How Far Can We Go Beyond Linear Cryptanalysis?

Thomas Baignères, Pascal Junod, and Serge Vaudenay

EPFL
http://lasecwww.epfl.ch

Abstract. Several generalizations of linear cryptanalysis have been proposed in the past, as well as very similar attacks in a statistical point of view. In this paper, we define a rigorous general statistical framework which allows to interpret most of these attacks in a simple and unified way. Then, we explicitely construct optimal distinguishers, we evaluate their performance, and we prove that a block cipher immune to classical linear cryptanalysis possesses some resistance to a wide class of generalized versions, but not all. Finally, we derive tools which are necessary to set up more elaborate extensions of linear cryptanalysis, and to generalize the notions of bias, characteristic, and piling-up lemma.
Keywords: Block ciphers, linear cryptanalysis, statistical cryptanalysis.

1 A Decade of Linear Cryptanalysis

Linear cryptanalysis is a known-plaintext attack proposed in 1993 by Matsui [21, 22] to break DES [26], exploiting specific correlations between the input and the output of a block cipher. Namely, the attack traces the statistical correlation between one bit of information about the plaintext and one bit of information about the ciphertext, both obtained linearly with respect to $\mathrm{GF}(2)^L$ (where L is the block size of the cipher), by means of *probabilistic linear expressions*, a concept previously introduced by Tardy-Corfdir and Gilbert [30].

Soon after, several attempts to generalize linear cryptanalysis are published: Kaliski and Robshaw [13] demonstrate how it is possible to combine several independent linear correlations depending on the same key bits. In [31], Vaudenay defines another kind of attack on DES, called χ^2-attack, and shows that one can obtain an attack slightly less powerful than a linear cryptanalysis, but without the need to know precisely what happens in the block cipher. Harpes, Kramer, and Massey [7] replace the linear expressions with so-called I/O sums, i.e., balanced binary-valued functions; they prove the potential effectiveness of such a generalization by exhibiting a block cipher secure against conventional linear cryptanalysis but vulnerable to their generalization. Practical examples are the attack of Knudsen and Robshaw [15] against LOKI91 and the one of Shimoyama and Kaneko [28] against DES which both use non-linear approximations.

In [8], Harpes and Massey generalize the results of [7] by considering *partitions pairs* of the input and output spaces. Let $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n\}$ and

$\mathcal{Y} = \{\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_n\}$ be *partitions* of the input and output sets respectively, where $\mathcal{X}_i$ and $\mathcal{Y}_i$ are called *blocks*. The pair $(\mathcal{X}, \mathcal{Y})$ is called a *partition-pair* if all blocks of $\mathcal{X}$ (respectively $\mathcal{Y}$) contain the same number of plaintexts (respectively ciphertexts). A partitioning cryptanalysis exploits the fact that the probabilities $\Pr\left[(X, \mathsf{f}_k(X)) \in (\mathcal{X}, \mathcal{Y})\right]$ may not be uniformly distributed for a block cipher f_k when the plaintext X is uniformly distributed. In order to characterize the non-uniformity of a sample distribution, Harpes and Massey consider two "measures" called *peak imbalance* and *squared Euclidean imbalance*. Furthermore, they observe on toy-examples that the latter seems to lead to more successful attacks. These results are completed by Jakobsen and Harpes in [10, 9], where they develop useful bounds to estimate the resistance of block ciphers to partitioning cryptanalysis, with the help of spectral techniques; these bounds are relative to the squared Euclidean imbalance only, but this choice is not motivated in a formal way. To the best of our knowledge, the first practical example of partitioning cryptanalysis breaking a block cipher is the attack known as "stochastic cryptanalysis" [24] proposed by Minier and Gilbert against Crypton [17, 18].

In recent papers, Junod and Vaudenay [12, 11] consider linear cryptanalysis in a purely statistical framework, as it was done for the first time by Murphy et al. [25], for deriving optimal key ranking procedures and asymptotic bounds on the success probability of optimal linear distinguishers. A somewhat similar approach is chosen by Coppersmith et al. [1], except that it is adapted to stream ciphers. One can note that tight results about optimal distinguishers allow furthermore to derive useful security criteria.

Finally, the NESSIE effort resulted in a few papers investigating the power of linear (or non-linear) approximations based on different algebraic structures, like $\mathbb{Z}_4$. For instance, Parker [27] shows how to approximate constituent functions of an S-box by *any* linear function over *any* weighted alphabet. However, Parker observes that it is not straightforward to piece these generalized linear approximations together. In [29], Standaert et al. take advantage of approximations in $\mathbb{Z}_4$ by *recombining* the values in order to reduce the problem to the well-known binary case; they obtain more interesting biases comparatively to a classical linear cryptanalysis.

Notation. Throughout this paper, random variables $X, Y, \ldots$ are denoted by capital letters, whilst their realizations $x \in \mathcal{X}, y \in \mathcal{Y}, \ldots$ are denoted by small letters. The cardinal of a set $\mathcal{X}$ is denoted $|\mathcal{X}|$. The probability function of a random variable X following a distribution D is denoted $\Pr_{\mathsf{D}}[x]$ or abusively $\Pr_X[x]$, when the distribution is clear from the context. For convenience, sequence $X_1, X_2, \ldots, X_n$ of n random variables is denoted $\boldsymbol{X}^n$. Similarly, a sequence $x_1, x_2, \ldots, x_n$ of realizations is denoted $\boldsymbol{x}^n$. We call *support* of a distribution D the set of all $x \in \mathcal{X}$ such that $\Pr_{\mathsf{D}}[x] \neq 0$. As usual, "iid" means "independent and identically distributed". The transpose of a linear function h is denoted ${}^t h$. $\mathbb{1}_{\mathrm{A}}$ is 1 if the predicate A is true, 0 otherwise. Finally, "$\cdot$" denotes the inner product. The distribution function of the standard normal distribution is denoted

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{1}{2}u^2} du \ .$$

2 Optimal Distinguisher Between Two Sources

In this section, we shall consider a source generating a sequence of n iid random variables $\boldsymbol{Z}^n$ following a distribution D and taking values in a set $\mathcal{Z}$. We wonder whether $\mathsf{D} = \mathsf{D}_0$ or $\mathsf{D} = \mathsf{D}_1$ (where D_1 is referred to as an "ideal distribution"), knowing that one of these two hypotheses is true. An algorithm which takes a sequence of n realizations $\boldsymbol{z}^n$ as input and outputs either 0 or 1 is known as a *distinguisher* limited to n samples. It can be defined by an *acceptance region* $\mathcal{A} \subset \mathcal{Z}^n$ such that the distinguisher outputs 0 (respectively 1) when $\boldsymbol{z}^n \in \mathcal{A}$ (respectively $\boldsymbol{z}^n \notin \mathcal{A}$). The ability to distinguish a distribution from another is known as the *advantage* of the distinguisher and is defined by

$$\mathrm{Adv}^n_{\mathcal{A}} = \left| \Pr_{\mathsf{D}_0^n}[\mathcal{A}] - \Pr_{\mathsf{D}_1^n}[\mathcal{A}] \right| ,$$

which is a quantity an adversary would like to maximize. The distinguisher can make two types of mistakes: it can either output 0 when $\mathsf{D} = \mathsf{D}_1$ or output 1 when $\mathsf{D} = \mathsf{D}_0$. We denote α and β the respective error probabilities and $\mathrm{P_e} = \frac{1}{2}(\alpha+\beta)$ the *overall probability of error*. We can assume without loss of generality that $\mathrm{P_e} \leq \frac{1}{2}$; we easily obtain that $\mathrm{Adv}^n_{\mathcal{A}} = 1 - 2\mathrm{P_e}$.

2.1 Deriving an Optimal Distinguisher

We describe here how to derive an optimal distinguisher for the scenario described below [1, 11]. Clearly, $\mathrm{P_e} = \frac{1}{2} - \frac{1}{2}\sum_{\boldsymbol{z}^n \in \mathcal{A}} \left(\Pr_{\mathsf{D}_0^n}[\boldsymbol{z}^n] - \Pr_{\mathsf{D}_1^n}[\boldsymbol{z}^n]\right)$, and therefore that the set minimizing[1] $\mathrm{P_e}$ is

$$\mathcal{A} = \{\boldsymbol{z}^n \in \mathcal{Z}^n \ : \ \mathrm{LR}(\boldsymbol{z}^n) \geq 1\} \quad \text{where} \quad \mathrm{LR}(\boldsymbol{z}^n) = \frac{\Pr_{\mathsf{D}_0^n}[\boldsymbol{z}^n]}{\Pr_{\mathsf{D}_1^n}[\boldsymbol{z}^n]} \tag{1}$$

stands for *likelihood ratio*[2]. It defines an optimal distinguisher, i.e., with maximum advantage given a bounded number of samples and with no assumption on the computational power of the adversary.

In order to take a decision, a distinguisher defined by (1) has to keep in memory the results of the n queries, which is not feasible in practice if n grows. Fortunately, it is possible to derive an equivalent distinguisher with $|\mathcal{Z}|$ counter values $\mathrm{N}(a|\boldsymbol{z}^n)$, each one counting the number of occurrence of a certain symbol a of $\mathcal{Z}$ in the sequence $\boldsymbol{z}^n$. We summarize this in the following result.

Proposition 1. ***(Optimal Distinguisher)*** *The optimal acceptance region to test* $\mathsf{D} = \mathsf{D}_0$ *against* $\mathsf{D} = \mathsf{D}_1$ *is* $\mathcal{A}_{\mathrm{opt}} = \{\boldsymbol{z}^n \in \mathcal{Z}^n \ : \ \mathrm{LLR}(\boldsymbol{z}^n) \geq 0\}$ *where*

$$\mathrm{LLR}(\boldsymbol{z}^n) = \sum_{\substack{a \in \mathcal{Z} \\ \text{s.t. } \mathrm{N}(a|\boldsymbol{z}^n)>0}} \mathrm{N}(a|\boldsymbol{z}^n) \log \frac{\Pr_{\mathsf{D}_0}[a]}{\Pr_{\mathsf{D}_1}[a]}$$

[1] Note that we could have equivalently chosen a strict inequality in (1).
[2] The likelihood ratio builds the core of the Neyman-Pearson lemma [2, Ch. 12].

is the logarithmic likelihood ratio*, with the convention that* $\log \frac{0}{p} = -\infty$ *and* $\log \frac{p}{0} = +\infty$ *(the* $\log \frac{0}{0}$ *case can be ignored), and where* $\mathrm{N}(a|\boldsymbol{z}^n)$ *is the number of times the symbol* a *occurs in the sequence* $\boldsymbol{z}^n \in \mathcal{Z}^n$*.*

Given the number of realizations n, we can compute the exact advantage of the optimal distinguisher. Let $[\mathsf{D}_0]^n$ and $[\mathsf{D}_1]^n$ be the vectors defined by

$$[\mathsf{D}_j]^n_{(z_1,z_2,\ldots,z_n)} = \Pr_{\mathsf{D}_j^n}[z_1, z_2, \ldots, z_n] \quad \text{with} \quad j \in \{0,1\} \ ,$$

which are a specific case of n-wise distribution matrices of the Decorrelation Theory [33] in a simplified case as we have no input here, only outputs z_i. The probability that the distinguisher outputs 0 when $\mathsf{D} = \mathsf{D}_j$ is $\sum_{\boldsymbol{z}^n \in \mathcal{A}} [\mathsf{D}_j]^n_{\boldsymbol{z}^n}$, for $j \in \{0,1\}$. The advantage is thus $\left|\sum_{\boldsymbol{z}^n \in \mathcal{A}} ([\mathsf{D}_0]^n_{\boldsymbol{z}^n} - [\mathsf{D}_1]^n_{\boldsymbol{z}^n})\right|$. Since $\mathcal{A}_{\text{opt}}$ maximizes the sum, we obtain

$$\mathrm{Adv}^n_{\mathcal{A}_{\text{opt}}} = \frac{1}{2} \parallel [\mathsf{D}_0]^n - [\mathsf{D}_1]^n \parallel_1 \ ,$$

where the norm $\| \cdot \|_1$ of a vector $\boldsymbol{A}$ is defined by $\parallel \boldsymbol{A} \parallel_1 = \sum_i |\boldsymbol{A}_i|$. Note that the statistical framework of Coppersmith et al. [1] is based on this norm.

2.2 Complexity Analysis

In this section, we compute the number of queries the optimal distinguisher needs in order to distinguish D_0 from D_1, given a fixed error probability $\mathrm{P_e}$.

Definition 2. *The* relative entropy *or* Kullback-Leibler distance *between two distributions* D_0 *and* D_1 *is defined as*

$$D(\mathsf{D}_0 \parallel \mathsf{D}_1) = \sum_{z \in \mathcal{Z}} \Pr_{\mathsf{D}_0}[z] \log \frac{\Pr_{\mathsf{D}_0}[z]}{\Pr_{\mathsf{D}_1}[z]} \ ,$$

with the convention that $0 \log \frac{0}{p} = 0$ *and* $p \log \frac{p}{0} = +\infty$ *for* $p > 0$*.*

We will refer to this notion using the term relative entropy as, being non-symmetric, it is not exactly a distance. Nevertheless, it is always positive since $-\log$ is convex. Using this notation, the following proposition can be proved.

Proposition 3. *Considering that* $Z_1, Z_2, \ldots$ *is a sequence of iid random variables of distribution* D *and that* D_0 *and* D_1 *share the same support,*

$$\Pr\left[\frac{\mathrm{LLR}(\boldsymbol{Z}^n) - n\mu}{\sigma\sqrt{n}} < t\right] \xrightarrow{n \to \infty} \Phi(t) \ , \tag{2}$$

assuming that $\mu = \mu_j$ *with* $\mu_0 = D(\mathsf{D}_0 \parallel \mathsf{D}_1) \geq 0$ *and* $\mu_1 = -D(\mathsf{D}_1 \parallel \mathsf{D}_0) \leq 0$*, and that* σ^2 *is*

$$\sigma_j{}^2 = \sum_{z \in \mathcal{Z}} \Pr_{\mathsf{D}_j}[z] \left(\log \frac{\Pr_{\mathsf{D}_0}[z]}{\Pr_{\mathsf{D}_1}[z]}\right)^2 - \mu_j^2 \ , \tag{3}$$

when $\mathsf{D} = \mathsf{D}_j$ *for* $j \in \{0,1\}$*.*

Proof. We first note that the logarithmic likelihood ratio can be expressed as a sum $\mathrm{LLR}(\boldsymbol{Z}^n) = R_1 + \cdots + R_n$ where

$$R_i = \sum_{z \in \mathcal{Z}} \mathbb{1}_{Z_i = z} \log \frac{\Pr_{\mathsf{D}_0}[z]}{\Pr_{\mathsf{D}_1}[z]} \ ,$$

where every Z_i follows distribution D_j (so that the R_i's are iid). The Central Limit Theorem then states that $\Pr\left[(\mathrm{LLR}(\boldsymbol{Z}^n) - n\mu_j)/(\sigma_j\sqrt{n}) < t\right]$ converges in distribution towards $\Phi(t)$, where $\mu_j = \mathrm{E}_{\mathsf{D}_j}[R_i]$ and ${\sigma_j}^2 = \mathrm{Var}_{\mathsf{D}_j}[R_i]$. Some straightforward computations lead to the announced result. Note that the assumption that both distributions share the same support is necessary for μ_j and σ_j to be well defined. □

We now assume that the distributions D_0 and D_1 are close to each other, since it is the usual encountered case in practice.

Assumption 4. *Considering that* D_0 *is close to* D_1*, we can write*

$$\forall z \in \mathcal{Z} \ : \ \Pr_{\mathsf{D}_0}[z] = p_z + \epsilon_z \quad \textit{and} \quad \Pr_{\mathsf{D}_1}[z] = p_z \quad \textit{with} \quad |\epsilon_z| \ll p_z \ .$$

Note that in such a case we can approximate $\mathrm{LLR}(\boldsymbol{z}^n)$ by $\sum_a \mathrm{N}(a|\boldsymbol{z}^n)\epsilon_a/p_a$. Proposition 3 can now be simplified using Taylor series.

Proposition 5. *Under the hypothesis of Proposition 3 and of Assumption 4 we have, at order two:*

$$\mu_0 \approx -\mu_1 \approx \frac{1}{2}\sum_{z \in \mathcal{Z}} \frac{\epsilon_z^2}{p_z} \quad \textit{and} \quad {\sigma_0}^2 \approx {\sigma_1}^2 \approx \sum_{z \in \mathcal{Z}} \frac{\epsilon_z^2}{p_z} \ .$$

We can finally derive a heuristic theorem giving the number of samples the distinguisher needs, together with the implied probability of error, in order to distinguish close distributions with same support.

Theorem 6. *Let* $Z_1, \ldots, Z_n$ *be iid random variables over the set* $\mathcal{Z}$ *of distribution* D*,* D_0 *and* D_1 *be two distributions of same support which are close to each other, and* n *be the number of samples of the best distinguisher between* $\mathsf{D} = \mathsf{D}_0$ *or* $\mathsf{D} = \mathsf{D}_1$*. Let* d *be a real number such that*

$$n \ = \ \frac{d}{\displaystyle\sum_{z \in \mathcal{Z}} \frac{\epsilon_z^2}{p_z}} \ \approx \ \frac{d}{2D(\mathsf{D}_0 \parallel \mathsf{D}_1)} \tag{4}$$

(where $p_z = \Pr_{\mathsf{D}_1}[z]$ *and* $p_z + \epsilon_z = \Pr_{\mathsf{D}_0}[z]$*). Then, the overall probability of error is* $\mathrm{P_e} \approx \Phi(-\sqrt{d}/2)$*.*

Proof. If d is such that $\tilde{\mu} = \frac{1}{2}\sqrt{d/n}\ \tilde{\sigma}$, where $\tilde{\mu}$ and $\tilde{\sigma}$ respectively denote the approximation of μ_0 and σ_0 at order 2, we obtain (4). By definition $\mathrm{P_e} = \frac{1}{2}(1 - \Pr_{\mathsf{D}_1}[\mathrm{LLR} < 0] + \Pr_{\mathsf{D}_0}[\mathrm{LLR} < 0])$. However

$$\Pr_{\mathsf{D}_j}[\mathrm{LLR} < 0] = \Pr_{\mathsf{D}_j}\left[\frac{\mathrm{LLR} - n\mu_j}{\sigma_j\sqrt{n}} < -\frac{\sqrt{n}\mu_j}{\sigma_j}\right] \approx \Phi\left(-\frac{\sqrt{n}\mu_j}{\sigma_j}\right) \ ,$$

where we make the usual approximation that the left hand side of (2) can be approximated by $\Phi(t)$. Therefore, as Proposition 5 states that $\mu_0 \approx -\mu_1 \approx \tilde{\mu}$ and that $\sigma_0 \approx \sigma_1 \approx \tilde{\sigma}$, we have $\mathrm{P_e} \approx \frac{1}{2}\left(1 - \Phi(\sqrt{d}/2) + \Phi(-\sqrt{d}/2)\right) = \Phi(-\sqrt{d}/2)$. □

Note that it may be possible to obtain strict tight bounds instead of an approximation for $\mathrm{P_e}$ using, for instance, Chernoff bounds.

2.3 Case where the Ideal Source is Uniform

From now on, we assume that D_1 is the uniform distribution. When D_0 is a distribution whose support is $\mathcal{X}$ itself and which is close to D_1, Theorem 6 can be rewritten with

$$n = \frac{d}{|\mathcal{Z}| \sum_{z \in \mathcal{Z}} \epsilon_z^2} \ .$$

This shows that the distinguishability can be measured by means of the Euclidean distance between D_0 and D_1. In the very specific case where $\mathcal{Z} = \{0, 1\}$, we have $\epsilon_0 = -\epsilon_1 = \epsilon$ and one can see that n is proportional to ϵ^{-2}. It is a well accepted fact that the complexity of linear cryptanalysis is linked to the inverse of the square of the bias [21] which is, as we can see, a consequence of Theorem 6. We now recall what appears to be the natural measure of the bias of a distribution, considering the needed number of samples and Assumption 4.

Definition 7. *Let* $\epsilon_z = \Pr_{\mathsf{D}_0}[z] - \frac{1}{|\mathcal{Z}|}$. *The* Squared Euclidean Imbalance[3] *(*SEI*)* $\Delta(\mathsf{D}_0)$ *of a distribution* D_0 *of support* $\mathcal{Z}$ *from the uniform distribution is defined by*

$$\Delta(\mathsf{D}_0) = |\mathcal{Z}| \sum_{z \in \mathcal{Z}} \epsilon_z^2 \ .$$

It is well-known (see [6, 14]) that a χ^2 cryptanalysis needs $O(1/\Delta(\mathsf{D}_0))$ queries to succeed, which is by no means worse, up to a *constant term*, than an optimal distinguisher. Junod observed [11] that a χ^2 statistical test is asymptotically equivalent to a generalized likelihood-ratio test developed for a multinomial distribution; although such tests are not optimal in general, they usually perform reasonably well. Our results confirm this fact: a cryptanalyst will not loose any *essential* information in the case she can describe *only one* of the two distributions, but the precise knowledge of *both* distributions allows to derive an optimal attack. In other words, when it is impossible to derive both probability distributions, or when an attack involves many different distributions and only one is known, the best practical alternative to an optimal distinguisher seems to be a χ^2 attack, as proposed in [31]. This fact corroborates the intuition stipulating that χ^2 attacks are useful when one does not know precisely what happens in the attacked block cipher.

[3] Although this appellation coincide with the one of [7], note that the definitions slightly differ.

2.4 Case where the Source Generates Boolean Vectors

We assume here that random variables are bitstrings[4], so that $\mathcal{Z} = \{0,1\}^\ell$.

Definition 8. *Following the notations of Assumption 4, let* D_0 *be the distribution defined by the set* $\{\epsilon_z\}_{z\in\mathcal{Z}}$, D_1 *being the uniform distribution on* $\mathcal{Z}$. *We define the Fourier transform of* D_0 *at point* $u \in \mathcal{Z}$ *as*

$$\widehat{\epsilon}_u = \sum_{z\in\mathcal{Z}} (-1)^{u\cdot z}\ \epsilon_z \ . \tag{5}$$

The involution property of the Fourier transform leads to

$$\epsilon_z = \frac{1}{2^\ell} \sum_{u\in\mathcal{Z}} (-1)^{u\cdot z}\ \widehat{\epsilon}_u \ . \tag{6}$$

The next property can be compared to Parseval's Theorem.

Proposition 9. *In the case where* D_1 *is the uniform distribution over* $\mathcal{Z} = \{0,1\}^\ell$, *the* SEI *and the Fourier coefficients are related by:*

$$\Delta(\mathsf{D}_0) = \sum_{u\in\mathcal{Z}} \widehat{\epsilon}_u^2 \ .$$

We now recall the definition of the linear probability [23], which plays a central role in the context of linear cryptanalysis.

Definition 10. *The* linear probability *of a boolean random variable* B *is*

$$\mathsf{LP}(B) = (\Pr\,[B=0] - \Pr\,[B=1])^2 = (2\Pr\,[B=0]-1)^2 = \left(\mathrm{E}\left[(-1)^B\right]\right)^2 \ .$$

Proposition 11. *Let* $\mathcal{Z} = \{0,1\}^\ell$. *If* $Z \in \mathcal{Z}$ *is a random variable of distribution* D_0, *the* SEI *and the linear probability are related by:*

$$\Delta(\mathsf{D}_0) \ = \sum_{w\in\mathcal{Z}\setminus\{0\}} \mathsf{LP}(w\cdot Z) \ .$$

Proof. By using (5) we have $\widehat{\epsilon}_u = \mathrm{E}_{\mathsf{D}_0}\left[(-1)^{u\cdot Z}\right] - \mathbb{1}_{u=0}$. Proposition 9 gives $\Delta(\mathsf{D}_0) = \sum_{u\in\mathcal{Z}\setminus\{0\}} \left(\mathrm{E}_{\mathsf{D}_0}\left[(-1)^{u\cdot Z}\right]\right)^2 = \sum_{w\in\mathcal{Z}\setminus\{0\}} \mathsf{LP}(w\cdot Z)$. □

Corollary 12. *Let* Z *be a random variable over* $\mathcal{Z} = \{0,1\}^\ell$ *of distribution* D_0 *and let*[5] $\mathsf{LP}^Z_{\max}$ *be the maximum of* $\mathsf{LP}(w\cdot Z)$ *over* $w \in \mathcal{Z}\setminus\{0\}$. *We have*

$$\Delta(\mathsf{D}_0) \ \le\ \left(2^\ell - 1\right)\ \mathsf{LP}^Z_{\max} \ .$$

[4] Note that all the study below extends in a straightforward way to $\mathcal{Z} = \mathrm{GF}(p)^\ell$ for a prime p by replacing (-1) by $e^{\frac{2i\pi}{p}}$ and by using the conjugates of ϵ_z and $\widehat{\epsilon}_z$ in (5) and (6) respectively. For simplicity we restrict ourselves to GF(2).

[5] We make a slight abuse of notation since $\mathsf{LP}^Z_{\max}$ is not a random variable depending on Z, but a real value depending on the *distribution* of Z.

Theorem 6 and Corollary 12 together mean that the complexity of the best distinguisher between two distributions of random bit strings can decrease with a factor up to 2^ℓ when compared to the best linear distinguisher. It is interesting to note that there are cases where this bound is tight. For example if D_0 is such that $\Pr_{\mathsf{D}_0}[z]$ is $\frac{1}{2^\ell} + \left(1 - \frac{1}{2^\ell}\right)\gamma$ if $z = 0$, and $\frac{1}{2^\ell} - \frac{1}{2^\ell}\gamma$ otherwise (where γ is a positive constant), it can be shown that $\mathsf{LP}(w \cdot Z) = \gamma^2$ for all $w \neq 0$. Hence $\Delta(\mathsf{D}_0) = (2^\ell - 1)\gamma^2$ and $\mathsf{LP}_{\max} = \gamma^2$.

2.5 Statistical Distinguishers

In the last section, we have been trying to distinguish two random variables following two distinct distributions in a set $\mathcal{Z} = \{0,1\}^\ell$ where ℓ should not be too large from an implementation point of view. If we try to distinguish two random variables distributed in some set $\{0,1\}^L$ of large cardinality (e.g. where $L = 128$), we won't be able to implement the best distinguisher of Proposition 1 as the memory requirement would be too high. Instead, we can reduce the source space to a smaller space $\mathcal{Z} = \{0,1\}^\ell$ by means of a *projection*[6] $h : \{0,1\}^L \to \mathcal{Z}$ defining, for a random variable $S \in \{0,1\}^L$ of distribution $\widetilde{\mathsf{D}}$, a random variable $Z = h(S)$ of distribution D. Here we consider that h is a balanced function and that $\widetilde{\mathsf{D}}_1$ is a uniform distribution, so that D_1 is a uniform distribution as well. This is a typical construction in a real-life block cipher cryptanalysis, where the block length is quite large. Now, even though we know which distinguisher is the best to use in order to distinguish D_0 from D_1, it is still not clear how the projection h has to be chosen. Probably the most classical example arises when $\ell = 1$ and $h(S) = a \cdot S$ for some non-zero $a \in \{0,1\}^\ell$. We then talk about a *linear distinguisher*. In this case, we note that $\Delta(\mathsf{D}_0) = \mathsf{LP}(a \cdot S) \leq \mathsf{LP}^S_{\max}$. Modern ciphers protect themselves against that type of distinguisher by bounding the value of $\mathsf{LP}^S_{\max}$. A natural extension of the previous scheme would be to consider any linear projection onto wider spaces, e.g. to consider $h(S) \in \mathcal{Z} = \{0,1\}^\ell$ (where $\ell > 1$ is still small) such that h is GF(2)-linear. We then talk about an *extended linear distinguisher*. It seems natural to wonder about the complexity gap between linear cryptanalysis and this extension. The following theorem proves that if a cipher provably resists classical linear cryptanalysis, it is (to some extent) protected against extended linear cryptanalysis.

Theorem 13. *Let S be a random variable over $\{0,1\}^L$. Whenever the source space is reduced by a projection $h : \{0,1\}^L \to \{0,1\}^\ell$ in a* GF(2)*-linear way, we have $\Delta(h(S)) \leq (2^\ell - 1)\mathsf{LP}^S_{\max}$.*

Proof. We use Proposition 11 and the fact that $w \cdot h(S) = {}^t h(w) \cdot S$. □

A classical example of a linear space reduction arises when considering *concatenation* of several projections. For example, denoting $\mathsf{D}_0^{(i)} = h^{(i)}(\widetilde{\mathsf{D}}_0)$ for $i \in \{1, \ldots, \ell\}$ where $h^{(i)} : \{0,1\}^L \to \{0,1\}$ is linear, we consider $h(S) =$

[6] We borrow this appellation from Vaudenay [31]; the same expression is used within Wagner's unified view of block cipher cryptanalysis [34] as well.

$\left(h^{(1)}(S), \ldots, h^{(n)}(S)\right)$. This corresponds to the works of Kaliski and Robshaw [13] (where different linear characteristics involving *identical* key bits are merged) and of Junod and Vaudenay [12] (where different linear characteristics involving *different* key bits are merged). In the latter situation, if no assumption is made about the dependency among the $\Delta(\mathsf{D}_0^{(i)})$'s, Theorem 13 tells us $\Delta(\mathsf{D}_0^{(1)} \times \cdots \times \mathsf{D}_0^{(\ell)}) \leq (2^\ell - 1)\mathsf{LP}_{\max}^S$. The following proposition tells us what happens in general when the $\mathsf{D}_0^{(i)}$'s are independent but do not necessarily come from a linear projection nor a Boolean projection.

Proposition 14. *Consider the case where* $\mathsf{D}_0 = \mathsf{D}_0^{(1)} \times \cdots \times \mathsf{D}_0^{(\ell)}$. *If* $\mathsf{D}_0^{(1)}, \ldots, \mathsf{D}_0^{(\ell)}$ *are* independent *distributions, then* $\Delta(\mathsf{D}_0)+1 = \prod_{i=1}^{\ell} \left(\Delta(\mathsf{D}_0^{(i)}) + 1\right)$. *Therefore,* $\Delta(\mathsf{D}_0)$ *can be approximated by the sum of the* $\Delta(\mathsf{D}_0^{(i)})$*'s.*

Proof. For the sake of simplicity, we restrict this proof to the case where $\mathsf{D}_0 = \mathsf{D}_0^{(a)} \times \mathsf{D}_0^{(b)}$. Let $Z = (A, B)$ where A and B are two independent random variable following distributions $\mathsf{D}_0^{(a)}$ and $\mathsf{D}_0^{(b)}$ respectively. As in Proposition 11, we have

$$\begin{aligned}
\Delta(\mathsf{D}_0^{(a)} \times \mathsf{D}_0^{(b)}) &= \sum_{(v,w)\in\mathcal{Z}^2\setminus\{\mathbf{0}\}} \left(\mathrm{E}\left[(-1)^{v\cdot A \oplus w\cdot B}\right]\right)^2 \\
&= \sum_{(v,w)\in\mathcal{Z}^2\setminus\{\mathbf{0}\}} \left(\mathrm{E}\left[(-1)^{v\cdot A}\right]\right)^2 \left(\mathrm{E}\left[(-1)^{w\cdot B}\right]\right)^2 \\
&= \left(\Delta(\mathsf{D}_0^{(a)}) + 1\right)\left(\Delta(\mathsf{D}_0^{(b)}) + 1\right) - 1 \ .
\end{aligned}$$

□

This result tells us that merging ℓ independent biases should only be considered when their respective amplitudes are within the same order of magnitude.

In the light of the preceeding discussion, the cryptanalyst may wonder if it is possible to find a distinguisher with a high advantage even though the value of $\mathsf{LP}_{\max}^S$ is very small. We provide an example for which it is indeed the case.

Example. Consider a source generating a random variable $S = (X_1, \ldots, X_{n+1}) \in \mathbb{Z}_4^{n+1}$, where n is some odd large integer, and we represent $\mathbb{Z}_4$ by $\{0,1\}^2$ in binary. Here we have $L = 2n+2$. If the source follows distribution D_0, then $X_1, \ldots, X_n \in \mathbb{Z}_4$ are uniform iid random variables and $X_{n+1} = (Y + \sum_{i=1}^n X_i) \bmod 4$, where $Y \in \{0,1\}$ is a uniformly distributed random variable independent of $X_1, \ldots, X_n$. If the source follows distribution D_1, $S \in \mathbb{Z}_4^{n+1}$ is uniformly distributed. It can be shown (see Appendix A) that $\mathsf{LP}_{\max}^S = 2^{-(n+1)}$. On the other hand, if we let $h : \mathbb{Z}_4^{n+1} \to \mathbb{Z}_2$ be such that $h(S) = \mathrm{msb}\left((X_{n+1} - \sum_{i=0}^n X_i) \bmod 4\right)$ (where msb stands for most significant bit), we have $\ell = 1$ and a SEI equal to 1, so that $\Delta(\mathsf{D}_0) \gg \mathsf{LP}_{\max}^S$: D_0 can be distinguished from D_1 despite $\mathsf{LP}_{\max}^S$ is small.

This example shows that Theorem 13 tells us nothing about the SEI whenever the plaintext space is reduced by a non-linear projection. Therefore, even though

$\mathsf{LP}^S_{\max}$ is very low, there may exist some tricky non-linear projections which lead to significant breakdown of the complexity of distinguishers, i.e., there may be something beyond linear cryptanalysis.

3 Optimal Distinguisher Between Two Oracles

So far we discussed how to distinguish random sources. Now we investigate applications for distinguishing random oracles, such as block ciphers, and how to transform this into the previous problem.

We consider the random variable Z taking values in $\mathcal{Z}$ to be a couple of random variables (X, Y) taking values in $\mathcal{X} \times \mathcal{Y}$. As discussed in Sect. 2.5, the couple (X, Y) can be seen like the image of a plaintext/ciphertext couple (P, C) by some balanced projections ϕ and ψ (which actually define the statistical cryptanalysis in use); in other words, the adversary queries the oracle for known-plaintext pairs and compute the projections ϕ and ψ to sample (X, Y). For simplicity reasons, we focus our study on known-plaintext attacks (such as linear cryptanalysis) and thus, we consider that X is uniformly distributed. The distribution of Y is defined by a transition matrix $\boldsymbol{T}$ such that

$$[\boldsymbol{T}]_{x,y} = \Pr\ [Y = y | X = x] = \Pr\ [\psi(C) = y | \phi(P) = x] \ .$$

The transition matrix $\boldsymbol{T}$ can either be $\boldsymbol{T}_0$ or $\boldsymbol{T}_1$, where $\boldsymbol{T}_1$ is the uniform transition matrix (i.e., $[\boldsymbol{T}_1]_{x,y} = \frac{1}{|\mathcal{Y}|}$). The distribution D of Z depends on the transition matrix $\boldsymbol{T}$. We will denote it D_0 (respectively D_1) when $\boldsymbol{T} = \boldsymbol{T}_0$ (respectively $\boldsymbol{T} = \boldsymbol{T}_1$). We can see that if $\boldsymbol{T} = \boldsymbol{T}_1$, as X is uniformly distributed, the distribution D_1 of Z is also uniform. Therefore, all the results presented so far can be applied to the particular case we study here. Indeed, if we note that

$$\Pr_{\mathsf{D}}[z] = \Pr\ [X = x, Y = y] = [\boldsymbol{T}]_{x,y} \Pr\ [X = x] \ .$$

we can express Proposition 1 in terms of the transition matrices.

Proposition 15. ***(Optimal Binary Hypothesis Test with Transition Matrices)*** *The optimal acceptance region to test* $\mathsf{D} = \mathsf{D}_0$ *against* $\mathsf{D} = \mathsf{D}_1$ *(where* D_1 *is the uniform distribution), that is to test* $\boldsymbol{T} = \boldsymbol{T}_0$ *against* $\boldsymbol{T} = \boldsymbol{T}_1$*, is*

$$\mathcal{A}_{\text{opt}} = \{(\boldsymbol{x}^n, \boldsymbol{y}^n) \in \mathcal{X}^n \times \mathcal{Y}^n \ : \ \text{LLR}(\boldsymbol{x}^n, \boldsymbol{y}^n) \geq 0\}$$

where

$$\text{LLR}(\boldsymbol{x}^n, \boldsymbol{y}^n) = \sum_{\substack{(x,y) \in \mathcal{X} \times \mathcal{Y} \\ \text{s.t. } \text{N}((x,y)|\boldsymbol{z}^n) > 0}} \text{N}((x, y)|\boldsymbol{z}^n) \log \frac{[\boldsymbol{T}_0]_{x,y}}{[\boldsymbol{T}_1]_{x,y}}$$

with the conventions used in Proposition 1.

In the next sections, we derive the complexity of this distinguisher, discuss the relationship between our model and previous work, and study how Matsui's *Piling-up Lemma* [21] extends to our model.

3.1 Cryptanalysis Complexity

We introduce the notion of *bias matrix* $\boldsymbol{B} = \boldsymbol{T}_0 - \boldsymbol{T}_1$. Note that $\sum_{x\in\mathcal{X}}[\boldsymbol{B}]_{x,y} = 0$ when X is uniformly distributed and that $\sum_{y\in\mathcal{Y}}[\boldsymbol{B}]_{x,y} = 0$ in any case. Similarly to Definition 8, the Fourier transform $\widehat{\boldsymbol{B}}$ of the bias matrix $\boldsymbol{B}$ is such that

$$\left[\widehat{\boldsymbol{B}}\right]_{u,v} = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} (-1)^{u\cdot x \oplus v\cdot y} \left[\boldsymbol{B}\right]_{x,y} \ .$$

Furthermore, we define LPM, the *linear probability matrix*, by $[\mathsf{LPM}]_{u,v} = 0$ if $u = v = 0$ and by $[\mathsf{LPM}]_{u,v} = \mathsf{LP}(u \cdot X \oplus v \cdot Y)$ otherwise. It can be noted that $\left[\widehat{\boldsymbol{B}}\right]^2_{u,v} = |\mathcal{X}|^2 [\mathsf{LPM}]_{u,v}$. With the notations we just introduced, it is possible to derive the complexity of the best distinguisher between two oracles as a simple consequence of Theorem 6 and of Proposition 11.

Proposition 16. *Let n be the number of queries of the best distinguisher between $\boldsymbol{T}_0$ and $\boldsymbol{T}_1$, which are supposed to be close to each other and of same support. Then the overall probability of error is $\mathrm{P_e} \approx 1 - \Phi(\sqrt{d}/2)$, where d is a real number such that $n = d/\Delta(\mathsf{D}_0)$. Furthermore, as*

$$\Delta(\mathsf{D}_0) \ = \ \frac{|\mathcal{Y}|}{|\mathcal{X}|} \parallel \boldsymbol{B} \parallel_2^2 = \ \frac{1}{|\mathcal{X}|^2} \parallel \widehat{\boldsymbol{B}} \parallel_2^2 = \sum_{(u,v)\in\mathcal{X}\times\mathcal{Y}} [\mathsf{LPM}]_{u,v} \ ,$$

n can be equivalently expressed in terms of the bias matrix, of its Fourier transform, or of the linear probability matrix (and thus, of the linear probabilities).

Matsui's linear expressions are a very particular case of the transition matrices we have defined at the beginning of Sect. 3. Indeed, choosing balanced *linear* projections $\phi, \psi : \{0,1\}^L \to \{0,1\}$ is equivalent to choose input/output masks on the plaintext/ciphertext bits. The respective shapes of the corresponding bias matrix, of its Fourier transform, and of the LPM matrix are

$$\boldsymbol{B} = \begin{pmatrix} \epsilon & -\epsilon \\ -\epsilon & \epsilon \end{pmatrix} \quad , \quad \widehat{\boldsymbol{B}} = \begin{pmatrix} 0 & 0 \\ 0 & 4\epsilon \end{pmatrix} \quad \text{, and} \quad \mathsf{LPM} = \begin{pmatrix} 0 & 0 \\ 0 & 4\epsilon^2 \end{pmatrix} \ ,$$

where ϵ is nothing but the bias of Matsui's linear expressions. According to Proposition 16, we see that the complexity of the distinguishing attack is proportional to $\parallel \boldsymbol{B} \parallel_2^{-2}$, which is a well known result in linear cryptanalysis, for which $\parallel \boldsymbol{B} \parallel_2^2 = 4\epsilon^2$.

There is an intuitive link between linear probability matrices and *correlation matrices* [3]. Recall that the correlation matrix of a Boolean function $f : \{0,1\}^n \to \{0,1\}^m$ is the $2^m \times 2^n$ matrix $\mathsf{C}^{(f)}$ such that $\left[\mathsf{C}^{(f)}\right]_{u,v} = 2\Pr\left[u \cdot f(P) \oplus v \cdot P\right] - 1$, where the probability holds over the uniform distribution of P, so that $\left[\mathsf{C}^{(f)}\right]^2_{u,v} = \mathsf{LP}(u \cdot f(P) \oplus v \cdot P)$. We see that correlation matrices are strongly related to the linear probability matrices in the specific case where ϕ and ψ are identity functions (i.e., no reduction is performed on the plaintext space).

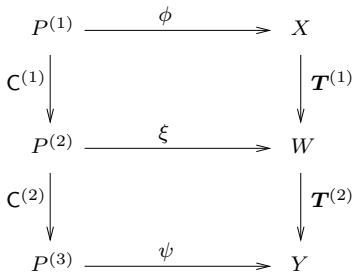

Fig. 1. Two rounds of an iterated block cipher

3.2 Piling-up Transition Matrices

A distinguishing attack on an iterated cipher is practical on condition that the cryptanalyst knows a transition matrix spanning several rounds. In practice, the cryptanalyst will derive a transition matrix on each round and, provided that the projections were chosen carefully, pile them in order to obtain a transition matrix on several rounds of the cipher.

We consider the scenario where a block cipher is represented by a random permutation C over $\{0,1\}^L$ (L denotes the block size of the cipher), where the randomness comes from the key. Moreover we suppose that the block cipher is made of two rounds corresponding to the succession of two random permutations $\mathsf{C}^{(1)}$ and $\mathsf{C}^{(2)}$. In other words $\mathsf{C} = \mathsf{C}^{(2)} \circ \mathsf{C}^{(1)}$. We denote $P^{(1)}, P^{(2)} \in \{0,1\}^L$ the respective inputs of $\mathsf{C}^{(1)}$ and $\mathsf{C}^{(2)}$, whereas $P^{(3)}$ denote the output of $\mathsf{C}^{(2)}$. The random variables X, W, and Y respectively denote $\phi(P^{(1)})$, $\xi(P^{(2)})$, and $\psi(P^{(3)})$, where ϕ, ξ, and ψ are projections onto $\mathcal{X}$, $\mathcal{W}$, and $\mathcal{Y}$, respectively. With these notations, the respective transition matrices of $\mathsf{C}^{(1)}$, $\mathsf{C}^{(2)}$, and C are

$$\left[\boldsymbol{T}^{(1)}\right]_{x,w} = \Pr_{W|X}[w \mid x] \quad , \quad \left[\boldsymbol{T}^{(2)}\right]_{w,y} = \Pr_{Y|W}[y \mid w] \quad ,$$
$$\text{and} \quad [\boldsymbol{T}]_{x,y} = \Pr_{Y|X}[y \mid x] \quad .$$

This situation is represented on Fig. 1. Note that we use a representation which is very similar to Wagner's commutative diagrams [34]. Under the assumption that $X \leftrightarrow W \leftrightarrow Y$ is a Markov chain (as in [34]), it can easily be shown that successive transition matrices are multiplicative, i.e., $\boldsymbol{T} = \boldsymbol{T}^{(1)} \times \boldsymbol{T}^{(2)}$. Note that this situation is idealistic as, even under the classical assumption that $P^{(1)} \leftrightarrow P^{(2)} \leftrightarrow P^{(3)}$ is a Markov chain [16, 32], $X \leftrightarrow W \leftrightarrow Y$ may not be a Markov chain unless the projection are chosen with care. Nevertheless, under the suitable suppositions, the following lemma shows how the Piling-Up Lemma extends to our model.

Lemma 17. *Let $\boldsymbol{B}^{(1)}$, $\boldsymbol{B}^{(2)}$, and $\boldsymbol{B}$ be the bias matrices associated with $\boldsymbol{T}^{(1)}$, $\boldsymbol{T}^{(2)}$, and $\boldsymbol{T}$ respectively, such that $\boldsymbol{T} = \boldsymbol{T}^{(1)} \times \boldsymbol{T}^{(2)}$. In the case of a known-plaintext attack, $\boldsymbol{B} = \boldsymbol{B}^{(1)} \times \boldsymbol{B}^{(2)}$ and $\widehat{\boldsymbol{B}} = \frac{1}{|\mathcal{W}|}\, \widehat{\boldsymbol{B}^{(1)}} \times \widehat{\boldsymbol{B}^{(2)}}$. Therefore,*

$\| \boldsymbol{B} \|_2^2 \leq \| \boldsymbol{B}^{(1)} \|_2^2 \| \boldsymbol{B}^{(2)} \|_2^2$, *with equality if, and only if we can write* $\left[\boldsymbol{B}^{(1)}\right]_{x,w} = \alpha_x \gamma_w$ *and* $\left[\boldsymbol{B}^{(2)}\right]_{w,y} = \gamma_w \beta_y$*, for some* $\boldsymbol{\alpha} \in \mathbb{R}^{|\mathcal{X}|}$*,* $\boldsymbol{\beta} \in \mathbb{R}^{|\mathcal{Y}|}$ *and* $\boldsymbol{\gamma} \in \mathbb{R}^{|\mathcal{W}|}$*.*

Proof. As $\boldsymbol{T} = \boldsymbol{T}^{(1)} \times \boldsymbol{T}^{(2)}$, we have

$$[\boldsymbol{B}]_{x,y} = [\boldsymbol{T}]_{x,y} - \frac{1}{|\mathcal{Y}|} = \sum_{w \in \mathcal{W}} \left(\left[\boldsymbol{B}^{(1)}\right]_{x,w} + \frac{1}{|\mathcal{W}|} \right) \left(\left[\boldsymbol{B}^{(2)}\right]_{w,y} + \frac{1}{|\mathcal{Y}|} \right) - \frac{1}{|\mathcal{Y}|} \ .$$

As $\sum_w \left[\boldsymbol{B}^{(1)}\right]_{x,w} = 0$, we obtain $[\boldsymbol{B}]_{x,y} = \left[\boldsymbol{B}^{(1)} \times \boldsymbol{B}^{(2)}\right]_{x,y} + \frac{1}{|\mathcal{W}|} \sum_w \left[\boldsymbol{B}^{(2)}\right]_{w,y}$. The fact that $P^{(1)}$ is uniformly distributed implies that $P^{(2)}$ and $P^{(3)}$ are uniformly distributed and thus, as ϕ, ξ, and ψ are balanced, that X, Z, and Y are also uniformly distributed. In that case, we know that $\sum_{w \in \mathcal{W}} \left[\boldsymbol{B}^{(2)}\right]_{w,y} = 0$, which proves that $\boldsymbol{B} = \boldsymbol{B}^{(1)} \times \boldsymbol{B}^{(2)}$. We also have

$$\begin{aligned}
\left[\widehat{\boldsymbol{B}^{(1)}} \times \widehat{\boldsymbol{B}^{(2)}}\right]_{u,v} &= \sum_{a \in \mathcal{W}} \left[\widehat{\boldsymbol{B}^{(1)}}\right]_{u,a} \left[\widehat{\boldsymbol{B}^{(2)}}\right]_{a,v} \\
&= \sum_{\substack{(x,w) \in \mathcal{X} \times \mathcal{W} \\ (w',y) \in \mathcal{W} \times \mathcal{Y}}} (-1)^{u \cdot x \oplus v \cdot y} \left[\boldsymbol{B}^{(1)}\right]_{x,w} \left[\boldsymbol{B}^{(2)}\right]_{w',y} \sum_{a \in \mathcal{W}} (-1)^{a \cdot (w \oplus w')} \\
&= |\mathcal{W}| \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} (-1)^{u \cdot x \oplus v \cdot y} \sum_{w \in \mathcal{W}} \left[\boldsymbol{B}^{(1)}\right]_{x,w} \left[\boldsymbol{B}^{(2)}\right]_{w,y} \\
&= |\mathcal{W}| \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} (-1)^{u \cdot x \oplus v \cdot y} [\boldsymbol{B}]_{x,y} \\
&= |\mathcal{W}| \left[\widehat{\boldsymbol{B}}\right]_{u,v} \ ,
\end{aligned}$$

which proves that $\widehat{\boldsymbol{B}} = \frac{1}{|\mathcal{W}|} \widehat{\boldsymbol{B}^{(1)}} \times \widehat{\boldsymbol{B}^{(2)}}$. Finally, from Cauchy-Schwarz inequality:

$$\begin{aligned}
\| \boldsymbol{B}^{(1)} \times \boldsymbol{B}^{(2)} \|_2^2 &= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \left(\sum_{w \in \mathcal{W}} \left[\boldsymbol{B}^{(1)}\right]_{x,w} \left[\boldsymbol{B}^{(2)}\right]_{w,y} \right)^2 \\
&\leq \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \left(\sum_{w \in \mathcal{W}} \left[\boldsymbol{B}^{(1)}\right]_{x,w}^2 \right) \left(\sum_{w' \in \mathcal{W}} \left[\boldsymbol{B}^{(2)}\right]_{w',y}^2 \right) \\
&= \| \boldsymbol{B}^{(1)} \|_2^2 \| \boldsymbol{B}^{(2)} \|_2^2 \ ,
\end{aligned}$$

with equality if, and only if, for all $x, y \in \mathcal{X} \times \mathcal{Y}$ there exists some $\lambda_{x,y}$ such that $\left[\boldsymbol{B}^{(1)}\right]_{x,w} = \lambda_{x,y} \left[\boldsymbol{B}^{(2)}\right]_{w,y}$, so that $\left[\boldsymbol{B}^{(1)}\right]_{x,w} = \lambda_{x,0} \left[\boldsymbol{B}^{(2)}\right]_{w,0} = \alpha_x \gamma_w$. Taking β_y equal to $\alpha_0 / \lambda_{0,y}$ when $\lambda_{0,y} \neq 0$ and to zero otherwise leads to the announced result. □

How to find projections ϕ, ψ and ξ on larger spaces exhibiting such a Markovian property in a given block cipher remains however an open question to us. We may hope to *approximate* such a Markovian process.

4 Distinguishers Versus Key Recovery

In this section we show that our framework can adapt to key recovery instead of distinguishers. Let us consider a process which generates *independent* random variables $Z_{1,K}, \ldots, Z_{n,K}$ depending on some key $K \in \{0,1\}^k$. We assume that for one unknown value $K = K_0$ all $Z_{i,K}$'s follow distribution D_0, whereas when $K \neq K_0$ all $Z_{i,K}$'s follow distribution D_1. We consider the simple key ranking procedure which, for all possible $K \in \{0,1\}^k$, instantiates $\boldsymbol{z}_K^n$ and ranks K according to the grade $\mathrm{G}_K = \mathrm{LLR}(\boldsymbol{z}_K^n)$. For any $K \neq K_0$ we obtain (similarly to what we had in Theorem 6) that $\mathrm{G}_{K_0} - \mathrm{G}_K$ is approximatively normally distributed with expected value $n\Delta(\mathsf{D}_0)$ and standard deviation $\sqrt{2n\Delta(\mathsf{D}_0)}$. Hence we obtain $\mathrm{G}_{K_0} < \mathrm{G}_K$ (i.e., a wrong key K has a better rank than the right key K_0) with probability approximatively $\Phi\left(-\sqrt{n\Delta(\mathsf{D}_0)/2}\right)$. Let d be such that $n = d/\Delta(\mathsf{D}_0)$. This probability becomes $\Phi\left(-\sqrt{d/2}\right)$ which is approximatively $e^{-d/4}/\sqrt{2\pi}$ when d is large. So K_0 gets the highest grade with probability approximatively equal to $\left(1 - e^{-d/4}/\sqrt{2\pi}\right)^{2^k-1} \approx \exp\left(-2^k \cdot e^{-d/4}/\sqrt{2\pi}\right)$, which is high provided that $d \geq 4k \log 2$. Hence we need

$$n \geq \frac{4k \log 2}{\Delta(\mathsf{D}_0)} \; .$$

This formula is quite useful to estimate the complexity of many attacks, e.g. [19, 20][7]. We can finally note that the expected rank of K_0 (from 1 up to 2^k) is $1 + (2^k - 1)\Phi\left(-\sqrt{n\Delta(\mathsf{D}_0)/2}\right)$.

5 Conclusion

Most modern block ciphers are proven to be resistant to linear cryptanalysis in some sense. In this paper, we wonder how this resistance extends to (both known and unknown) generalizations of linear cryptanalysis. For this, we define a sound and rigorous statistical framework which allows us to interpret most of these attacks in a simple and unified way; secondly, we develop a set of useful statistical tools to describe such attacks and to analyze their performance. Recently, our results on GF(2)-linear projections were exploited by [20] to obtain a small improvement factor in an attack on E0, and by [19] in another attack against two-level E0 [19]. In the sequel of this paper, we observe that resistance to linear cryptanalysis implies (a somewhat weaker) resistance to generalizations based on GF(2)-*linear projections*; however this resistance does not extend to *all* statistical cryptanalysis, as demonstrated by our example exploiting correlations in $\mathbb{Z}_4$, which confirms observations of Parker and Standaert et al. [27, 29]. The next natural step, which we hope to have rendered easier, will be to exhibit such a practical statistical cryptanalysis against a block cipher immune to linear cryptanalysis, like AES [4].

[7] Note that [19, 20] use slightly different notations: $\Delta(\mathsf{D}_0)$ denotes the Euclidean Imbalance instead of the *Squared* Euclidean Imbalance.

References

[1] D. Coppersmith, S. Halevi, and C. Jutla. Cryptanalysis of stream ciphers with linear masking. In M. Yung, editor, *Advances in Cryptology* - CRYPTO *'02*, volume 2442 of *LNCS*, pages 515–532. Springer-Verlag, 2002.

[2] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications. John Wiley & Sons, 1991.

[3] J. Daemen, R. Govaerts, and J. Vandewalle. Correlation matrices. In B. Preneel, editor, *Fast Software Encryption* - FSE *'94*, volume 1008 of *LNCS*, pages 275–285. Springer-Verlag, 1995.

[4] J. Daemen and V. Rijmen. *The Design of Rijndael*. Information Security and Cryptography. Springer-Verlag, 2002.

[5] W. Feller. *An Introduction to Probability Theory and Its Applications*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, third edition, 1968.

[6] H. Handschuh and H. Gilbert. χ^2 cryptanalysis of the SEAL encryption algorithm. In E. Biham, editor, *Fast Software Encryption* - FSE *'97*, volume 1267 of *LNCS*, pages 1–12. Springer-Verlag, 1997.

[7] C. Harpes, G. Kramer, and J. Massey. A generalization of linear cryptanalysis and the applicability of Matsui's piling-up lemma. In L.C. Guillou and J.-J. Quisquater, editors, *Advances in Cryptology* - EUROCRYPT *'95*, volume 921 of *LNCS*, pages 24–38. Springer-Verlag, 1995.

[8] C. Harpes and J. Massey. Partitioning cryptanalysis. In E. Biham, editor, *Fast Software Encryption* - FSE *'97*, volume 1267 of *LNCS*, pages 13–27. Springer-Verlag, 1997.

[9] T. Jakobsen. *Higher-order cryptanalysis of block ciphers*. PhD thesis, Department of Mathematics, Technical University of Denmark, 1999.

[10] T. Jakobsen and C. Harpes. Non-uniformity measures for generalized linear cryptanalysis and partitioning cryptanalysis. In J. Pribyl, editor, PRAGOCRYPT *'96*. CTU Publishing House, 1996.

[11] P. Junod. On the optimality of linear, differential and sequential distinguishers. In E. Biham, editor, *Advances in Cryptology* - EUROCRYPT *'03*, volume 2656 of *LNCS*, pages 17–32. Springer-Verlag, 2003.

[12] P. Junod and S. Vaudenay. Optimal key ranking procedures in a statistical cryptanalysis. In T. Johansson, editor, *Fast Software Encryption* - FSE *'03*, volume 2887 of *LNCS*, pages 235–246. Springer-Verlag, 2003.

[13] B. Kaliski and M. Robshaw. Linear cryptanalysis using multiple approximations. In Y.G. Desmedt, editor, *Advances in Cryptology* - CRYPTO *'94*, volume 839 of *LNCS*, pages 26–39. Springer-Verlag, 1994.

[14] J. Kelsey, B. Schneier, and D. Wagner. mod$n$ cryptanalysis, with applications against RC5P and M6. In L. Knudsen, editor, *Fast Software Encryption* - FSE *'99*, volume 1636 of *LNCS*, pages 139–155. Springer-Verlag, 1999.

[15] L. Knudsen and M. Robshaw. Non-linear approximations in linear cryptanalysis. In U. Maurer, editor, *Advances in Cryptology* - EUROCRYPT *'96*, volume 1070 of *LNCS*, pages 224–236. Springer-Verlag, 1996.

[16] X. Lai, J. Massey, and S. Murphy. Markov ciphers and differential cryptanalysis. In D.W. Davies, editor, *Advances in Cryptology* - EUROCRYPT *'91*, volume 547 of *LNCS*, pages 17–38. Springer-Verlag, 1991.

[17] C.H. Lim. CRYPTON: A new 128-bit block cipher. In *The First AES Candidate Conference*. National Institute for Standards and Technology, 1998.

[18] C.H. Lim. A revised version of CRYPTON: CRYPTON V1.0. In L. Knudsen, editor, *Fast Software Encryption* - FSE*'99*, volume 1636 of *LNCS*, pages 31–45. Springer-Verlag, 1999.
[19] Y. Lu and S. Vaudenay. Cryptanalysis of Bluetooth Keystream Generator Two-level E0. In *Advances in Cryptology* - ASIACRYPT*'04*, LNCS. Springer-Verlag, 2004.
[20] Y. Lu and S. Vaudenay. Faster correlation attack on Bluetooth keystream generator E0. In M. Franklin, editor, *Advances in Cryptology* - CRYPTO*'04*, volume 3152 of *LNCS*, pages 407–425. Springer-Verlag, 2004.
[21] M. Matsui. Linear cryptanalysis method for DES cipher. In T. Helleseth, editor, *Advances in Cryptology* - EUROCRYPT*'93*, volume 765 of *LNCS*, pages 386–397. Springer-Verlag, 1993.
[22] M. Matsui. The first experimental cryptanalysis of the Data Encryption Standard. In Y.G. Desmedt, editor, *Advances in Cryptology* - CRYPTO*'94*, volume 839 of *LNCS*, pages 1–11. Springer-Verlag, 1994.
[23] M. Matsui. New structure of block ciphers with provable security against differential and linear cryptanalysis. In D. Gollman, editor, *Fast Software Encryption* - FSE*'96*, volume 1039 of *LNCS*, pages 205–218. Springer-Verlag, 1996.
[24] M. Minier and H. Gilbert. Stochastic cryptanalysis of Crypton. In B. Schneier, editor, *Fast Software Encryption* - FSE*'00*, volume 1978 of *LNCS*, pages 121–133. Springer-Verlag, 2000.
[25] 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.
[26] National Institute of Standards and Technology, U. S. Department of Commerce. *Data Encryption Standard*, NIST FIPS PUB 46-2, 1993.
[27] M. Parker. Generalized S-Box linearity. Technical report `nes/doc/uib/wp5/020/a`, NESSIE Project, 2003. Available on https://www.cryptonessie.org.
[28] T. Shimoyama and T. Kaneko. Quadratic relation of S-Box and its application to the linear attack of full round DES. In H. Krawczyk, editor, *Advances in Cryptology* - CRYPTO*'98*, volume 1462 of *LNCS*, pages 200–211. Springer-Verlag, 1998.
[29] F.-X. Standaert, G. Rouvroy, G. Piret, J.-J. Quisquater, and J.-D. Legat. Key-dependent approximations in cryptanalysis: an application of multiple Z4 and non-linear approximations. In *24th Symposium on Information Theory in the Benelux*, 2003.
[30] A. Tardy-Corfdir and H. Gilbert. A known plaintext attack of FEAL-4 and FEAL-6. In J. Feigenbaum, editor, *Advances in Cryptology* - CRYPTO*'91*, volume 576 of *LNCS*, pages 172–182. Springer-Verlag, 1992.
[31] S. Vaudenay. An experiment on DES statistical cryptanalysis. In *3rd ACM Conference on Computer and Communications Security*, pages 139–147. ACM Press, 1996.
[32] S. Vaudenay. On the security of CS-cipher. In L. Knudsen, editor, *Fast Software Encryption* - FSE*'99*, volume 1636 of *LNCS*, pages 260–274. Springer-Verlag, 1999.
[33] S. Vaudenay. Decorrelation: a theory for block cipher security. *Journal of Cryptology*, 16(4):249–286, 2003.
[34] D. Wagner. Towards a unifying view of block cipher cryptanalysis. In B. Roy and W. Meier, editors, *Fast Software Encryption* - FSE*'04*, volume 3017 of *LNCS*, pages 16–33. Springer-Verlag, 2004.

A A Strange Distribution

We consider a source generating a random variable $S = (X_1, \ldots, X_{n+1}) \in \mathbb{Z}_4^{n+1}$, where n is some large integer, which follows either D_0 or D_1 (the uniform distribution). The distribution D_0 is such that the $X_1, \ldots X_n$ are uniformly distributed iid random variables and $X_{n+1} = (Y + \sum_{i=1}^{n} X_i) \bmod 4$, where $Y \in \{0,1\}$ is uniformly distributed and independent of $X_1, \ldots, X_n$. We claim that the linear probability of the best linear distinguisher with one query is very small (equal to $2^{-(n+1)}$) whereas it is still possible to find a projection h such that $Z = h(S)$ has a high SEI. In order to simplify the proof, we will suppose that $n+1$ is a multiple of 4.

Proposition 18. *Let $h : \mathbb{Z}_4^{n+1} \to \mathbb{Z}_2$ be defined by $h(S) = \mathrm{msb}\left((X_{n+1} - \sum_{i=1}^{n} X_i) \bmod 4\right)$. Then the* SEI *of $Z = h(S)$ is 1.*

The following lemmas will be used to prove that the best linear distinguisher is drastically less powerful than the distinguisher of Proposition 18.

Lemma 19. *Let $\boldsymbol{u} = u_1 u_2 \ldots u_n$ be a string of n bits. If we denote w the Hamming weight of $\boldsymbol{u}$ then we have*

$$\sum_{1 \le j < k \le n} u_j u_k = \frac{w(w-1)}{2} \ .$$

Lemma 20. *For any positive integer N, we have:*

$$\sum_{j=0}^{\lfloor N/4 \rfloor} \binom{N}{4j} = \frac{1}{4}\left(2^N + (1+i)^N + (1-i)^N\right) \quad \textit{and}$$

$$\sum_{j=0}^{\lfloor (N-1)/4 \rfloor} \binom{N}{4j+1} = \frac{1}{4}\left(2^N - i(1+i)^N + i(1-i)^N\right) \ ,$$

where i is the imaginary unit equal to $\sqrt{-1}$.

Proposition 21. *When S follows D_0 we have $\mathsf{LP}^S_{max} = 2^{-(n+1)}$.*

Proof. Each X_i is in $\mathbb{Z}_4$ so that it can be described by two bits, denoted $X_i^{\mathrm{H}} X_i^{\mathrm{L}}$. If S is considered like a bit string, a linear distinguisher will be defined by a hash function h such that

$$h(S) = \left(\bigoplus_{j=1}^{n+1} a_j X_j^{\mathrm{L}}\right) \oplus \left(\bigoplus_{j=1}^{n+1} b_j X_j^{\mathrm{H}}\right) \ ,$$

where $a_1, \ldots, a_{n+1}, b_1, \ldots, b_{n+1} \in \{0,1\}$ with at least one non-zero value. We easily prove that

$$X_{n+1}^{\mathrm{L}} \oplus Y = \bigoplus_{j=1}^{n} X_j^{\mathrm{L}} \ \text{ and } \ X_{n+1}^{\mathrm{H}} = \left(\bigoplus_{j=1}^{n} X_j^{\mathrm{H}}\right) \oplus \left(\bigoplus_{j<k\le n} X_j^{\mathrm{L}} X_k^{\mathrm{L}}\right) \oplus \left(\bigoplus_{j=1}^{n} X_j^{\mathrm{L}} Y\right) .$$

Thus, if B denotes the value of the bit $h(S)$, we have

$$B = \left(\bigoplus_{j=1}^{n}(a_j \oplus a_{n+1})X_j^{\mathrm{L}}\right) \oplus \left(\bigoplus_{j=1}^{n}(b_j \oplus b_{n+1})X_j^{\mathrm{H}}\right) \oplus a_{n+1}Y \\ \oplus \left(b_{n+1}\bigoplus_{1\le j<k\le n} X_j^{\mathrm{L}}X_k^{\mathrm{L}}\right) \oplus \left(b_{n+1}\bigoplus_{j=1}^{n} X_j^{\mathrm{L}}Y\right) \ .$$

If $b_{n+1} = 0$ we can see that (as at least one of the $a_1, \ldots, a_{n+1}, b_1, \ldots, b_n$ is strictly positive) $\Pr_{\mathsf{D}_0}[B = 0] = \frac{1}{2}$, hence $\mathsf{LP}(B) = 0$. If $b_{n+1} = 1$, we have

$$B = \left(\bigoplus_{j=1}^{n}(a_j \oplus a_{n+1})X_j^{\mathrm{L}}\right) \oplus \left(\bigoplus_{j=1}^{n}\overline{b_j}X_j^{\mathrm{H}}\right) \oplus a_{n+1}Y \\ \oplus \left(\bigoplus_{1\le j<k\le n} X_j^{\mathrm{L}}X_k^{\mathrm{L}}\right) \oplus \left(\bigoplus_{j=1}^{n} X_j^{\mathrm{L}}Y\right) \ .$$

If one of the $\overline{b_j}$'s is non-zero, then B is uniformly distributed so $\mathsf{LP}(B) = 0$. We now assume that $b_j = 1$ for all $j = 1, \ldots, n$. We have

$$B = \left(\bigoplus_{j=1}^{n}(a_j \oplus a_{n+1})X_j^{\mathrm{L}}\right) \oplus a_{n+1}Y \oplus \left(\bigoplus_{1\le j<k\le n} X_j^{\mathrm{L}}X_k^{\mathrm{L}}\right) \oplus \left(\bigoplus_{j=1}^{n} X_j^{\mathrm{L}}Y\right) \ .$$

Let us define $U_j = X_j^{\mathrm{L}} \oplus a \oplus a_j$ for $j \in \{0, \ldots, n\}$ and $U_j = Y \oplus a$ for $j = n+1$, with $a = \bigoplus_{j=1}^{n+1} a_j$. We can show that

$$B = \left(\bigoplus_{1\le j<k\le n+1} U_jU_k\right) \oplus c \ ,$$

where $c \in \{0, 1\}$ is a constant. Using Lemma 19 and denoting W the Hamming weight of the random string of bits $U_1, \ldots, U_{n+1}$ we obtain

$$\Pr[B = c] = \Pr\left[\frac{W(W-1)}{2} \equiv 0 \pmod 2\right] = \Pr[W \bmod 4 = 0 \text{ or } 1] \\ = \frac{1}{2^{n+1}}\sum_{j=0}^{\frac{n+1}{4}}\binom{n+1}{4j} + \frac{1}{2^{n+1}}\sum_{j=0}^{\lfloor\frac{n}{4}\rfloor}\binom{n+1}{4j+1} \ .$$

Using Lemma 20 we deduce

$$\Pr[B = c] = \frac{1}{2} + \frac{(1+i)^n + (1-i)^n}{4 \times 2^n} = \frac{1}{2} + \frac{\cos\left(\frac{n\pi}{4}\right)}{2^{\frac{n}{2}+1}} = \frac{1}{2} + \frac{(-1)^{\frac{n+1}{4}}}{2^{\frac{n+3}{2}}} \ ,$$

where we used the fact that $n+1$ is a multiple of 4. Finally, $\mathsf{LP}^S_{\max} = 2^{-(n+1)}$. □