On the Optimality of Linear, Differential and Sequential Distinguishers

Pascal Junod

Security and Cryptography Laboratory
Swiss Federal Institute of Technology
CH-1015 Lausanne, Switzerland
pascal.junod@epfl.ch

This paper is the long version of
P.Junod, "On the optimality of linear, differential and sequential distinguishers", to appear in
Advances in Cryptology – EUROCRYPT'03
May 4-8, 2003, Warsaw - Poland
Lecture Notes in Computer Science, Springer-Verlag, 2003

Abstract. In this paper, we consider the statistical decision processes behind a linear and a differential cryptanalysis. By applying techniques and concepts of statistical hypothesis testing, we describe precisely the shape of optimal linear and differential distinguishers and we improve known results of Vaudenay concerning their asymptotic behaviour. Furthermore, we formalize the concept of "sequential distinguisher" and we illustrate potential applications of such tools in various statistical attacks.

Keywords: Distinguishers, Statistical Hypothesis Testing, Linear Cryptanalysis, Differential cryptanalysis

1 Introduction

Historically, statistical procedures are indissociable of cryptanalytic attacks against block ciphers. One of the first attack exploiting statistical correlations in the core of DES [24] is Davies and Murphy's attack [10]. Biham and Shamir's differential cryptanalysis [1–3], Matsui's attack against DES [17, 18], Vaudenay's statistical and χ^2 cryptanalysis [29], Harpes and Massey's partitioning cryptanalysis [14], and Gilbert-Minier stochastic cryptanalysis [21] are attacks using statistical procedures in their core.

To the best of our knowledge, Murphy *et al.*, in an unpublished report [22], proposed for the first time a general statistical framework for the analysis of block ciphers using the technique of *likelihood estimation*. Other examples can be found in the field of cryptology: recently, Coppersmith, Halevi and Jutla [7] have devised a general statistical framework for analysing stream ciphers; they use the concept of statistical hypothesis testing for systematically distinguishing

a stream cipher from a random function. Other examples (this list being non-exhaustive) include Maurer's analysis of Simmon's authentication theory [19, 20] and Cachin's theoretical treatment of steganography [4, 5].

In a parallel way, some attempts of formalizing resistance of block ciphers towards cryptanalytic attacks have been proposed: for instance, Pornin [25] proposes a general criterion of resistance against the Davies and Murphy attack; for this purpose, he makes use of statistical hypothesis testing. Vaudenay, in a sequence of papers (*e.g.* [30, 31, 28]) proposes the *decorrelation theory* as a generic technique for estimating the strength of block ciphers against various kinds of attacks. In these papers, he notably derives bounds on the best advantage of any linear and differential distinguishers, however without using statistical hypothesis testing concepts.

As pointed out by many authors, statistical hypothesis tests are convenient in the analysis of statistical problems, since, in certain cases, well-known optimality results (like the Neyman-Pearson lemma, for instance) can be applied.

1.1 Contributions of this Paper

In this paper, we consider the resistance of block ciphers against linear and differential cryptanalysis as a statistical hypothesis testing problem, which allows us to improve Vaudenay's asymptotic bounds on the best advantage of any linear and differential distinguishers and to give optimality results on the decision processes involved during these attacks.

For this, we recall some well-known statistical concepts in Section §2. In Section §3, we treat linear distinguishers and we derive a Chernoff-like bound, which gives the right asymptotic behaviour of the best advantage of such distinguishers. In §4, we do the same for differential distinguishers. In §5, we formalize the notion of *sequential distinguisher*; this kind of statistical procedure has been recognized quite early as potentially useful (in [22, 10], for instance). We restate this by showing, with help of a toy-example (a linear cryptanalysis of 5-rounds DES), that sequential sampling procedures may divide the needed number of plaintext-ciphertext pairs by a non-negligible factor in certain statistical cryptanalysis. In §6, we discuss potential applications of statistical hypothesis testing concepts in various attacks, and finally, we conclude in §7.

1.2 Notation

The following notation will be used throughout this paper. Random variables[1] $X, Y, \ldots$ are denoted by capital letters, while *realizations* $x \in \mathcal{X}, y \in \mathcal{Y}, \ldots$ of random variables are denoted by small letters; random vectors $\mathbf{X}, \mathbf{Y}, \ldots$ and their realizations $\mathbf{x}, \mathbf{y}, \ldots$ are denoted in bold characters. The fact for a random variable X to follow a distribution D is denoted $X \leftarrow \mathsf{D}$, while its probability function is denoted by $\Pr_X[x]$. Finally, as usual, "iid" means "independent and identically distributed".

[1] In this paper, we are only dealing with discrete random variables.

2 Statistical Hypothesis Testing

We recall some well-known facts about statistical hypothesis testing, both in the classical and in the Bayesian approaches; details can be found in any good book on statistics (*e.g.* see [26]).

2.1 Classical Approach

Let D_0 and D_1 be two different probability distributions defined on the same finite set $\mathcal{X}$. In a *binary hypothesis testing problem*, one is given an element $x \in \mathcal{X}$ which was drawn according either to D_0 or to D_1 and one has to decide which is the case. For this purpose, one defines a so-called *decision rule*, which is a function $\delta : \mathcal{X} \to \{0,1\}$ taking a sample of X as input and defining what should be the guess for each possible $x \in \mathcal{X}$. Associated to this decision rule are two different types of error probabilities: $\alpha \triangleq \Pr_{X_0}[\delta(x) = 1]$ and $\beta \triangleq \Pr_{X_1}[\delta(x) = 0]$. The decision rule δ defines a partition of $\mathcal{X}$ in two subsets which we denote by $\mathcal{A}$ and $\overline{\mathcal{A}}$, *i.e.* $\mathcal{A} \cup \overline{\mathcal{A}} = \mathcal{X}$; $\mathcal{A}$ is called the *acceptance region of* δ. We recall now the Neyman-Pearson lemma which derives the shape of the optimal statistical test δ between two simple hypotheses, *i.e.* which gives the optimal decision region $\mathcal{A}$ (in terms of error probability).

Lemma 1 (Neyman-Pearson). *Let X be a random variable drawn according to a probability distribution D and let be the decision problem corresponding to hypotheses $X \leftarrow \mathsf{D}_0$ and $X \leftarrow \mathsf{D}_1$. For $\tau \geq 0$, let $\mathcal{A}$ be defined by*

$$\mathcal{A} \triangleq \left\{ x \in \mathcal{X} : \frac{\Pr_{X_0}[x]}{\Pr_{X_1}[x]} \geq \tau \right\} \tag{1}$$

Let $\alpha^* \triangleq \Pr_{X_0}[\overline{\mathcal{A}}]$ and $\beta^* \triangleq \Pr_{X_1}[\mathcal{A}]$. Let $\mathcal{B}$ be any other decision region with associated probabilities of error α and β. If $\alpha \leq \alpha^*$, then $\beta \geq \beta^*$.

Hence, the Neyman-Pearson lemma indicates that the optimum test (regarding error probabilities) in case of a binary decision problem is the *likelihood-ratio test*. All these considerations are summarized in Definition 1.

Definition 1 (Optimal Binary Hypothesis Test). *To test $X \leftarrow \mathsf{D}_0$ against $X \leftarrow \mathsf{D}_1$, choose a constant $\tau > 0$ depending on α and β and define the likelihood ratio*

$$\mathrm{lr}(x) \triangleq \frac{\Pr_{X_0}[x]}{\Pr_{X_1}[x]} \tag{2}$$

The optimal decision function is then defined by

$$\delta_{\mathrm{opt}} \triangleq \begin{cases} 0 \ (i.e \text{ accept } X \leftarrow \mathsf{D}_0) & \text{if } \mathrm{lr}(x) \geq \tau \\ 1 \ (i.e. \text{ accept } X \leftarrow \mathsf{D}_1) & \text{if } \mathrm{lr}(x) < \tau \end{cases} \tag{3}$$

We note that Lemma 1 does not consider any special hypothesis on the observed random variable X. In the following, we will assume that we are interested in taking a decision about the distribution of a random vector $\mathbf{X} \triangleq (X_1, \ldots, X_n)$

where $X_1, \ldots, X_n$ are iid random variables, *i.e.* $\mathbf{X} \leftarrow \mathsf{D}^n$ is a random vector of n independent samples of the random variable X. This is a typical situation during a known-plaintext attack.

When dealing with error probabilities, one usually proceeds as follows in the classical approach: one of the two possible error probabilities is fixed, and one minimizes the other error probability. In this case, Stein's lemma (we refer to [8] for more details) gives the best error probability expression. As this approach lacks symmetry, we won't describe it in more details.

2.2 Bayesian Approach

The other possibility is to follow a Bayesian approach and to assign *prior* probabilities π_0 and π_1 to both hypotheses, respectively, and *costs* $\kappa_{i,j} \geq 0$ to the possible decisions $i \in \{0,1\}$ and states of nature $j \in \{0,1\}$. In this case, we would like to minimize the *expected cost*. If we assign $\kappa_{0,0} = \kappa_{1,1} \triangleq 0$ and $\kappa_{0,1} = \kappa_{1,0} \triangleq 1$, *i.e.* correct decisions are not penalized, while incorrect decisions are penalized equally, then the optimal Bayesian decision rule is given by

$$\delta(\mathbf{x}) \triangleq \begin{cases} 0 & \text{if } \pi_0 \Pr_{X_0^n}[\mathbf{x}] \geq \pi_1 \Pr_{X_1^n}[\mathbf{x}] \\ 1 & \text{if } \pi_0 \Pr_{X_0^n}[\mathbf{x}] < \pi_1 \Pr_{X_1^n}[\mathbf{x}] \end{cases} \tag{4}$$

Clearly, the overall error probability $P_e^{(n)} \triangleq \pi_0 \alpha^{(n)} + \pi_1 \beta^{(n)}$ of such an optimal Bayesian distinguisher must decrease towards zero as the number n of samples increases. It turns out that the decrease asymptotically approaches an exponential in the number of samples drawn before the decision, the exponent being given by the so-called *Chernoff bound* (see Theorem 1; in Appendix A, we give some information-theoretic results justifying this bound, and we refer to [8] for a detailed and complete treatment).

Theorem 1 (Chernoff). *The best probability of error of the Bayesian decision rule defined in* (4) *satisfies*

$$\lim_{n \to +\infty} \frac{1}{n} \log \frac{P_e^{(n)}}{2^{-n\nu}} = 0 \tag{5}$$

where $\nu = C(\mathsf{D}_0, \mathsf{D}_1)$ *is the* Chernoff information *between* D_0 *and* D_1 *defined by*

$$C(\mathsf{D}_0, \mathsf{D}_1) \triangleq -\min_{0 \leq \lambda \leq 1} \log \left(\sum_{x \in \mathcal{X}} \Pr_{X_0}[x]^{\lambda} \Pr_{X_1}[x]^{1-\lambda} \right) \tag{6}$$

Note that the Bayesian error exponent does not depend on the actual value of π_0 and π_1, as long as they are non-zero: essentially, the effect of the prior is washed out for large sample sizes.

3 Linear Distinguishers

In this section, we consider the classical model of a linear distinguisher and we present several new results derived using tools of statistical hypothesis testing.

3.1 Introduction

A *linear distinguisher* δ_{lin} is a (possibly computationally unbounded) Turing machine which can play with an oracle Ω implementing a permutation C; δ_{lin} is bounded in the number n of queries to the oracle Ω. Furthermore, it uses a *linear characteristic* $(\mathbf{a}, \mathbf{b})$ which is a pair of boolean vectors. Algorithm 1 defines the classical modelization of a linear distinguisher (see [30, 31, 28]).

Parameters: a complexity n, a characteristic $(\mathbf{a}, \mathbf{b})$, an acceptance region $\mathcal{A}^{(n)}$
Input: an oracle Ω which implements a permutation C
Initialize a counter u to 0.
for $i = 1 \dots n$ **do**
  Pick uniformly at random x and query $C(x)$ to the oracle Ω.
  if $\mathbf{a} \cdot x = \mathbf{b} \cdot C(x)$ **then**
    Increment u
  end if
end for
if $u \in \mathcal{A}^{(n)}$ **then**
  Output 0
else
  Output 1
end if

Algorithm 1: Modelization of a linear distinguisher δ_{lin}.

The statistical game is the following. One gives an oracle Ω to Algorithm 1, which is with probability $\pi_0 = \frac{1}{2}$ the permutation C or, with probability $\pi_1 = \frac{1}{2}$, a permutation $C^* \in_U \mathcal{C}_m$ drawn uniformly at random from the set $\mathcal{C}_m$ of all permutations over inputs of size m (C^* is often refereed as the "Perfect Cipher"). The goal of Algorithm 1 is to decide whether Ω implements C or C^*. One measures the performance of a distinguisher δ_{lin} by the expression

$$\text{Adv}^n_{\delta_{\text{lin}}}(C, C^*) \triangleq \left| \Pr_{C}[\delta_{\text{lin}}(\mathbf{x}) = 1] - \Pr_{C^*}[\delta_{\text{lin}}(\mathbf{x}) = 1] \right| = \left| 2P_e^{(n)} - 1 \right| \qquad (7)$$

where $\mathbf{x} = (x_1, \dots, x_n)$ is the vector of the values queried to the oracle. The distinguisher's core is the *acceptance region* $\mathcal{A}^{(n)}$: it defines the set of values $(x_1, \dots, x_n)$ which lead to output 0 (*i.e.* it decides that the oracle implements C) or 1 (*i.e.* it decides that the oracle implements C^*).

As pointed out by Chabaud and Vaudenay in [6], linear cryptanalysis is based on the quantity

$$\text{LP}^C(\mathbf{a}, \mathbf{b}) \triangleq \left(2 \cdot \Pr_X[\mathbf{a} \cdot X = \mathbf{b} \cdot C(X)] - 1 \right)^2 \qquad (8)$$

This value depends of the (fixed) permutation C and of the distribution of plaintext, which is usually defined to be uniform. Actually, most of the time, a cryptanalyst does not possess any information about the permutation (*i.e.* about the

key), so one is more interested in the *average* $\mathrm{LP}^C(\mathbf{a},\mathbf{b})$ over the permutation space $\mathcal{C}_m$ (or, equivalently, over the key space $\mathcal{K}$); this quantity is denoted

$$\mathrm{ELP}(\mathbf{a},\mathbf{b}) \triangleq \mathrm{E}\left[\mathrm{LP}^C(\mathbf{a},\mathbf{b})\right] \tag{9}$$

where the expectation is taken over the permutation distribution.

When studying linear distinguishers, one is interested in bounding the advantage of any linear distinguisher in terms of $\mathrm{ELP}(\mathbf{a},\mathbf{b})$. We review now a known result of Vaudenay (see [28], for instance).

Theorem 2 (Vaudenay). *For any distinguisher in the model of Algorithm 1*

$$\mathrm{BestAdv}^n_{\delta_{\mathrm{lin}}}(C,C^*) \leq 2.78\sqrt[3]{n\cdot\mathrm{ELP}(\mathbf{a},\mathbf{b})} + 2.78\sqrt[3]{\frac{n}{2^m-1}} \tag{10}$$

where m is the block size of the permutation.

In the case of a practical linear cryptanalysis of DES [18], we have $\mathrm{ELP}(\mathbf{a},\mathbf{b}) \approx 4\cdot\left(1.19\cdot 2^{-21}\right)^2 \approx 1.288\cdot 10^{-12}$ and $m=64$, which means that (10) is useful as long as $n \leq 2^{35}$. Thus, although of great theoretical interest, we note that (10) is not tight for large n, or, in other words, does not capture the asymptotical behavior of the advantage. In the next part, we reconsider this problem in the statistical hypothesis testing framework and we derive an asymptotically tight Chernoff-like bound on the best advantage of any linear distinguisher.

3.2 New Asymptotic Bounds

First, we note that if δ_{lin} is optimal, then $P_e^{(n)} \leq \frac{1}{2}$ for all $n>0$ (otherwise, we could modify it such that it outputs the opposite decision as defined in Algorithm 1 and get a smaller error probability). Thus, we have

$$\mathrm{Adv}^n_{\delta_{\mathrm{lin}}}(C,C^*) = 1 - 2P_e^{(n)} \tag{11}$$

As outlined before, the crucial part of δ_{lin} is the acceptance region $\mathcal{A}^{(n)}$. The following lemma, which is a direct application of Lemma 1, gives the optimal $\mathcal{A}^{(n)}_{\mathrm{opt}}$, *i.e.* the region producing the smallest overall error probability. Without loss of generality, we assume that

$$\mathrm{E}\left[\Pr_X[\mathbf{a}\cdot X = \mathbf{b}\cdot C(X)]\right] \triangleq \frac{1}{2}+\epsilon \quad \text{with } \epsilon>0 \tag{12}$$

where the expectation is taken over the key space $\mathcal{K}$ in case of a uniformly distributed plaintext space $\mathcal{X}$.

Lemma 2. *The optimal acceptance region for δ_{lin} is*

$$\mathcal{A}^{(n)}_{\mathrm{opt}} = \left\{u\in\{0,\ldots,n\} : u \geq n\cdot\frac{\log_2(1-2\epsilon)}{\log_2(1-2\epsilon)-\log_2(1+2\epsilon)}\right\} \tag{13}$$

where u is defined in Algorithm 1.

Proof. Following Lemma 1, the optimal decision region is given by (4) where $\pi_0 = \pi_1 = \frac{1}{2}$. In other words, δ_{lin} must decide that Ω implements C if

$$\left(\frac{1}{2}+\epsilon\right)^u \left(\frac{1}{2}-\epsilon\right)^{n-u} \geq \frac{1}{2^n} \tag{14}$$

which is equivalent to

$$u \cdot \log_2\left(\frac{1+2\epsilon}{1-2\epsilon}\right) + n \cdot \log_2(1-2\epsilon) \geq 0 \tag{15}$$

The lemma follows if we take into account that $\epsilon > 0$. □

Note that, for ϵ small, one can approximate (13) with

$$\mathcal{A}^{(n)}_{\text{opt}} \approx \left\{ u \in \{0,\ldots,n\} : u \geq n \cdot \left(\frac{1}{2}+\frac{\epsilon}{2}\right) \right\} \tag{16}$$

Using a precise version of Chernoff's theorem 1, we can bound the advantage of the best linear distinguisher as follows:

Theorem 3. *Let m be the block size of the involved permutations. For any distinguisher in the model of Algorithm 1*

$$1 - \frac{(n+1)}{2^{n\nu-1}} \leq \text{BestAdv}^n_{\delta_{\text{lin}}}(C, C^*) \leq 1 - \frac{1}{(n+1)\cdot 2^{n\nu-1}} \tag{17}$$

where $\nu = C(\mathsf{D}_0, \mathsf{D}_1)$ is the Chernoff information between D_0, a binary distribution having a bias equal to $\max\{\frac{1}{2^m-1}, \epsilon\}$ such that $\text{ELP}^C(\mathbf{a},\mathbf{b}) = 4\epsilon^2$ and the uniform binary distribution D_1.

Proof. From this point, $\mathcal{X} = \mathcal{K} \triangleq \{0,1\}$. We modelize by binary random variables U and V the events whether the counter u (see lines 6-7 of Algorithm 1) is incremented or not after an oracle response when the oracle implements C and C^*, respectively:

$$\begin{aligned} U &\triangleq \begin{cases} 0 & \text{if } \mathbf{a}\cdot x \neq \mathbf{b}\cdot C(x) \\ 1 & \text{if } \mathbf{a}\cdot x = \mathbf{b}\cdot C(x) \end{cases} \\ V &\triangleq \begin{cases} 0 & \text{if } \mathbf{a}\cdot x \neq \mathbf{b}\cdot C^*(x) \\ 1 & \text{if } \mathbf{a}\cdot x = \mathbf{b}\cdot C^*(x) \end{cases} \end{aligned} \tag{18}$$

We have to distinguish between two cases: if the oracle implements C^*, then V depends on two random values: the plaintext $x \in \mathcal{X}^m$ and the permutation, which is a permutation drawn uniformly at random in $\mathcal{C}_m$ and is parametered by a key $k \in \mathcal{K}^\ell$, where ℓ is the key length; in the second case, if the oracle implements C, the key k is fixed and U depends on the plaintext x only. More precisely,

$$U : \begin{cases} \Pr[U=0] \triangleq \frac{1}{2} - \epsilon \\ \Pr[U=1] \triangleq \frac{1}{2} + \epsilon \end{cases} \quad \text{with } \epsilon \neq 0 \tag{19}$$

and

$$V : \begin{cases} \Pr[V=0] \triangleq \frac{1}{2} - \epsilon(K) \\ \Pr[V=1] \triangleq \frac{1}{2} + \epsilon(K) \end{cases} \tag{20}$$

Here, the ϵ-value of (19) is related to $\mathrm{LP}^C(\mathbf{a}, \mathbf{b})$ as

$$\mathrm{LP}^C(\mathbf{a}, \mathbf{b}) = 4\epsilon^2 \tag{21}$$

for uniformly distributed plaintexts. We note that we lose information about ϵ, *i.e.* its sign. However, this does not play a role during the derivation of the Chernoff information between the two distributions of interest.

As usually, we will assume that the keys are modelized by a uniformly distributed random variable on $\mathcal{K}$ and that they are statistically independent of the plaintext. We have

$$\mathrm{E}[\epsilon(K)] = 0 \tag{22}$$

where the expectation is taken over the key space. To summarize, we have to distinguish between a uniformly distributed binary random variable (when Ω implements C^*) and a biased binary random variable (when Ω implements C).

In order to show the bounds given in Theorem 3, we use a more precise version of Sanov's Theorem tailored to binary random variables. Let $\mathcal{A}_{\mathrm{opt}}^{(n)}$ be the optimal acceptance region for δ_{lin} defined in Lemma 2. Let $\mathcal{E}_{\alpha^{(n)}} \in \mathcal{P}_n$ be the set of types (see Appendix A for more information about the method of types) such that

$$\mathcal{E}_{\alpha^{(n)}} \triangleq \left\{ \mathbf{x} \in \mathcal{P}_n : \mathsf{D}_{\mathbf{x}} \notin \mathcal{A}_{\mathrm{opt}}^{(n)} \right\} \tag{23}$$

when $\mathbf{x} \leftarrow \mathsf{D}_{X_0^n}$. Similarly,

$$\mathcal{E}_{\beta^{(n)}} \triangleq \left\{ \mathbf{x} \in \mathcal{P}_n : \mathsf{D}_{\mathbf{x}} \in \mathcal{A}_{\mathrm{opt}}^{(n)} \right\} \tag{24}$$

when $\mathbf{x} \leftarrow \mathsf{D}_{X_1^n}$. Then,

$$\begin{aligned}
\Pr_{X_0^n}[\mathcal{E}_{\alpha^{(n)}}] &= \sum_{\mathsf{D}_X \in \mathcal{E}_{\alpha^{(n)}} \cap \mathcal{P}_n} \Pr_{X_0^n}[\mathcal{T}(\mathsf{D}_X)] && (25)\\
&\leq \sum_{\mathsf{D}_X \in \mathcal{E}_{\alpha^{(n)}} \cap \mathcal{P}_n} 2^{-nD(\mathsf{D}_X||X_0)} && (26)\\
&\leq \sum_{\mathsf{D}_X \in \mathcal{E}_{\alpha^{(n)}} \cap \mathcal{P}_n} \max_{\mathsf{D}_X \in \mathcal{E}_{\alpha^{(n)}} \cap \mathcal{P}_n} 2^{-nD(\mathsf{D}_X||X_0)} && (27)\\
&= \sum_{\mathsf{D}_X \in \mathcal{E}_{\alpha^{(n)}} \cap \mathcal{P}_n} 2^{-n \min_{\mathsf{D}_X \in \mathcal{E}_{\alpha^{(n)}} \cap \mathcal{P}_n} D(\mathsf{D}_X||X_0)} && (28)\\
&\leq \sum_{\mathsf{D}_X \in \mathcal{E}_{\alpha^{(n)}} \cap \mathcal{P}_n} 2^{-n \min_{\mathsf{D}_X \in \mathcal{E}_{\alpha^{(n)}}} D(\mathsf{D}_X||X_0)} && (29)\\
&= \sum_{\mathsf{D}_X \in \mathcal{E}_{\alpha^{(n)}} \cap \mathcal{P}_n} 2^{-nD(\mathsf{D}_{X^*}||X_0)} && (30)\\
&\leq (n+1) \cdot 2^{-nD(\mathsf{D}_{X^*}||X_0)} && (31)
\end{aligned}$$

where the last inequality comes from

$$|\mathcal{P}_n| = \binom{n + |\mathcal{X}| - 1}{|\mathcal{X}| - 1} \tag{32}$$

The computation for upper bounding $\Pr_{X_1^n}[\mathcal{E}_{\beta^{(n)}}]$ are similar.

For the lower bound, we need a set $\mathcal{E}_{\alpha^{(n)}}$ such that for all large n, we can find a distribution in $\mathcal{E}_{\alpha^{(n)}} \cap \mathcal{P}_n$ which is close to D_{X^*}. As $\mathcal{E}_{\alpha^{(n)}}$ is the closure of its interior (thus the interior must be non-empty), then since $\bigcup_n \mathcal{P}_n$ is dense in the set of all distributions, it follows that $\mathcal{E}_{\alpha^{(n)}} \cap \mathcal{P}_n$ is non-empty for all $n \geq n_0$ for some n_0. We can then find a sequence of distributions D_{X_n} such that $\mathsf{D}_{X_n} \in \mathcal{E}_{\alpha^{(n)}} \cap \mathcal{P}_n$ and $D(\mathsf{D}_{X_n}||\mathsf{D}_{X_0^n}) \to D(\mathsf{D}_{X^*}||_{X_0^n})$. For each $n \geq n_0$,

$$\Pr_{X_0^n}[\mathcal{E}_{\alpha^{(n)}}] = \sum_{\mathsf{D}_X \in \mathcal{E}_{\alpha^{(n)}} \cap \mathcal{P}_n} \Pr_{X_0^n}[\mathcal{T}(\mathsf{D}_X)] \tag{33}$$

$$\geq \Pr_{X_0^n}[\mathcal{T}(\mathsf{D}_X)] \tag{34}$$

$$\geq \frac{2^{-nD(\mathsf{D}_{X_n}||\mathsf{D}_{X_0})}}{n+1} \tag{35}$$

Consequently,

$$\liminf \frac{1}{n} \Pr_{X_0^n}[\mathcal{E}_{\alpha^{(n)}}] \geq \liminf \left(-\frac{\log(n+1)}{n} - D(\mathsf{D}_{X_n}||X_0) \right) \tag{36}$$

$$= -D(\mathsf{D}_{X^*}||\mathsf{D}_{X_0}) \tag{37}$$

The computations are similar for lower bounding $\Pr_{X_1^n}[\mathcal{E}_{\beta^{(n)}}]$ Combining the upper bounds derived before and this lower bound, and the computations of Appendix A.3 yields Theorem 3. □

Generally, the Chernoff information cannot be expressed explicitely, because one has to solve a transcendental equation. However, in the case which interests us, $\nu(\lambda) \triangleq 2^{-\lambda} \cdot ((\frac{1}{2}+\epsilon)^\lambda + (\frac{1}{2}-\epsilon)^\lambda)$ and

$$C(\mathsf{D}_0, \mathsf{D}_1) = \nu(\lambda^*) \text{ for } \lambda^* = \frac{\log\left(-\frac{\log(1-2\epsilon)}{\log(1+2\epsilon)}\right)}{\log\left(\frac{1+2\epsilon}{1-2\epsilon}\right)} \tag{38}$$

We give now a numerical illustration: for $\epsilon = 1.19 \cdot 2^{-21}$ (which is the bias of the best linear approximation of 14 rounds of DES), we obtain a useful lower bound only for $n \geq 2^{48.2}$; unfortunately, even if it captures the asymptotic exponential shape of the best advantage curve, it is not practically useful for "interesting" values of n; for which concerns the upper bound, it is useful for all n but it is not tight: one may give a tighter lower bound using Bernstein's inequality (see Theorem 4 and [12] for a proof). In the following, we will assume that ϵ is small and thus that one is using (16) as acceptance region.

Theorem 4 (Bernstein's Inequality). *Let X_i be iid discrete random variables following a Bernoulli law with parameter $0 \leq p \leq 1$ and let $S_n \triangleq \sum_i X_i$. Then*

$$\Pr\left[S_n \geq n(p+\epsilon)\right] \leq e^{-\frac{1}{4}n\epsilon^2} \text{ for } \epsilon > 0 \tag{39}$$

This allows to derive in an easy way the following lower bound:

Theorem 5. *Let m be the block size of the involved permutations. For any distinguisher in the model of Algorithm 1*

$$\mathrm{BestAdv}^n_{\delta_{\mathrm{lin}}}(C, C^*) \geq 1 - e^{-\frac{n\overline{\epsilon}^2}{16}} \tag{40}$$

where $\overline{\epsilon} \triangleq \max\{\frac{1}{2^m-1}, \epsilon\}$ such that $\mathrm{ELP}^C(\mathbf{a}, \mathbf{b}) = 4\epsilon^2$.

4 Differential Distinguishers

Similarly, one can study *differential distinguishers* with the same tools. A differential distinguisher δ_{diff} is a (possibly computationally unbounded) Turing machine which is able to submit chosen *pairs* of plaintexts to an oracle Ω, implementing with probability $\pi_0 = \frac{1}{2}$ a fixed permutation C or, with probability $\pi_1 = \frac{1}{2}$, a permutation drawn uniformly at random from the set $\mathcal{C}_m$ of all permutations on m-bit blocks. Although the cryptanalytic settings are quite different (δ_{diff} can submit *chosen pairs* of plaintext), in a statistical point of view, the distinguishing process is very similar to linear distinguishers. In Algorithm 2, the classical modelization of a differential distinguisher [30, 31] is given.

If we look at Algorithm 2, we note that, although the complexity n is given in advance as input and is (implicitly) *fixed*, the effective number of queries to the oracle is merely a random variable. In other words, δ_{diff} does not make use of all the information that it could exploit. In fact, we can see the class of distinguishers submitting a *random* number of queries to the oracle as a generalization of the class of distinguishers submitting a *fixed* number of queries. We will call this generalization *sequential distinguishers*; this new concept is formalized and studied in Section 5.

1: **Parameters**: a complexity n, a characteristic (a, b)
2: **Input**: an oracle Ω which implements a permutation C
3: **for** $i = 1 \ldots n$ **do**
4: Pick uniformly at random x and query $C(x)$ and $C(x+a)$ to the oracle Ω.
5: **if** $C(x+a) = C(x) + b$ **then**
6: Output 0 and stop.
7: **end if**
8: **end for**
9: Output 1.

Algorithm 2: Classical modelization of a differential distinguisher δ_{diff}.

In order to better understand the statistical decision processes, we give in Algorithm 3 an "unorthodox" modelization, denoted δ'_{diff}, which is very similar to the linear one. As for linear distinguishers, it is well-known [23] that differential

1: **Parameters**: a complexity n, a characteristic (a, b), an acceptance region $\mathcal{A}^{(n)}$
2: **Input**: an oracle Ω which implements a permutation C
3: Initialize a counter u to 0.
4: **for** $i = 1 \dots n$ **do**
5: Pick uniformly at random x and query $C(x)$ and $C(x + a)$ to the oracle Ω.
6: **if** $C(x + a) = C(x) + b$ **then**
7: Increment u
8: **end if**
9: **end for**
10: **if** $u \in \mathcal{A}^{(n)}$ **then**
11: Output 0
12: **else**
13: Output 1
14: **end if**

Algorithm 3: Unorthodox modelization of a differential distinguisher δ'_{diff}.

cryptanalysis depends on the quantity $\mathrm{DP}^C(a, b) \triangleq \Pr_X[C(X + a) = C(X) + b]$, where the plaintext space $\mathcal{X}$ is uniformly distributed. As this value depends on the choice of the cipher (*i.e.* on the key), one defines $\mathrm{EDP}(a, b) \triangleq \mathrm{E}\left[\mathrm{DP}^C(a, b)\right]$, where the expectation is taken over the permutation space. We note that Algorithm 2 outputs 1 if and only if no *differential event* occurs. As for linear distinguishers, and considering this time Algorithm 3, one can define the optimal acceptance region using Lemma 1 and which is given by Lemma 3. As $\mathrm{EDP}^{C^*}(a, b) = \frac{1}{2^m - 1}$ (where m is the block size of the permutation), and, typically, $\mathrm{DP}^C(a, b) \triangleq \frac{1+\epsilon}{2^m - 1}$ with $0 < \epsilon \leq 2^m - 2$, we can note that the *optimal* acceptance region will make δ_{diff} output 0 if

$$\binom{n}{u} \left(\frac{1+\epsilon}{2^m - 1}\right)^u \left(1 - \frac{1+\epsilon}{2^m - 1}\right)^{n-u} \geq \binom{n}{u} \left(\frac{1}{2^m - 1}\right)^u \left(\frac{2^m - 2}{2^m - 1}\right)^{n-u}$$

which gives the following result.

Lemma 3. *The optimal acceptance region for* δ'_{diff} *is*

$$\mathcal{A}^{(n)}_{\text{opt}} = \left\{ u \in \{0, \dots, n\} : u \geq n \cdot \frac{\log(2^m - 2) - \log(2^m - 2 - \epsilon)}{\log((2^m - 2)(1 + \epsilon)) - \log(2^m - 2 - \epsilon)} \right\} \quad (41)$$

where u *is defined in Algorithm 3.*

For small ϵ, (41) may be approximated by

$$\mathcal{A}^{(n)}_{\text{opt}} \approx \left\{ u \in \{0, \dots, n\} : u \geq n \cdot \left(\frac{1}{2^m - 1} + \frac{2^{m-1} - 1}{(2^m - 2)(2^m - 1)} \cdot \epsilon \right) \right\} \quad (42)$$

Thus, we have

Corollary 1. *δ_{diff} is an optimal differential distinguisher submitting n queries to the oracle if and only if* (41) *is satisfied for all $u \in \mathbb{N}$ with $1 < u \leq n$ and for all $0 < \epsilon \leq 2^m - 2$.*

It is not difficult to build artificially a situation where Algorithm 2 is not optimal: it is sufficient to take a characteristic (a, b) with $\mathrm{DP}^C(a, b)$ having a *very high probability*. In this case, it is not sufficient for δ_{diff} to wait for only *one* differential event and to stop, since if it is unique during the n samplings, it would have been better to output 1. However, if we have a look at (42), we can note that Algorithm 2 captures well real-world situations, where exploited differential probabilities are only slightly greater than ideal ones.

A very similar proof of Theorem 3 leads to

Theorem 6. *For any distinguisher in the model of δ'_{diff},*

$$1 - \frac{n+1}{2^{n\nu-1}} \leq \mathrm{BestAdv}^n_{\delta'_{\text{diff}}}(C, C^*) \leq 1 - \frac{1}{(n+1) \cdot 2^{n\nu-1}} \tag{43}$$

where $\nu = C(\mathsf{D}_0, \mathsf{D}_1)$ is the Chernoff information between D_0, a binary distribution with $\Pr_{X_0}[X_0 = 0] = 1 - \Pr_{X_0}[X_0 = 1] = \mathrm{DP}^C(a, b)$, and D_1, a binary distribution with $\Pr_{X_1}[X_1 = 0] = 1 - \Pr_{X_0}[X_1 = 1] = \frac{1}{2^m - 1}$.

Usually, in the context of differential cryptanalysis, one encounters the concept of *signal-to-noise ratio*, which was used by Biham and Shamir in the papers defining the differential cryptanalysis [1–3]; it is defined as being the ratio of probability of the right (sub-)key being suggested by a right pair and the probability of a random (sub-)key being suggested by a random pair, given the initial difference. By empirical evidence, they suggested that when this ratio is around 1-2, about 20-40 right pairs are sufficient for a successful attack, and when this ratio is higher, even 3-4 right pairs are enough; clearly, this is a (implicitly defined) likelihood-ratio test, which turns out to be optimal in terms of error probabilities.

5 Sequential Distinguishers

In this section, we formalize the concepts of *generic sequential non-adaptive distinguisher (GSNAD)* and of *n-limited generic sequential non-adaptive distinguisher (n-limited GSNAD)*. These kinds of distinguishers use sequential sampling procedures as their statistical core. We note that this idea was used earlier by Davies and Murphy (see Appendix of [10]) in an attempt to decrease the complexity of their attack against DES.

In the Luby-Rackoff model [16], a non-adaptive attacker (which may be modelized by an n-limited GNAD as described in Algorithm 4) is an infinitely powerful Turing machine which has access to an oracle Ω. It aims at distinguishing a cipher C from the "Perfect Cipher" C^* by querying Ω, and with a limited number n of inputs. The attacker must finally take a decision; usually, one is interested in measuring the ability (*i.e.* the advantage as defined in (7)) to distinguish C from C^* for a given, *fixed* amount n of queries. Clearly, in this model, one is interested in maximizing the advantage given a fixed number of queries.

In a more "real-life" situation, a cryptanalyst proceeds usually in an inverse manner: given a fixed success probability (*i.e.* a given advantage), she may look for minimizing the amount of queries to Ω, since such queries are typically expensive. With this model in head, we can now define a *n-limited generic sequential non-adaptive distinguisher* (see Algorithm 5), which turns out to be more efficient in terms of the average number of oracle queries than Algorithm 4 given a fixed advantage. In fact, such a distinguisher is adaptive regarding the decision process.

After having received the i-th response from the oracle, the distinguisher compare the i responses it has at disposal towards an acceptance set $\mathcal{A}_i$ and a rejection set $\mathcal{B}_i$, which depend on the number of queries and on the (fixed in advance) advantage, and can then take *three* different decisions: either it decides to output "0" or "1" and to stop, or to query one more question to the oracle and to repeat the decision process, until it has queried n questions. Note that $\mathcal{A}_i \subseteq \mathcal{Y}^i$ and $\mathcal{B}_i \subseteq \mathcal{Y}^i$ are disjoint sets for all $1 \leq i \leq n$ and that $\mathcal{A}_n \cup \mathcal{B}_n = \mathcal{Y}^n$. In statistics, this process is known as a *sequential decision procedure*.

We note that Algorithm 2 can be viewed as a sequential differential distinguisher which does not take explicitely into account a decision region, since it always outputs 0 as soon as it observes a "differential event".

5.1 Sequential Statistical Inference

We describe now formally the *sequential decision procedure* behind Algorithm 5. Let $\mathcal{D}$ be the set of possible decisions.

Definition 2 (Sequential decision procedure). *Let $X_1, X_2, \ldots$ be random variables observed sequentially. A* sequential decision procedure *consists in:*

1. *a* stopping rule σ_n *which specifies whether a decision must be taken without taking any further observation. If at least one observation is taken, this rule specifies for every set of observed values $(x_1, \ldots, x_n)$, with $n \geq 1$, whether to stop sampling and take a decision out of $\mathcal{D}$ or to take another observation x_{n+1}.*
2. *a* decision rule δ_n *which specifies the decision to be taken. If $n \geq 1$ observations have been taken, then one takes an action $\delta_n(x_1, \ldots, x_n) \in \mathcal{D}$. Once a decision has been taken, the sampling process is stopped.*

1: **Parameters**: a complexity n, an acceptance set $\mathcal{A}$.
2: **Input**: an oracle Ω implementing a permutation C
3: Compute some messages $\mathbf{x} = (x_1, \ldots, x_n)$.
4: Query $\mathbf{y} = (C(x_1), \ldots, C(x_n))$ to Ω.
5: **if** $\mathbf{y} \in \mathcal{A}$ **then**
6: Output 0
7: **else**
8: Output 1
9: **end if**

Algorithm 4: A n-limited generic non-adaptive distinguisher (GNAD)

1: **Parameters**: a complexity n, acceptance sets $\mathcal{A}_i, 1 \leq i \leq n$ and rejection sets $\mathcal{B}_i, 1 \leq i \leq n$.
2: **Input**: an oracle Ω implementing a permutation C
3: $i \leftarrow 1$
4: **repeat**
5: Select non-adaptively a message x_i and get $y_i = C(x_i)$.
6: **if** $(y_1, \ldots, y_i) \in \mathcal{A}_i$ **then**
7: Output 0 and stop.
8: **else if** $(y_1, \ldots, y_i) \in \mathcal{B}_i$ **then**
9: Output 1 and stop.
10: **end if**
11: $i \leftarrow i + 1$
12: **until** $i = n - 1$
13: Select non-adaptively a message x_n and get $y_n = C(x_n)$.
14: **if** $(y_1, \ldots, y_n) \in \mathcal{A}_n$ **then**
15: Output 0.
16: **else if** $(y_1, \ldots, y_n) \in \mathcal{B}_n$ **then**
17: Output 1.
18: **end if**

Algorithm 5: A n-limited sequential generic non-adaptive distinguisher

If we consider Algorithm 5 at the light of this formalism, $\mathcal{D} = \{0, 1\}$,

$$\delta_n(x_1, \ldots, x_n) = \begin{cases} 0 \text{ if } (x_1, \ldots, x_n) \in \mathcal{A}_n \\ 1 \text{ if } (x_1, \ldots, x_n) \in \mathcal{B}_n \end{cases} \tag{44}$$

and

$$\sigma_n(x_1, \ldots, x_n) = \begin{cases} \text{continue sampling if } & (x_1, \ldots, x_n) \notin \mathcal{A}_n \cup \mathcal{B}_n \\ \text{stop sampling if } & (x_1, \ldots, x_n) \in \mathcal{A}_n \cup \mathcal{B}_n \end{cases} \tag{45}$$

5.2 Sequential Decision Procedures

We have seen that Lemma 1 defines the shape of the optimal acceptance region for binary hypothesis testing. Theoretically, if one is able to compute the exact joint probability distribution of the oracle's responses when it implements both ciphers, one is able to compute the optimal acceptance region $\mathcal{A}$ for a generic n-limited distinguisher. Practically, one should notice that it seems considerably easier to compute joint probability distributions when the distinguisher is *non-adaptive*, since one can use some (maybe heuristic) statistical independence assumptions.

A sequential likelihood-ratio test uses exactly the same process to define *two* types of acceptance regions, denoted $\mathcal{A}$ and $\mathcal{B}$, respectively. So, it is always possible to define a sequential test when one has a classical test at disposal. In few words, a sequential test has three alternatives once it has received a response from the oracle: either it can conclude for one of both hypotheses, or it can decide to query more samples. In its simpler definition, a sequential ratio test has the

possibility to query *as many samples as it is needed to take a decision*, given a fixed error probability. The *expected* number of queries required to reach one of the two possible decision turns out to be less than it would need in order to make the same decision on the basis of a single *fixed-size* sample set. Of course it may happen that the sequential procedure will take more queries than the fixed-size one, but sequential sampling is a definitely economical procedure.

One may define Algorithm 5, as a *truncated sequential test*, *i.e.* one fixes an upper-bound n on the number of queries; it is still clear that such a sequential procedure cannot be worse than a fixed-size sampling procedure. In the following, we state some definitions and results about sequential hypothesis tests.

Definition 3 (Sequential Likelihood-Ratio Test). *To test* $\mathbf{X} \leftarrow \mathsf{D}_0$ *against* $\mathbf{X} \leftarrow \mathsf{D}_1$*, define two constants* $\tau_{\text{up}} > \tau_{\text{down}} > 0$ *depending on* α *and* β*, and define the likelihood ratio*

$$\text{lr}(\mathbf{x}) \triangleq \frac{f_{\mathbf{X}_1}(\mathbf{x})}{f_{\mathbf{X}_0}(\mathbf{x})}$$

The decision function at i-th step is

$$\delta_{\text{opt}} \triangleq \begin{cases} 1 \ (i.e \text{ accept } \mathbf{X} \leftarrow \mathsf{D}_1) & \text{if } \text{lr}(\mathbf{x}^{(i)}) \geq \tau_{\text{up}} \\ 0 \ (i.e. \text{ accept } \mathbf{X} \leftarrow \mathsf{D}_0) & \text{if } \text{lr}(\mathbf{x}^{(i)}) \leq \tau_{\text{down}} \\ \emptyset \ \text{ query another sample} & \text{otherwise} \end{cases} \tag{46}$$

When the observations are *independent and identically distributed*, then sequential likelihood-ratio tests have the following nice property (we refer to [27] as an excellent treatment of sequential procedure and for the proof of the following three theorems):

Theorem 7. *For testing a simple hypothesis against a simple alternative with independent, identically distributed observations, a sequential probability ratio test is optimal in the sense of* minimizing the expected sample size *among all tests having no larger error probabilities.*

The following results relate error probabilities α and β to τ_{up} and τ_{down}, and give an approximation of the expected number of samples.

Theorem 8. *Let be a sequential likelihood-ratio test with stopping bounds* τ_{up} *and* τ_{down}*, with* $\tau_{\text{up}} > \tau_{\text{down}}$ *and error probabilities* $0 < \alpha < 1$ *and* $0 < \beta < 1$*, then*

$$\tau_{\text{down}} \geq \frac{\beta}{1-\alpha} \quad \text{and} \quad \tau_{\text{up}} \leq \frac{1-\beta}{\alpha} \tag{47}$$

The approximation $\tau_{\text{down}} \triangleq \frac{\beta}{1-\alpha}$ and $\tau_{\text{up}} \triangleq \frac{1-\beta}{\alpha}$ is known as "Wald's approximation". The following theorem gives some credit to this approximation.

Theorem 9. *Let us assume we select for given* $\alpha, \beta \in]0,1[$*, where* $\alpha+\beta \leq 1$*, the stopping bounds* $\tau'_{\text{down}} \triangleq \frac{\beta}{1-\alpha}$ *and* $\tau'_{\text{up}} \triangleq \frac{1-\beta}{\alpha}$*. Then it holds that the sequential likelihood-ratio test with stopping bounds* τ'_{down} *and* τ'_{up} *has error probabilities* α' *and* β' *where*

$$\alpha' \leq \frac{\alpha}{1-\beta}, \quad \beta' \leq \frac{\beta}{1-\alpha} \quad \text{and} \quad \alpha' + \beta' \leq \alpha + \beta \tag{48}$$

By taking into account Wald's approximation, we can compute approximations of the expected number of queries:

$$\mathrm{E}_{\mathbf{X}_0}[N] \approx \frac{\alpha \log\left(\frac{1-\beta}{\alpha}\right) + (1-\alpha)\log\left(\frac{\beta}{1-\alpha}\right)}{\mathrm{E}_{X_1}[\log(f_{X_0}(x)) - \log(f_{X_0}(x))]} \tag{49}$$

and

$$\mathrm{E}_{\mathbf{X}_1}[N] \approx \frac{(1-\beta)\log\left(\frac{1-\beta}{\alpha}\right) + \beta\log\left(\frac{\beta}{1-\alpha}\right)}{\mathrm{E}_{X_1}[\log(f_{X_1}(x)) - \log(f_{X_0}(x))]} \tag{50}$$

5.3 A Toy-Example on DES

In order to illustrate advantages of sequential linear distinguishers, we have implemented a linear cryptanalysis of DES reduced to five rounds which uses a sequential distinguisher for deciding the parity of the linear approximation, *i.e.* the parity of the sum of involved key bits.

Using a static test, we needed 2800 known plaintext-ciphertext pairs in order to get a success probability of 97 %. Using a sequential strategy and for the same success probability, only 1218 samples were necessary on average. We give here both the static and the sequential decision rules.

Let S_n denote the number of times that Matsui's best linear characteristic [17] on 5-rounds DES evaluates to 0, where n is the number of known plaintext-ciphertext pairs at disposal. This linear approximation holds with probability $\frac{1}{2} + 0.01907$. The static decision rule is given by

$$\begin{cases} \text{Output ``key parity} = 0\text{'' if } S_n \geq \frac{n}{2} \\ \text{Output ``key parity} = 1\text{'' if } S_n < \frac{n}{2} \end{cases} \tag{51}$$

With 2800 known pairs at disposal, the static rule is successful in 97% of the cases.

For $\alpha = \beta \triangleq 0.025$, Wald's approximation gives $\tau_{\mathrm{up}} = 48$ and $\tau_{\mathrm{down}} = \frac{1}{48}$. The sequential rule is then defined by

$$\begin{cases} \text{Output ``key parity} = 1\text{'' if } S_n \leq \frac{n}{2} - \frac{\log \tau_{\mathrm{up}}}{2\log\left(\frac{1+2\epsilon}{1-2\epsilon}\right)} \\ \text{Output ``key parity} = 0\text{'' if } S_n \geq \frac{n}{2} + \frac{\log \tau_{\mathrm{down}}}{2\log\left(\frac{1-2\epsilon}{1+2\epsilon}\right)} \\ \text{Query another sample,} \quad \text{otherwise.} \end{cases} \tag{52}$$

where $\epsilon = 0.01907$. We repeated this experiment 1'000'000 times for 5 different keys and got the following results:

	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5
Pr[static distinguisher successful]	0.9689	0.9687	0.9684	0.9686	0.9688
Pr[sequential distinguisher successful]	0.9686	0.9684	0.9683	0.9682	0.9684
Average number of queries	1218.7	1218.7	1218.3	1219.1	1218.8

6 Links to Other Statistical Attacks

Potential applications in cryptanalysis of sequential distinguishers are numerous. As soon as one is able to derive underlying probability distributions, it is possible to define likelihood-ratios, and thus to use a sequential distinguisher. However, deriving even approximations of probability distributions may not be a trivial task in certain cases.

Furthermore, even if one has the probability distributions in hand, one should not neglect the amount of computations necessary to get the information which will be fed into the likelihood-ratio.

Under the light of the hypothesis testing paradigm, several known statistical attacks can be summarized (for which concerns their *decisional part*), and thus potentially analyzed in a simple way. The χ^2 statistical test, proposed in [29] for the first time and then used in many cryptanalytic contributions (*e.g.* see [13, 15, 11, 21]), is closely related to *generalized likelihood-ratio tests*.

Indeed, as outlined in Section §2, likelihood ratio tests are optimal for testing a simple versus a simple hypothesis. It is possible to develop a generalization of this test for use in situations in which the hypotheses are not simple (*e.g.* one tests a probability distribution depending of a parameter $\theta \in \omega_0$ against $\theta \in \omega_1$ where ω_0 and ω_1 are disjoint subsets of possible parameters. Such tests are not generally optimal, but they are typically non-optimal in situations for which no optimal test exists, and they usually perform reasonably well.

It is well-known (see for instance [26]) that Pearson's χ^2 statistic and a generalized likelihood-ratio test for a *multinomial distribution* are asymptotically equivalent. Thus, the underlying statistical decision processes in linear, differential, statistical, χ^2- and stochastic cryptanalysis are all equivalent in a statistical point of view: they try to distinguish two different (families of) probability distributions with help of a generalized likelihood-ratio test.

Another interesting attack is Harpes and Massey's *partitioning cryptanalysis* [14]. In such an attack, one defines the *imbalance* of a random variable as being a non-uniformity measure, *i.e.* as measure of distance between a uniform distribution and the distribution obtained through the partitioning process. In [14], two different imbalance measures are considered, namely the *peak imbalance* and the *squared Euclidean imbalance*: one could consider a χ^2-value or, equivalently, a generalized likelihood-ratio value as well (and maybe slightly improve its performances). Thus, the statistical problem behind this attack remains the same.

7 Conclusion

In this paper, we have used the power of some tools proposed by the theory of statistical tests for considering various situations in well-known cryptanalytic attacks, like linear and differential cryptanalysis; we improve known bounds on the asymptotical behavior of the best advantage of distinguishers implementing these attacks. Furthermore, we formalize the concept of "sequential distinguisher" and

we illustrate its potential power in a toy-example. Finally, we discuss the application of the statistical tools in a couple of known attacks; this suggests that statistical hypothesis testing theory may be a mean to unify, to characterize and to analyze most of the known attacks against block ciphers.

Acknowledgments

The author would like to thank Serge Vaudenay for many interesting and enlightning discussions.

References

1. E. Biham and A. Shamir, *Differential cryptanalysis of DES-like cryptosystems (extended abstract)*, Advances in Cryptology - CRYPTO'90, LNCS, vol. 537, Springer-Verlag, 1990, pp. 2–21.
2. ———, *Differential cryptanalysis of DES-like cryptosystems*, Journal of Cryptology **4** (1991), no. 1, 3–72.
3. ———, *Differential cryptanalysis of the Data Encryption Standard*, Springer-Verlag, 1993.
4. C. Cachin, *An information-theoretic model for steganography*, Information Hiding, 2nd International Workshop, LNCS, vol. 1525, Springer-Verlag, 1998, pp. 306–318.
5. ———, *An information-theoretic model for steganography*, Available on `http://eprint.iacr.org/2000/028/`, 2000.
6. F. Chabaud and S. Vaudenay, *Links between differential and linear cryptanalysis*, Advances in Cryptology - EUROCRYPT'94, LNCS, vol. 950, Springer-Verlag, 1995, pp. 356–365.
7. D. Coppersmith, S. Halevi, and C. Jutla, *Cryptanalysis of stream ciphers with linear masking*, Advances in Cryptology - CRYPTO'02, LNCS, vol. 2442, Springer-Verlag, 2002, pp. 515–532.
8. T. M. Cover and J. A. Thomas, *Information theory*, Wiley Series in Telecommunications, Wiley, 1991.
9. I. Csiszár and J. Körner, *Information theory: coding theorems for discrete memoryless systems*, Academic Press, 1981.
10. D. Davies and S. Murphy, *Pairs and triples of DES S-boxes*, Journal of Cryptology **8** (1995), no. 1, 1–25.
11. H. Gilbert, H. Handschuh, A. Joux, and S. Vaudenay, *A statistical attack on RC6*, Fast Software Encryption FSE'00, LNCS, vol. 1978, Springer-Verlag, 2000, pp. 65–74.
12. G.R. Grimmett and D.R. Stirzaker, *Probability and random processes*, Oxford University Press, 2001, 3rd edition.
13. H. Handschuh and H. Gilbert, χ^2 *cryptanalysis of the SEAL encryption algorithm*, Fast Software Encryption FSE'97, LNCS, vol. 1267, Springer-Verlag, 1997, pp. 1–12.
14. C. Harpes and J. Massey, *Partitioning cryptanalysis*, Fast Software Encryption FSE'97, LNCS, vol. 1267, Springer-Verlag, 1997, pp. 13–27.
15. L. Knudsen and W. Meier, *Correlations in RC6 with a reduced number of rounds*, Fast Software Encryption FSE'00, LNCS, vol. 1978, Springer-Verlag, 2000, pp. 94–108.

16. M. Luby and C. Rackoff, *How to construct pseudorandom permutations from pseudorandom functions*, SIAM Journal on Computing **17** (1988), no. 2, 373–386.
17. M. Matsui, *Linear cryptanalysis method for DES cipher*, Advances in Cryptology - EUROCRYPT'93, LNCS, vol. 765, Springer-Verlag, 1993, pp. 386–397.
18. ———, *The first experimental cryptanalysis of the Data Encryption Standard*, Advances in Cryptology - CRYPTO'94, LNCS, vol. 839, Springer-Verlag, 1994, pp. 1–11.
19. U. Maurer, *A unified and generalized treatment of authentication theory*, Proc. 13th Symp. on Theoretical Aspects of Computer Science (STACS'96), LNCS, vol. 1046, Springer-Verlag, 1996, pp. 387–398.
20. ———, *Authentication theory and hypothesis testing*, IEEE Transactions on Information Theory **46** (2000), no. 4, 1350–1356.
21. M. Minier and H. Gilbert, *Stochastic cryptanalysis of Crypton*, Fast Software Encryption FSE'00, LNCS, vol. 1978, Springer-Verlag, 2000, pp. 121–133.
22. S. Murphy, F. Piper, M. Walker, and P. Wild, *Likelihood estimation for block cipher keys*, Technical report, Information Security Group, University of London, England, 1995.
23. K. Nyberg, *Perfect nonlinear S-boxes*, Advances in Cryptology - EUROCRYPT'91, LNCS, vol. 547, Springer-Verlag, 1991, pp. 378–386.
24. National Bureau of Standards, *Data Encryption Standard*, U. S. Department of Commerce, 1977.
25. T. Pornin, *Optimal resistance against the Davies and Murphy attack*, Advances in Cryptology - ASIACRYPT'98, LNCS, vol. 1514, Springer-Verlag, 2000, pp. 148–159.
26. J. A. Rice, *Mathematical statistics and data analysis*, Duxbury Press, 1995.
27. D. Siegmund, *Sequential analysis - tests and confidence intervals*, Springer-Verlag, 1985.
28. S. Vaudenay, *Decorrelation: a theory for block cipher security*, to appear in the Journal of Cryptology, Available on `http://lasecwww.epfl.ch`.
29. ———, *An experiment on DES statistical cryptanalysis*, 3rd ACM Conference on Computer and Communications Security, ACM Press, 1996, pp. 139–147.
30. ———, *Provable security for block ciphers by decorrelation*, Proceedings of STACS'98, LNCS, vol. 1373, Springer-Verlag, 1998, Invited talk, pp. 249–275.
31. ———, *Resistance against general iterated attacks*, Advances in Cryptology - EUROCRYPT'99, LNCS, vol. 1592, Springer-Verlag, 1999, pp. 255–271.

A Statistical Information Theory

In this section, we recall some well-known results about Csiszár and Körner's method of types [9] and we apply them to derive Chernoff's information. We closely follow the organization of Chapter 12 in [8].

A.1 Method of Types

The *type* $\mathsf{D}_{\mathbf{x}}$ (or *empirical probability distribution*) of a sequence

$$\mathbf{x} = (x_1, \ldots, x_n) \quad \text{with } x_i \in \mathcal{X} \quad \text{for all } i \in \{1, \ldots, n\} \tag{53}$$

of n symbols from a set $\mathcal{X} = \{a_1, \ldots, a_{|\mathcal{X}|}\}$ is the relative proportion of occurrences of each symbol of $\mathcal{X}$, *i.e.*

$$\Pr_{\mathbf{x}}[a] \triangleq \frac{N(a|\mathbf{x})}{n} \quad \forall a \in \mathcal{X} \tag{54}$$

where $N(a|\mathbf{x})$ is the number of times the symbol a occurs in the sequence $\mathbf{x} \in \mathcal{X}^n$. We denote by $\mathcal{P}_n$ the set of types with denominator n. If $\mathsf{D}_P \in \mathcal{P}_n$, then the set of sequences of length n and type D_P is called the *type class of* D_P, and is noted $\mathcal{T}(\mathsf{D}_P)$, *i.e.*

$$\mathcal{T}(\mathsf{D}_P) \triangleq \{\mathbf{x} \in \mathcal{X}^n : \mathsf{D}_{\mathbf{x}} = \mathsf{D}_P\} \tag{55}$$

The essential power of the method of types arises from the following result, which shows that the number of types is at most polynomial in n.

Theorem 10.

$$|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|} \tag{56}$$

From this point, we will assume that the sequence $X_1, \ldots, X_n$ is drawn independently and identically distributed according to a distribution D_P. All sequences with the same type have the same probability, as shown in the following theorem.

Theorem 11. *If $X_1, \ldots, X_n$ are drawn iid according to D_P, then the probability of $\mathbf{x}$ depends only on its type and is given by*

$$\Pr_{P^n}[\mathbf{x}] = \prod_{i=1}^{n} \Pr_{P}[x_i] = 2^{-n(H(\mathbf{x})+D(\mathsf{D}_{\mathbf{x}}||\mathsf{D}_P))} \tag{57}$$

where $H(\mathbf{x})$ is the entropy[2] *of $\mathbf{x}$ and $D(\mathsf{D}_{\mathbf{x}}||\mathsf{D}_P)$ is the* Kullback-Leibler distance[3] *between the distributions $\mathsf{D}_{\mathbf{x}}$ and D_P.*

The following theorem allows to give useful bounds on the size of a type class.

Theorem 12. *For any $\mathsf{D}_P \in \mathcal{P}_n$,*

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(\mathsf{D}_P)} \leq |\mathcal{T}(\mathsf{D}_P)| \leq 2^{nH(\mathsf{D}_P)} \tag{58}$$

With help of Theorem 12, it is possible to prove the following result.

Theorem 13. *For any $\mathsf{D}_P \in \mathcal{P}_n$, and any distribution D_Q, the probability of the type class $\mathcal{T}(\mathsf{D}_P)$ under D_{Q^n} satisfies*

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(\mathsf{D}_P||\mathsf{D}_Q)} \leq \Pr_{Q^n}[\mathcal{T}(\mathsf{D}_P)] \leq 2^{-nD(\mathsf{D}_P||\mathsf{D}_Q)} \tag{59}$$

---

[2] The *entropy* of a discrete random variable $X \leftarrow \mathsf{D}_X$ is defined by $H(X) \triangleq -\sum_{x \in \mathcal{X}} \Pr_X[x] \log_2(\Pr_X[x])$.

[3] The *Kullback-Leibler distance* between two discrete probability distributions D_P and D_Q is defined to be $D(\mathsf{D}_P||\mathsf{D}_Q) \triangleq \sum_{x \in \mathcal{X}} \Pr_P[x] \log_2 \left(\frac{\Pr_P[x]}{\Pr_Q[x]}\right)$.

A.2 Sanov's Theorem

The method of types and above summarized results can be used to show Sanov's Theorem (see Theorem 14). We recall first some notions of topology. A family τ of subsets of a set $\mathcal{X}$ is a *topology* of $\emptyset \in \tau$, if $\mathcal{X} \in \tau$, if any union of sets of τ belongs to τ, and if any finite intersection of elements of τ belongs to τ. Sets that belongs to τ are called *open sets*, while complements of open sets are called *closed sets*. The *interior* of a subset $\mathcal{A} \subset \mathcal{X}$ is the union of the open subsets of $\mathcal{A}$. The *closure of* $\mathcal{A}$, is the intersection of all closed sets containing $\mathcal{A}$.

Theorem 14 (Sanov). *Let $X_1, \ldots, X_n$ be n iid random variables distributed according D_Q. Let $\mathcal{E} \subseteq \mathcal{P}_n$ be a set of probability distributions. Then*

$$\Pr_{Q^n}[\mathcal{E}] = \Pr_{Q^n}[\mathcal{E} \cap \mathcal{P}_n] \leq (n+1)^{|\mathcal{X}|} 2^{-nD(\mathsf{D}_{P^*}||\mathsf{D}_Q)} \tag{60}$$

where

$$\mathsf{D}_{P^*} = \arg\min_{\mathsf{D}_P \in \mathcal{E}} D(\mathsf{D}_P||\mathsf{D}_Q) \tag{61}$$

is the distribution in $\mathcal{E}$ that is closest to D_Q in relative entropy. If, in addition, the set $\mathcal{E}$ is the closure of its interior, then

$$\lim_{n\to+\infty} \frac{1}{n} \log \Pr_{Q^n}[\mathcal{E}] = -D(\mathsf{D}_{P^*}||\mathsf{D}_Q) \tag{62}$$

A.3 Chernoff's Information

We recall now the derivation of the highest achievable exponent for the probability of error of an optimal decision region when sampling n times the same random variable. From Lemma 1, we know that the optimum test is a likelihood-ratio test. We can rewrite this ratio $\mathrm{lr}(\mathbf{x})$ as

$$\frac{\Pr_{X_0^n}[\mathbf{x}]}{\Pr_{X_1^n}[\mathbf{x}]} \geq \tau \iff D(\mathsf{D}_{\mathbf{x}}||\mathsf{D}_{X_1^n}) - D(\mathsf{D}_{\mathbf{x}}||\mathsf{D}_{X_0^n}) \geq \frac{1}{n}\log\tau \tag{63}$$

or, in other words, it is possible to rewrite the log-likelihood ratio as the difference between th relative entropy distance of the sample type to each of the two possible distributions. Let $\mathcal{A}$ denote the set on which hypothesis $\mathbf{x} \leftarrow \mathsf{D}_{X_0^n}$ is accepted. Then, since the set $\overline{\mathcal{A}}$ is convex, one can use Theorem 14 to show that the error probability

$$\alpha^{(n)} = \Pr_{X_0^n}[\mathbf{x} \in \overline{\mathcal{A}}] \tag{64}$$

is essentially determined by the relative entropy of the closest member $\mathsf{D}_{X_0^*}$ of $\overline{\mathcal{A}}$ to D_{X_0}:

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{\alpha^{(n)}}{2^{-nD(\mathsf{D}_{X_0^*}||\mathsf{D}_{X_0})}} = 0 \tag{65}$$

Similarly,

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{\beta^{(n)}}{2^{-nD(\mathsf{D}_{X_1^*}||\mathsf{D}_{X_1})}} = 0 \tag{66}$$

where $\beta^{(n)} = \Pr_{X_1^n}[\mathbf{x} \in \mathcal{A}]$ and $\mathsf{D}_{X_1^*}$ is the closest element in $\mathcal{A}$ to distribution D_{X_1}.

Now, minimizing $D(\mathsf{D}_X||\mathsf{D}_{X_1})$ subject to the constraint

$$D(\mathsf{D}_X||\mathsf{D}_{X_1}) - D(\mathsf{D}_X||\mathsf{D}_{X_0}) \geq \frac{1}{n}\log\tau \tag{67}$$

will result in the type in $\mathcal{A}$ that is closest to D_{X_1}. Setting up the minimization of D_{X_1} subject to $D(\mathsf{D}_X||\mathsf{D}_{X_1}) - D(\mathsf{D}_X||\mathsf{D}_{X_0}) = \frac{1}{n}\log\tau$ using Lagrange multipliers, we obtain that the minimizing D_X is of the form

$$\Pr_{X_1^*}[x] \triangleq \Pr_{\lambda^*}[x] = \frac{\Pr_{X_0}[x]^\lambda \Pr_{X_1}[x]^{1-\lambda}}{\sum_{a\in\mathcal{X}} \Pr_{X_0}[a]^\lambda \Pr_{X_1}[a]^{1-\lambda}} \tag{68}$$

where λ is chosen so that $D(\mathsf{D}_{X_{\lambda^*}}||\mathsf{D}_{X_0}) - D(\mathsf{D}_{X_{\lambda^*}}||\mathsf{D}_{X_1}) = \frac{\log\tau}{n}$. Furthermore, from the symmetry of the above equation, we have $\mathsf{D}_{X_0^*} = \mathsf{D}_{X_1^*}$.

We come back to our decision problem. In the Bayesian case, the overall probability of error is the weighted sum of the two probabilities of error, and we have

$$\lim_{n\to+\infty} \frac{1}{n}\log \frac{\pi_0\alpha^{(n)} + \pi_1\beta^{(n)}}{2^{-n\min\{D(\mathsf{D}_{X_\lambda}||\mathsf{D}_{X_0}), D(\mathsf{D}_{X_\lambda}||\mathsf{D}_{X_1})\}}} \tag{69}$$

where D_{X_λ} has the form of (68). Since $D(\mathsf{D}_{X_\lambda}||\mathsf{D}_{X_0})$ increases with λ and $D(\mathsf{D}_{X_\lambda}||\mathsf{D}_{X_1})$ decreases with λ, the maximum value of

$$\min\{D(\mathsf{D}_{X_\lambda}||\mathsf{D}_{X_0}), D(\mathsf{D}_{X_\lambda}||\mathsf{D}_{X_1})\} \tag{70}$$

is attained when they are equal. So choosing λ such that

$$D(\mathsf{D}_{X_\lambda}||\mathsf{D}_{X_0}) = D(\mathsf{D}_{X_\lambda}||\mathsf{D}_{X_1}) \triangleq C(\mathsf{D}_{X_0}, \mathsf{D}_{X_1}) \tag{71}$$

yields the highest achievable exponent for the probability error and is called the *Chernoff's information*.