be used for instantiating the leakage functions of [25].

1.3 Outline

The paper is structured as follows. In Section 2, we introduce our models of side-channels and attackers and we review basics of information theory. In Section 3, we present measures for quantifying the information leakage in unknown-message attacks. In Section 4, we show how these measures can be computed for given implementations. We report on experimental results in Section 5 and compare different kinds of side-channel attacks in Section 6. We conclude in Section 7.

2 Preliminaries

We start by describing our models of side-channels and attackers, and we briefly recall some basic information theory.

2.1 Modeling Side-channels and Attackers

We consider systems that compute functions of type $F\colon K \times M \to D$ for finite sets K, M, and D. We assume that the attacker can make physical observations about F's implementation I_F that are associated with the computation of $F(k, m)$. We assume that the attacker can make one observation per invocation of the function F and that no measurement errors occur. Examples of such observations are the power or the time consumption of I_F during the computation (see [16, 21] and [15, 6, 4, 24], respectively).

A *side-channel* is a function $f_{I_F}\colon K \times M \to O$, where O denotes the set of possible observations. We assume that the attacker has full knowledge about the implementation I_F, i.e., f_{I_F} is known to the attacker. We will usually leave I_F implicit and abbreviate f_{I_F} by f.

Example 1. Suppose that F is implemented in synchronous (clocked) hardware and that the attacker is able to determine I_F's running times up to single clock ticks. Then the timing side-channel of I_F can be modeled as a function $f\colon K \times M \to \mathbb{N}$ that represents the number of clock ticks consumed by an invocation of F. A hardware simulation environment can be used to compute f.

Example 2. Suppose F is given in a description language for synchronous hardware. Power estimation techniques such as [23, 30] can be used to determine a function $f\colon K \times M \to \mathbb{R}^n$ that estimates an implementation's power consumption during n clock ticks.

In a side-channel attack, a malicious agent gathers side-channel observations $f(k, m_1), \ldots, f(k, m_n)$ for deducing k or narrowing down its possible values. Depending on the attack scenario, the attacker might additionally be able to see or choose the messages $m_i \in M$: an attack is *unknown-message* if the attacker cannot observe $m_i \in M$; an attack is *known-message* if the attacker can observe but cannot influence the choice of $m_i \in M$; an attack is *chosen-message* if the

attacker can choose $m_i \in M$. Most documented timing and cache attacks are chosen-message or known-message attacks [4, 13, 15], whereas power attacks are often unknown-message attacks [16, 28].

In this paper, we focus on the open problem of giving bounds on the side-channel leakage in unknown-message attacks. In Section 6, we will come back to the distinction between different attack types and formally compare them with respect to the quantity of information that they can extract from a system.

2.2 Information Theory Basics

Let A be a finite set and $p\colon A \to \mathbb{R}$ a probability distribution. For a random variable $\mathcal{X}\colon A \to X$, we define $p_{\mathcal{X}}\colon X \to \mathbb{R}$ as $p_{\mathcal{X}}(x) = \sum_{a \in \mathcal{X}^{-1}(x)} p(a)$, which is often denoted by $p(\mathcal{X} = x)$ in the literature.

The *(Shannon) entropy* of a random variable $\mathcal{X}\colon A \to X$ is defined as

$$H(\mathcal{X}) = -\sum_{x \in X} p_{\mathcal{X}}(x) \log_2 p_{\mathcal{X}}(x) \ .$$

The entropy is a lower bound for the average number of bits required for representing the results of independent repetitions of the experiment associated with $\mathcal{X}$. Thus, in terms of guessing, the entropy $H(\mathcal{X})$ is a lower bound for the average number of binary questions that need to be asked to determine $\mathcal{X}$'s value after the attack [5]. If $\mathcal{Y}\colon A \to Y$ is another random variable, $H(\mathcal{X}|\mathcal{Y} = y)$ denotes the entropy of $\mathcal{X}$ given $\mathcal{Y} = y$, i.e., with respect to the distribution $p_{\mathcal{X}|\mathcal{Y}=y}$. The *conditional entropy* $H(\mathcal{X}|\mathcal{Y})$ of $\mathcal{X}$ given $\mathcal{Y}$ is defined as the expected value of $H(\mathcal{X}|\mathcal{Y} = y)$ over all $y \in Y$, namely,

$$H(\mathcal{X}|\mathcal{Y}) = \sum_{y \in Y} p_{\mathcal{Y}}(y) H(\mathcal{X}|\mathcal{Y} = y) \ .$$

Entropy and conditional entropy are related by the equation $H(\mathcal{X}\mathcal{Y}) = H(\mathcal{Y}) + H(\mathcal{X}|\mathcal{Y})$, where $\mathcal{X}\mathcal{Y}$ is the random variable defined as $\mathcal{X}\mathcal{Y}(k) = (\mathcal{X}(k), \mathcal{Y}(k))$. *The mutual information* $I(\mathcal{X}; \mathcal{Y})$ of $\mathcal{X}$ and $\mathcal{Y}$ is defined as the reduction of uncertainty about $\mathcal{X}$ if one learns $\mathcal{Y}$, i.e., $I(\mathcal{X}; \mathcal{Y}) = H(\mathcal{X}) - H(\mathcal{X}|\mathcal{Y})$. The *relative entropy* or *Kullback-Leibler distance* $D(p_{\mathcal{X}} \parallel q_{\mathcal{X}})$ between two probability distributions $p_{\mathcal{X}}$ and $q_{\mathcal{X}}$ is given by $D(p_{\mathcal{X}} \parallel q_{\mathcal{X}}) = \sum_{x \in X} p_{\mathcal{X}}(x) \frac{p_{\mathcal{X}}(x)}{q_{\mathcal{X}}(x)}$. The relative entropy is always nonnegative, and it is zero if and only if $p_{\mathcal{X}} = q_{\mathcal{X}}$.

3 Information Leakage in Unknown-message Attacks

In this section, we first propose a novel information measure that expresses the information gain of an unknown-message attacker as a function of the number of side-channel observations made. Subsequently, we derive an explicit representation for the limit of this information gain for an unbounded number of

observations. This representation provides a characterization of the secret information that the side-channel eventually leaks. Moreover, it leads to a simple algorithm for computing this information.

3.1 Information Gain in n Observations

In the following, let $p_K\colon K \to \mathbb{R}$ and $p_M\colon M \to \mathbb{R}$ be probability distributions and let the random variables $\mathcal{K} = id_K$, $\mathcal{M} = id_M$ model the choice of keys and messages, respectively; we assume that p_M and p_K are known to the attacker. For $n \in N$, let $\mathcal{O}_n\colon K \times M^n \to O^n$ be defined by $\mathcal{O}_n(k, m_1, \ldots, m_n) = (f(k, m_1), \ldots, f(k, m_n))$, where $p_{KM^n}(k, m_1, \ldots, m_n) = p_K(k)p_M(m_1)\ldots p_M(m_n)$ is the probability distribution on $K \times M^n \to O^n$. The variable $\mathcal{O}_n$ captures that k remains fixed over all invocations of f, while the messages $m_1, \ldots, m_n$ are chosen independently.

An unknown-message attacker making n side-channel observations $\mathcal{O}_n$ may learn information about the value of $\mathcal{K}$, i.e., about the secret key. This information can be expressed as the reduction in uncertainty about the value of $\mathcal{K}$, i.e., $I(\mathcal{K}; \mathcal{O}_n) = H(\mathcal{K}) - H(\mathcal{K}|\mathcal{O}_n)$. An alternative viewpoint is to use the attacker's remaining uncertainty about the key $H(\mathcal{K}|\mathcal{O}_n)$ as a measure for quantifying the system's resistance against an attack. Focussing on $H(\mathcal{K}|\mathcal{O}_n)$ has the advantage of a precise interpretation in terms of guessing: it is a lower bound on the average number of binary questions that the attacker still needs to ask to determine $\mathcal{K}$'s value [5].

Definition 1. We define $\Lambda(n) = H(\mathcal{K}|\mathcal{O}_n)$ as the *resistance against unknown-message attacks* of n steps.

The function Λ is monotonically decreasing, i.e., more observations can only reduce the attacker's uncertainty about the key. If $\Lambda(n) = H(\mathcal{K})$, the first n side-channel observations contain no information about the key. If $\Lambda(n) = 0$, the key is completely determined by n side-channel observations. Clearly, $\Lambda(0) = H(\mathcal{K})$.

Since $\Lambda(n)$ is defined as the expected value of $H(\mathcal{K}|\mathcal{O}_n = o)$ over all $o \in O^n$, it expresses whether keys are, on the average, hard to determine after n side-channel observations. It is straightforward to adapt the resistance to accommodate worst-case guarantees [17] or to use alternative notions of entropy that correspond to different kinds of brute-force guessing [5]. For example, $\min\{H(\mathcal{K}|\mathcal{O}_n = o) \mid o \in O^n\}$ captures the uncertainty about the key after the side-channel observation that contains the most information.

In Section 4, we will give an algorithm for computing the resistance $\Lambda(n)$ against unknown-message attacks. The time complexity of this algorithm is, however, exponential in n, rendering computation for large values of n infeasible. To remedy this problem, we will now establish an explicit formula for $\lim_{n\to\infty} \Lambda(n)$, which will allow us to compute limits for the resistance without being faced with the exponential increase in n.

3.2 Bounds for Unlimited Observations

The core idea for computing the limit of Λ can be described as follows: for a large number $o_1, \ldots, o_n$ of side-channel observations and a fixed key k, the relative frequency of each $o \in O$ converges to the probability $p_{\mathcal{O}|\mathcal{K}=k}(o)$. Thus, making an unbounded number of observations corresponds to learning the distribution $p_{\mathcal{O}|\mathcal{K}=k}$. We next give a formal account of this idea.[1]

Define $k_1 \equiv k_2$ if and only if $p_{\mathcal{O}|\mathcal{K}=k_1} = p_{\mathcal{O}|\mathcal{K}=k_2}$. Then $\equiv$ constitutes an equivalence relation on K, and $K/_{\equiv}$ denotes the set of equivalence classes. The random variable $\mathcal{V}\colon K \to K/_{\equiv}$ defined by $\mathcal{V}(k) = [k]_{\equiv}$ maps every key to its $\equiv$-equivalence class. Knowledge of the value of $\mathcal{V}$ hence corresponds to knowledge of the distribution $p_{\mathcal{O}|\mathcal{K}=k}$ associated with k. Intuitively, an unbounded number of observations contains as much information about the key as its $\equiv$-equivalence class. This is formalized by the following theorem.

Theorem 1. *Let $\mathcal{K}, \mathcal{V}$ and $\mathcal{O}_n$ be defined as above. Then*

$$\lim_{n\to\infty} H(\mathcal{K}|\mathcal{O}_n) = H(\mathcal{K}|\mathcal{V}) \tag{1}$$

For space constraints, we only sketch the proof of Theorem 1. The full proof can be found in Appendix A.

Proof sketch: For the proof of Theorem 1, one first shows that (1) is equivalent to $\lim_{n\to\infty} H(\mathcal{V}|\mathcal{O}_n) = 0$. One then expands

$$H(\mathcal{V}|\mathcal{O}_n) = \sum_{B\in K/_{\equiv}} p_{\mathcal{V}}(B) \sum_{o\in O^n} p_{\mathcal{O}_n|\mathcal{V}=B}(o) \log \frac{p_{\mathcal{O}_n}(o)}{p_{\mathcal{O}_n|\mathcal{V}=B}(o) p_{\mathcal{V}}(B)} \ , \tag{2}$$

and splits the inner sum into the observations $o = (o_1, \ldots, o_n)$ whose empirical distribution has Kullback-Leibler distance $\leq \epsilon$ from $p_{\mathcal{O}_n|\mathcal{V}=B}$, and those with distance $> \epsilon$. Here, ϵ is chosen such that every o is in the ϵ-neighborhood of $p_{\mathcal{O}_n|\mathcal{V}=B'}$ for at most one B'. The probability that an observation $o \in O^n$ is close to the underlying probability distribution converges to 1, with a rate that is exponential in n. With this, we show that the inner sum in (2) over the o with empirical distribution close to $p_{\mathcal{O}_n|\mathcal{V}=B}$ is bounded from above by $(1 - c_1 2^{-n\epsilon}) \log (1 + 2^{-n\epsilon} c_2)$, for constants c_1, c_2. We also show that the inner sum in (2) over the remaining $o \in O$ is bounded from above by $n(n+1)^{|O|} 2^{-n\epsilon} c_3$, for a constant c_3. Both partial sums converge to 0 as $n \to \infty$. The outer sum in (2) is finite and independent of n, hence (2) also converges to 0. □

[1] For probabilities, this is a consequence of the law of large numbers. We are not aware of a corresponding result for the conditional entropy.

4 Computing the Resistance against Unknown-message Attacks

In this section, we show how $\Lambda(n)$ and $\lim_{n\to\infty}\Lambda(n)$ can be computed for given implementations I_F of cryptographic functions F. For this, we first need a representation of the side-channel $f = f_{I_F}$; second, we need to compute Λ from this representation.

4.1 Estimating Time and Power Consumption

We focus on implementations in synchronous hardware as time and power consumption are easy to determine in this setting. We use the hardware design environment GEZEL [27] for describing circuits and for building up value table representations of f.

For timing analysis, $f(k, m)$ is the number of clock ticks consumed by the computation of $F(k, m)$ and can be determined by the simulation environment. Specifications in the GEZEL language can be mapped into a synthesizeable subset of VHDL, an industrial-strength hardware description language. The mapping preserves the circuit's timing behavior within the granularity of clock cycles. In this way, the guarantees obtained by formal analysis translate to silicon implementations.

For power analysis, we use the simple, technology-independent approximation provided by GEZEL: we define $f(k, m)$ as the number of bit transitions during the computation of $F(k, m)$. This number serves as an estimate for the circuit's power consumption and can be computed by the simulation environment. We next show how $\Lambda(n)$ can be computed from the value table representation of f.

4.2 Computing $\Lambda(n)$

For computing $\Lambda(n)$ we first show how $\Lambda(n) = H(\mathcal{K}|\mathcal{O}_n)$ can be decomposed into a sum of terms of the form $p_{\mathcal{O}|\mathcal{K}=k}(o)$, with $k \in K$ and $o \in O$. Subsequently, we sketch how this decomposition can be used to derive a simple implementation for computing $\Lambda(n)$.

We have the following equalities

$$H(\mathcal{K}|\mathcal{O}_n) = -\sum_{o\in O^n} p_{\mathcal{O}_n}(o) \sum_{k\in K} p_{\mathcal{K}|\mathcal{O}_n=o}(k)\, \log_2 p_{\mathcal{K}|\mathcal{O}_n=o}(k) \tag{3}$$

$$p_{\mathcal{K}|\mathcal{O}_n=o}(k) = \frac{p_{\mathcal{O}_n|\mathcal{K}=k}(o)p_{\mathcal{K}}(k)p_{\mathcal{O}_n}(o)}{p_{\mathcal{O}_n}(o)} \tag{4}$$

$$p_{\mathcal{O}_n}(o) = \sum_{k\in K} p_{\mathcal{O}_n|\mathcal{K}=k}(o)p_{\mathcal{K}}(k) \tag{5}$$

$$p_{\mathcal{O}_n|\mathcal{K}=k}(o_1, \ldots, o_n) = \prod_{i=1}^{n} p_{\mathcal{O}|\mathcal{K}=k}(o_i)\ , \tag{6}$$

where (4) is Bayes' formula and (6) holds because, for a fixed key, the observations are independent and identically distributed. Furthermore, for uniformly distributed messages, $p_{\mathcal{O}|\mathcal{K}=k}(o) = |\{m \mid f(k,m) = o)\}|/|M|$, which can be computed using the value table representation of f given by GEZEL.

The decomposition in (3)-(6) of $H(\mathcal{K}|\mathcal{O}_n)$ into a combination of terms of the form $p_{\mathcal{O}|\mathcal{K}=k}(o)$ and $p_{\mathcal{K}}(k)$ for $k \in K$ and $o \in O$ can be expressed by list comprehensions. This is illustrated by the following code snippet in Haskell [3]. Here, `pO` computes $p_{\mathcal{O}}(o)$ according to (5) and (6) from a list of observations `obs`, a list representation `keys` of K, and an array `p` that stores the values $p_{\mathcal{O}|\mathcal{K}=k}(o)$:

```
pO obs = sum [ product [ p!(o,k) | o <- obs ]| k <- keys ]
             / length keys
```

The computation of $\Lambda(n)$ can be encoded in a similarly concise way. We have implemented this in Haskell and use this implementation to perform experiments in Section 5.

4.3 Computing $\lim_{n\to\infty} \Lambda(n)$

From Theorem 1 it follows that $\lim_{n\to\infty} = H(\mathcal{K}|\mathcal{V})$, where $k_1 \equiv k_2$ if and only if $p_{\mathcal{O}|\mathcal{K}=k_1} = p_{\mathcal{O}|\mathcal{K}=k_2}$, and $\mathcal{V}(k) = [k]_{\equiv}$. From Proposition 2 of [17] it follows that $H(\mathcal{K}|\mathcal{V}) = \frac{1}{|K|}\sum_{B\in K/\equiv} |B| \log |B|$ for uniformly distributed keys. Hence for computing $H(\mathcal{K}|\mathcal{V})$ it suffices to determine the sizes of the $\equiv$-equivalence classes.

The equivalence classes of an equivalence relation form a partition of the relation's domain. We compute the partition of K corresponding to $\equiv$ by refinement. For this, consider the equivalence relations $\equiv_o$ defined by $k_1 \equiv_o k_2$ if and only if $p_{\mathcal{O}|\mathcal{K}=k_2}(o) = p_{\mathcal{O}|\mathcal{K}=k_2}(o)$. Clearly, $k_1 \equiv k_2$ if and only if $\forall o \in O. k_1 \equiv_o k_2$. For partitioning a set $B \subseteq K$ with respect to $\equiv_o$, group together all $k \in B$ with the same value of $p_{O|K=k}(o)$. For refining a given partition P with respect to $\equiv_o$, partition all $B \in P$ according to $\equiv_o$. For computing the partition corresponding to $\equiv$, successively refine the partition $\{K\}$ with respect to all $o \in O$. The following Haskell program implements this idea:

```
partKeys keys obs = foldr refineBy [keys] obs
  where refineBy o part = concat (map (splitBlockByObs o) part)
```

Here, the refinement of a block by an observation is accomplished by the function `splitBlockByObs`. The function `refineBy` applies this procedure to every block in a given partition. The function `partKeys` refines the partition `[keys]` by all observations in `obs`.

Finally, we can compute $H(\mathcal{K}|\mathcal{V}) = \frac{1}{|K|}\sum_{B\in K/\equiv} |B| \log |B|$ from the partition `part` returned by `partKeys`:

```
entropy part = sum [ b * logBase 2 b | x <- bs ] / sum bs
  where bs = map length part
```

We use this simple prototype implementation in our experiments below.

	$n=0$	$n=1$	$n=2$	$n=3$	$n \to \infty$
$\Lambda(n)$	8	7.801	7.605	7.41	0

Figure 1: Resistance of an AES SBox to unknown-message power attacks

5 Experimental Results

We now report on case studies where we analyze implementations of cryptographic algorithms with respect to their resistance against timing and power attacks.

We compute the resistance against unknown-message power attacks of an AES SBox with key addition. We also compute the resistance against unknown-message timing attacks of a circuit for exponentiation in finite fields, which is relevant, for example, in the generalized ElGamal encryption scheme [20]. Furthermore, we show how this result can be used for evaluating state-of-the-art countermeasures against timing attacks.

5.1 Power Analysis of an AES SBox

We have analyzed the power consumption of a GEZEL implementation of the AES SBox with key addition from [28] (without any countermeasures against power attacks). The circuit computes $F\colon K \times M \to D$ with $M = K = D = \{0,1\}^8$, where $F(k,m) = SBox(m) \oplus k$.

We assume that the circuit is in a known initial state before invocation of F. In this way, the number of bit transitions is a function of the inputs to the circuit, which is the side-channel f that we analyze.

Results of the Analysis The results of our analysis are depicted in Figure 1. They show that the attacker learns ≈ 0.2 bits of secret information in each of the first three observations and that, in the limit, the entire key information is leaked. We conclude that the circuit is vulnerable to unknown-message power attacks.

5.2 Timing Analysis of a Finite-Field Exponentiation Algorithm

We have analyzed a GEZEL implementation of the finite-field exponentiation algorithm from [12]. It takes two arguments m and x and computes m^x in $\mathbb{F}_{2^w}$. The exponentiation is performed by square-and-multiply, where each multiplication corresponds to a multiplication of polynomials. The entire algorithm consists of three nested loops.

Computing $\Lambda(n)$ with the implementation presented in Section 4 is expensive and does not scale to large values of n and operands of large bit-widths. To overcome this problem, we use the following approximation technique: we

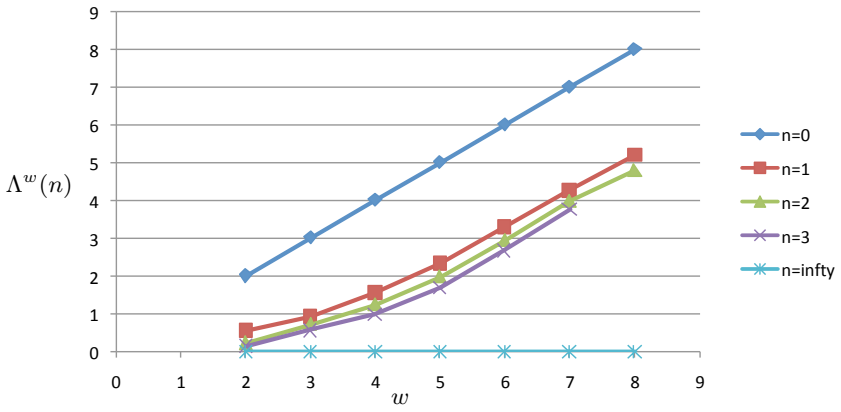

Figure 2: Resistance of a finite-field exponentiation algorithm to unknown-message timing attacks

parameterize each algorithm by the bit-width w of its operands. Our working assumption is that regularity in the values of Λ^w for $w \in \{2, \ldots, w_{\max}\}$ reflects the structural similarity of the algorithms. This permits the extrapolation to values of w beyond $w_{\max}$. To make this explicit, we will write Λ^w to denote that Λ is computed on w-bit operands.

Results of the Analysis The results of our analysis are given in Figure 2. The bit-width w of the operands is depicted along the horizontal axis and the entropy is depicted along the vertical axis. The different curves represent $\Lambda^w(n)$ for $n \in \{0, 1, 2, 3, \infty\}$.

We can draw the following conclusion from our data: the first timing observation reveals almost half of the secret information about the key. Subsequent observations reduce the uncertainty at a significantly slower rate. In the long run, however, the entire key information is leaked. Hence the circuit is vulnerable to unknown-message timing attacks.

5.3 Implications for the Security of Message-blinding

Timing attacks typically rely on the fact that the attacker can choose the input $m \in M$ and can measure the corresponding running time. Message-blinding, the state-of-the art countermeasure against timing attacks , renders this type of attack impractical by decoupling the algorithm's running time from m. Message-blinding has been proposed for exponentiation modulo n [15], but it can directly be applied to exponentiation in the field $\mathbb{F}_{2^w}$. We illustrate message-blinding for the common case of RSA.

Example 3. Consider an RSA decryption $x = m^k \mod n$, where m is chosen by the attacker, x the plaintext, n the modulus and k the secret key. Message-blinding decouples the running time of the exponentiation from m: In the *blind-*

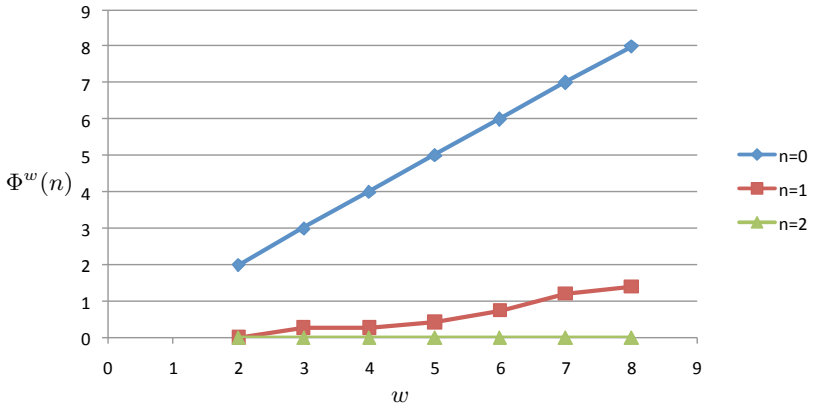

Figure 3: Resistance of a finite-field exponentiation algorithm to *chosen-message* attacks

ing phase one computes $m \cdot r^e \mod n$, where r is random and relatively prime to n, and e is the public key. The result of the decryption is $(m \cdot r^e)^k = x \cdot r \mod n$, which yields x after *unblinding*, i.e., after multiplication with $r^{-1} \mod n$.

The belief that message-blinding is secure is based on the assumption that the blinding and unblinding steps do not introduce new side-channels, and that $m \cdot r^e$ is sufficiently random. Analyzing the resistance of an exponentiation algorithm with respect to unknown-message attackers and uniformly distributed messages thus corresponds to analyzing the implementation with idealized message-blinding and with respect to chosen-message attacker.

This correspondence enables us to use Λ for evaluating the quality of message-blinding as a countermeasure for timing attacks against the finite-field exponentiation circuit from Section 5.2. Figure 3 is based on data from [17] and depicts the resistance of the same exponentiation algorithm with respect to *chosen-message* attacks. Here, $\Phi^w(n)$ denotes the remaining uncertainty after n steps of a chosen-message attack. The value $\Lambda^w(n) - \Phi^w(n)$, i.e., the difference between the curves in Figures 2 and 3, gives a formal account of what is gained by applying message-blinding as a countermeasure, namely that the information is leaked at a significantly slower rate. Figure 2 shows that $\lim_{n\to\infty} \Lambda(n) = 0$. This implies that, even with message-blinding applied, the timing side-channel eventually leaks the entire key information. To our knowledge, this is the first formal analysis of a countermeasure against timing attacks.

6 A Comparison of Side-Channel Attackers

In this section, we formally relate unknown-message, known-message and chosen-message attackers with respect to the information that they can extract from a given side-channel $f : K \times M \to O$. The main purpose of this compari-

son is a unified presentation that simplifies the task of picking the appropriate measure for a given attack scenario.

The result of the comparison is as expected: chosen-message attackers are stronger than known-message attackers, which are stronger than unknown-message attackers. All inclusions are shown to be strict. Before we formally state and prove this result, we begin with definitions of the resistance against known-message and chosen-message attacks.

Known-message attacks: We define the resistance against known-message attacks along the lines of Definition 1, where we express that the attacker knows the messages by conditioning the entropy if $\mathcal{K}$ on $\mathcal{M}_n$. Here, $\mathcal{M}_n$ models the n independent choices of messages from M.

Definition 2. We define $\Delta(n) = H(\mathcal{K}|\mathcal{O}_n\mathcal{M}_n)$ as the *resistance against known-message attacks* of n steps.

Note that Δ is an average-case measure, as $H(\mathcal{K}|\mathcal{O}_n\mathcal{M}_n)$ is the expected remaining uncertainty about $\mathcal{K}$ if the values of $\mathcal{O}_n$ and $\mathcal{M}_n$ are known. It can be adapted to accommodate worst-case guarantees by replacing the expected value by the minimal value over all n-tuples of messages or observations.

Chosen-message attacks: A measure for the resistance against chosen-message attacks has been defined in [17]. We next give a short account of this definition.

A chosen-message attack is formalized as a tree whose nodes are labeled with subsets of K. In this tree, an attack step is represented by a node v together with its children. The label A of v is the set of keys that could have led to the attacker's previous observations. The labels of the children of v form a partition of A. We require that this partition is of the form $\{A \cap f_m^{-1}(o) \mid o \in O\}$ for some $m \in M$, where $f_m(k) = f(k, m)$. This corresponds to the attacker's choice of a query m. By observing o, the attacker can narrow down the set of possible keys from A to $A'f^{-1}(o) \cap A$. The child of v with label A' is the starting point for subsequent attack steps.

Definition 3 ([17])**.** An *attack strategy against* f is a triple (T, r, L), where $T = (V, E)$ is a tree, $r \in V$ is the root, and $L\colon V \to 2^K$ is a node labeling with the following properties:

1. $L(r) = K$, and
2. for every $v \in V$, there is an $m \in M$ with $\{L(v) \cap f_m^{-1}(o) \mid o \in O\} = \{L(w) \mid (v, w) \in E\}$.

An attack strategy is of *length* l if T has height l.

A simple consequence of requirements 1 and 2 is that the labels of the leaves of an attack strategy $\mathfrak{a} = (T, r, L)$ form a partition $P_\mathfrak{a} = \{L(v) \mid v \text{ is a leaf of } T\}$ (the *induced partition*) of K. We denote by $\mathcal{V}_\mathfrak{a}$ the random variable that maps $k \in K$ to its enclosing block in $P_\mathfrak{a}$.

Definition 4 ([17])**.** We define $\Phi(n) = \min\{H(\mathcal{K}|\mathcal{V}_{\mathfrak{a}}) \mid \mathfrak{a} \text{ is of length } n\}$ as the *resistance against chosen-message attacks* of length n.

We are now ready to give a formal comparison of the three kinds of attackers.

A Hierarchy of Side-Channel Attacks

Theorem 2. *Let $f\colon K \times M \to O$ be a side-channel. Then, for all $n \in \mathbb{N}$,*

$$\Phi(n) \;\leq\; \Delta(n) \;\leq\; \Lambda(n)\ .$$

Proof. Conditioning on $\mathcal{M}_n$ does not increase the entropy, hence we have $\Delta(n) = H(\mathcal{K}|\mathcal{O}_n\mathcal{M}_n) \leq H(\mathcal{K}|\mathcal{O}_n) = \Lambda(n)$ for all $n \in \mathbb{N}$. For showing $\Phi(n) \leq \Delta(n)$ let $(m_1, \ldots, m_n) = argmin_{m \in M^n} H(\mathcal{K}|\mathcal{O}_n(\mathcal{M}_n = m))$ and observe that $H(\mathcal{K}|\mathcal{O}_n(\mathcal{M}_n = m)) \leq H(K|\mathcal{O}_n\mathcal{M}_n)$. Define $\mathfrak{a}$ as the attack strategy where, for each node of distance i from the root, the message m_i is chosen as a query. A simple calculation shows that $H(\mathcal{K}|\mathcal{V}_{\mathfrak{a}}) = \sum_{B \in P} p(B) H(\mathcal{K}|B) = H(\mathcal{K}|\mathcal{O}_n(\mathcal{M}_n = m))$ holds, where P is the partition of K given by $\bigcap_{i=1}^{n} \{f_{m_i}^{-1}(o) \mid o \in O\}$. Here, $\cap$ denotes the intersection of partitions, which is defined by $Q \cap Q' = \{B \cap B' \mid B \in Q, B' \in Q'\}$. Then $\Phi(n) \leq H(\mathcal{K}|\mathcal{V}_{\mathfrak{a}}) = H(\mathcal{K}|\mathcal{O}_n(\mathcal{M}_n = m)) \leq H(\mathcal{K}|\mathcal{O}_n\mathcal{M}_n) = \Delta(n)$, which concludes this proof. □

The inequalities in Theorem 2 are strict for some side-channels f, as the following example shows.

Example 4. Let $K = \{1, 2, 3, 4\}$, $M = \{m_1, m_2\}$, $O = \{1, 2\}$, and $f\colon K \times M \to O$ such that $f_{m_1}^{-1}(1) = \{1, 2\}$ and $f_{m_2}^{-1}(1) = \{2, 3\}$. With a uniform distribution on K, $\Phi(1) = 1$ and $\Phi(n) = 0$, for $n > 1$. According to Theorem 1, $\Lambda(n)$ is bounded from below by $H(\mathcal{K}|\mathcal{V})$. With a uniform distribution on M, we have $p_{\mathcal{O}|\mathcal{K}=1} = p_{\mathcal{O}|\mathcal{K}=3}$, hence $\Lambda(n) \geq H(\mathcal{K}|\mathcal{V}) = \frac{1}{2} H(\mathcal{K}|\mathcal{V} = [1]_{\equiv}) = \frac{1}{2}$. We have $\lim_{n\to\infty} \Delta(n) = 0$, but Δ will not reach its limit for a finite n as, e.g, $\mathcal{M}_n = (m_1, m_1, \ldots, m_1)$ is a possible choice of messages. Hence, $\Phi(n) < \Delta(n) < \Lambda(n)$ for the given f and large enough n.

We conclude that chosen-message attackers, known-message attackers, and unknown message attackers form a strict hierarchy in terms of the information that they can extract from a given side-channel.

7 Conclusions

We have presented a novel information measure to quantify the secret information that is revealed to unknown-message side-channel attackers. We have applied it to analyze hardware implementations with respect to their vulnerability against power and timing attacks. In particular, we have used it to perform the first formal analysis of message-blinding as a countermeasure against timing attacks. Finally, we have given a formal account of the intuition that more comprehensive attackers can extract more information from a given side-channel.

As future work, we plan to investigate whether techniques for entropy estimation [1, 2] can be used to approximate the value of Λ for implementations with operands of larger bit-widths. Furthermore, we plan to extend our work to capture Markov-chain models of side-channels. This will enable us to give bounds for the side-channel leakage in power attacks without the assumption that the system's initial state is known.

Acknowledgements We thank Patrick Schaumont for sharing his AES SBox implementation.

References

[1] G. Basharin. On a Statistical Estimate for the Entropy of a Sequence of Independent Random Variables. *Theory Probab. Appl.*, 47:333–336, 1959.

[2] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld. The complexity of approximating entropy. In *Proc. STOC '02*, pages 678–687. ACM, 2002.

[3] R. Bird. *Introduction to Functional Programming using Haskell*. Prentice Hall, second edition, 1998.

[4] D. Boneh and D. Brumley. Remote Timing Attacks are Practical. In *Proc. USENIX Security Symposium '03*.

[5] C. Cachin. Entropy Measures and Unconditional Security in Cryptography. PhD thesis, ETH Zürich, 1997.

[6] J. Cathalo, F. Koeune, and J.-J. Quisquater. A New Type of Timing Attack: Application to GPS. In *Proc. CARDIS '03*, LNCS 2779, pages 291–303. Springer.

[7] S. Chari, C. S. Jutla, J. R. Rao, and P. Rohatgi. Towards Sound Approaches to Counteract Power-Analysis Attacks. In *Proc. CRYPTO '99*, LNCS 1666, pages 398–412. Springer.

[8] S. Chari, J. R. Rao, and P. Rohatgi. Template Attacks. In *Proc. CHES '02*, LNCS 2523, pages 13–28. Springer.

[9] D. Clark, S. Hunt, and P. Malacaria. Quantitative Information Flow, Relations and Polymorphic Types. *J. Log. Comput.*, 18(2):181–199, 2005.

[10] M. Clarkson, A. Myers, and F. Schneider. Belief in Information Flow. In *Proc. CSFW '05*, pages 31– 45. IEEE.

[11] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, second edition, 2006.

[12] M. Davio, J. P. Deschamps, and A. Thayse. *Digital Systems with Algorithm Implementation*. John Wiley & Sons, Inc., 1983.

[13] J.-F. Dhem, F. Koeune, P.-A. Leroux, P. Mestre, J.-J. Quisquater, and J.-L. Willems. A Practical Implementation of the Timing Attack. In *Proc. CARDIS '98*, LNCS 1820, pages 167–182. Springer.

[14] K. Gandolfi, C. Mourtel, and F. Olivier. Electromagnetic analysis: Concrete results. In *Proc. CHES '01*, LNCS 2162, pages 251–261. Springer.

[15] P. Kocher. Timing Attacks on Implementations of Diffie-Hellman, RSA, DSS, and Other Systems. In *Proc. CRYPTO '96*, LNCS 1109, pages 104–113. Springer.

[16] P. Kocher, J. Jaffe, and B. Jun. Differential Power Analysis. In *Proc. CRYPTO '99*, LNCS 1666, pages 388–397. Springer.

[17] B. Köpf and D. Basin. An Information-Theoretic Model for Adaptive Side-Channel Attacks. In *Proc. CCS '07*, pages 286 – 296. ACM.

[18] G. Lowe. Quantifying Information Flow. In *Proc. CSFW '02*, pages 18–31. IEEE.

[19] F. Mace, F.-X. Standaert, and J.-J. Quisquater. An Informtion Theoretic Evaluation of Side-Channel Resistant Logic Styles. In *Proc. CHES '07*, LNCS 4727, pages 427–442. Springer.

[20] A. Menezes, P. van Oorschot, and S. Vanstone. *Handbook of Applied Cryptography*. CRC Press, 1996.

[21] T. S. Messerges, E. A. Dabbish, and R. H. Sloan. Power Analysis Attacks of Modular Exponentiation in Smartcards. In *Proc. CHES '99*, LNCS 1717, pages 144–157. Springer.

[22] S. Micali and L. Reyzin. Physically Observable Cryptography (Extended Abstract). In *Proc. TCC '04*, LNCS 2951, pages 278–296. Springer.

[23] F. N. Najm. A Survey of Power Estimation Techniques in VLSI Circuits. *IEEE Transactions on VLSI Systems*, 2(4):446–455, 1994.

[24] D. A. Osvik, A. Shamir, and E. Tromer. Cache Attacks and Countermeasures: the Case of AES. In *Proc. CT-RSA '06*, LNCS 3860, pages 1–20. Springer.

[25] C. Petit, F.-X. Standaert, O. Pereira, T. G. Malkin, and M. Yung. A Block Cipher based Pseudo Random Number Generator Secure Against Side-Channel Key Recovery. In *to appear in Proc. AsiaCCS '07*.

[26] J.-J. Quisquater and D. Samyde. ElectroMagnetic Analysis (EMA): Measures and Couter-Measures for Smard Cards. In *Proc. E-smart '01*, LNCS 2140, pages 200–210. Springer.

[27] P. Schaumont, D. Ching, and I. Verbauwhede. An Interactive Codesign Environment for Domain-Specific Coprocessors. *ACM Transactions on Design Automation for Electronic Systems*, 11(1):70–87, 2006.

[28] P. Schaumont and K. Tiri. Masking and Dual-Rail Logic Don't Add Up. In *Proc. CHES '07*, LNCS 4727, pages 95–106. Springer.

[29] F.-X. Standaert, E.Peeters, C. Archambeau, and J.-J. Quisquater. Towards Security Limits in Side-Channel Attacks. In *Proc. CHES '06*, LNCS 4249, pages 30–45. Springer.

[30] L. Zhong, S. Ravi, A. Raghunathan, and N. Jha. Power Estimation Techniques for Cycle-Accurate Functional Descriptions of Hardware. In *Proc. ICCAD '04*, pages 668– 675. ACM.

A Proof of Theorem 1

We first state two technical lemmas that we will need in the proof.

Lemma 1. *Let $k_1, k_2 \in K$ and $n \in \mathbb{N}$. Then*

1. $H(\mathcal{O}_n|\mathcal{K}) = H(\mathcal{O}_n|\mathcal{V})$,
2. $\lim_{n\to\infty} H(\mathcal{K}|\mathcal{O}_n) = H(\mathcal{K}|\mathcal{V})$ *iff* $\lim_{n\to\infty} H(\mathcal{V}|\mathcal{O}_n) = 0$.

Proof. For proving statement 1., observe that $p_{\mathcal{O}|\mathcal{K}=k} = p_{\mathcal{O}|\mathcal{V}=[k]}$ and, consequently, $H(\mathcal{O}_n|\mathcal{K}=k) = H(\mathcal{O}_n|\mathcal{V}=[k])$. Then

$$\begin{aligned} H(\mathcal{O}_n|\mathcal{K}) &= \sum_{k\in K} p_{\mathcal{K}}(k) H(\mathcal{O}_n|\mathcal{K}=k) = \sum_{B\in K/_\equiv} \sum_{k\in B} p_{\mathcal{K}}(k) H(\mathcal{O}_n|\mathcal{K}=k) \\ &= \sum_{B\in K} p_{\mathcal{V}}(B) H(\mathcal{O}_n|\mathcal{V}=B) = H(\mathcal{O}|\mathcal{V}) \ . \end{aligned}$$

For proving statement 2., observe that $H(\mathcal{K}\mathcal{V}) = H(\mathcal{V})$, as $\mathcal{V}$ is determined by $\mathcal{K}$. Hence $H(\mathcal{K}|\mathcal{O}_n) - H(\mathcal{K}|\mathcal{V}) = (H(\mathcal{O}_n|\mathcal{K}) + H(\mathcal{K}) - H(\mathcal{O}_n)) - (H(\mathcal{K}\mathcal{V}) - H(\mathcal{V})) = H(\mathcal{O}_n|\mathcal{K}) + H(\mathcal{V}) - H(\mathcal{O}_n)$, as $H(\mathcal{K}\mathcal{V}) = H(\mathcal{V})$. With the first part of Lemma 1, it follows that $H(\mathcal{O}_n|\mathcal{K}) + H(\mathcal{V}) = H(\mathcal{O}_n|\mathcal{V}) + H(\mathcal{V}) = H(\mathcal{O}_n\mathcal{V})$. As $H(\mathcal{O}_n\mathcal{V}) - H(\mathcal{O}_n) = H(\mathcal{V}|\mathcal{O}_n)$, Assertion (1) is equivalent to $\lim_{n\to\infty} H(\mathcal{V}|\mathcal{O}_n) = 0$ □

Let $\mathcal{X}_1, \ldots, \mathcal{X}_n$ be independent and identically distributed random variables with distribution $p_{\mathcal{X}}$. The *type* t_x of a sequence $x = (x_1, \ldots, x_n) \in X^n$ is the relative frequency of occurrences of each x_i. We define $U_\epsilon(p_{\mathcal{X}})$ as the set of sequences with types of Kullback-Leibler distance $\leq \epsilon$ from $p_{\mathcal{X}}$, i.e., $U_\epsilon(p_{\mathcal{X}}) = \{x \in X^n \mid D(t_x \parallel p_{\mathcal{X}}) \leq \epsilon\}$. Let $U_\epsilon^c(p_{\mathcal{X}})$ denote the set complement of $U_\epsilon(p_{\mathcal{X}})$, i.e., $U_\epsilon^c(p_{\mathcal{X}}) = X^n \setminus U_\epsilon(p_{\mathcal{X}_n})$. The following lemma from [11] shows that the probability of an observation (a set of observations, respectively) decreases exponentially with its Kullback-Leibler distance from the underlying distribution.

Lemma 2 ([11])**.** *Let $\mathcal{X}_1, \ldots, \mathcal{X}_n$ be independent and identically distributed random variables with distribution $p_{\mathcal{X}}$ and finite range X. Then*

1. $p_{\mathcal{X}_n}(x) = 2^{-n(H(t_x)+D(t_x \| p_{\mathcal{X}}))}$
2. $p_{\mathcal{X}_n}(U_\epsilon^c(p_{\mathcal{X}})) \leq (n+1)^{|X|} 2^{-n\epsilon}$

See [11] (pp. 349 and 356) for a proof of Lemma 2. We proceed with the proof of Theorem 1.

Proof of Theorem 1. According to Lemma 1, it suffices to show that $\lim_{n\to\infty} H(\mathcal{V}|\mathcal{O}_n) = 0$.

$$H(V|\mathcal{O}_n) = \sum_{o \in O^n} p_{\mathcal{O}_n}(o)\ H(\mathcal{V}|\mathcal{O}_n = o) \tag{7}$$

$$= \sum_{o \in O^n} p_{\mathcal{O}_n}(o) \sum_{B \in K/_\equiv} p_{\mathcal{V}|\mathcal{O}_n=o}(B) \log \left(p_{\mathcal{V}|\mathcal{O}_n=o}(B)\right)^{-1} \tag{8}$$

$$= \sum_{B \in K/_\equiv} p_{\mathcal{V}}(B) \sum_{o \in O^n} p_{\mathcal{O}_n|\mathcal{V}=B}(o) \log \frac{p_{\mathcal{O}_n}(o)}{p_{\mathcal{O}_n|\mathcal{V}=B}(o) p_{\mathcal{V}}(B)} \tag{9}$$

To simplify notation, we will from now on abbreviate $p_{\mathcal{O}_n|\mathcal{V}=B}$ by p_B. Let $\epsilon > 0$ such that $U_{2\epsilon}(p_B) \cap U_{2\epsilon}(p_{B'}) = \emptyset$ for all $B, B' \in K/_\equiv$ with $B \neq B'$. Such an ϵ exists, although the Kullback-Leibler distance is not a metric: the existence of such a ϵ follows from $D(p_B \parallel p_{B'}) \geq \frac{1}{2 \ln 2} \|p_B - p_{B'}\|_1^2$, where $\|\cdot\|_1$ is the $\mathcal{L}_1$-norm (see [11], p 370), and because $D(p_B \parallel p'_B) = 0$ only if $B = B'$. We next group the inner sum in (9) according to $O^n = U_\epsilon^c(p_B) \uplus U_\epsilon(p_B)$ and show that both parts converge to 0 as $n \to \infty$. As $|K/_\equiv|$ is finite and independent of n, the entire term in (9) also converges to 0 as $n \to \infty$. We perform a case split with respect to containment in $U_\epsilon(p_B)$

$\mathbf{o \in U_\epsilon(p_B)}$: From part 1 of Lemma 2 and the definition of ϵ it follows that

$$\frac{p_{B'}(o)}{p_B(o)} = \frac{2^{-n(H(t_o)+D(t_o \| p_{B'}))}}{2^{-n(H(t_o)+D(t_o \| p_B))}} \tag{10}$$

$$= 2^{-n(D(t_o \| p_{B'}) - D(t_o \| p_B))} \tag{11}$$

$$\leq 2^{-n\epsilon} \tag{12}$$

for $o \in U_\epsilon(p_B)$ and $B' \neq B$. Hence

$$0 \leq \sum_{o \in U_\epsilon(p_B)} p_B(o) \log \frac{p_{\mathcal{O}_n}(o)}{p_B(o) p_{\mathcal{V}}(B)} \tag{13}$$

$$\leq \sum_{o \in U_\epsilon(p_B)} p_B(o) \log \sum_{B' \in K/_\equiv} \frac{p_{B'}(o) p_{\mathcal{V}}(B')}{p_B(o) p_{\mathcal{V}}(B)} \tag{14}$$

$$\leq \sum_{o \in U_\epsilon(p_B)} p_B(o) \log \sum_{B' \in K/_\equiv} 2^{-n\epsilon} \frac{p_{\mathcal{V}}(B')}{p_{\mathcal{V}}(B)} \tag{15}$$

$$\leq (1 - (n+1)^{|O|} 2^{-n\epsilon}) \log 1 + 2^{-n\epsilon} \sum_{B' \neq B} \frac{p_{\mathcal{V}}(B')}{p_{\mathcal{V}}(B)} \ , \tag{16}$$

where the last two inequalities follow from (12) and Part 2 of Lemma 2, respectively. Finally, (16) and hence (13) converge to $1 \log 1 = 0$ as $n \to \infty$.

$\mathbf{o \in U_\epsilon^c(p_B)}$: Let $\delta = \min(\{p_B(o) | o \in O\} \setminus \{0\})$. Observe that, for $o \in O^n$, either $p_B(o) = 0$ or $p_B(o) \geq \delta^n$. W.l.o.g. assume $p_B(o) \neq 0$ in the following. Then

$$0 \leq \sum_{o \in U_\epsilon^c(p_B)} p_B(o) \log \frac{p_{\mathcal{O}_n}(o)}{p_B(o) p_{\mathcal{V}}(B)} \tag{17}$$

$$\leq \sum_{o \in U_\epsilon^c(p_B)} p_B(o) \log \frac{1}{\delta^n p_{\mathcal{V}}(B)} \tag{18}$$

$$\leq \sum_{o \in U_\epsilon^c(p_B)} p_B(o) (-n \log \delta)^{\frac{1}{p_{\mathcal{V}}(B)}} \tag{19}$$

$$\leq (n+1)^{|O|} 2^{-n\epsilon} (-n \log \delta)^{\frac{1}{p_{\mathcal{V}}(B)}} \tag{20}$$

where the last inequality follows from Lemma 2. Finally (20) and hence (17) converge to 0 as $n \to \infty$, which concludes our proof.

□