J Cryptol (2019) 32:1–34
https://doi.org/10.1007/s00145-018-9308-x

Journal of
CRYPTOLOGY

Multidimensional Linear Cryptanalysis*

Miia Hermelin
Finnish Defence Forces, P.O. Box 919, 00131 Helsinki, Finland
miia.hermelin@gmail.com

Joo Yeon Cho
ADVA Optical Networking, Fraunhoferstraße 9a, 82152 Martinsried/Munich, Germany
jooyeon.cho@gmail.com

Kaisa Nyberg
Aalto University, P.O. Box 15400, 00076 Aalto, Finland
kaisa.nyberg@aalto.fi

Communicated by Kenny Paterson.

Received 27 May 2009
Online publication 12 November 2018

Abstract. Linear cryptanalysis introduced by Matsui is a statistical attack which exploits a binary linear relation between plaintext, ciphertext and key, either in Algorithm 1 for recovering one bit of information of the secret key of a block cipher, or in Algorithm 2 for ranking candidate values for a part of the key. The statistical model is based on the expected and observed bias of a single binary value. Multiple linear approximations have been used with the goal to make the linear attack more efficient. More bits of information of the key can potentially be recovered possibly using less data. But then also more elaborated statistical models are needed to capture the joint behaviour of several not necessarily independent binary variables. Also more options are available for generalising the statistics of a single variable to several variables. The multidimensional extension of linear cryptanalysis to be introduced in this paper considers using multiple linear approximations that form a linear subspace. Different extensions of Algorithm 1 and Algorithm 2 will be presented and studied. The methods will be based on known statistical tools such as goodness-of-fit test and log-likelihood ratio. The efficiency of the different methods will be measured and compared in theory and experiments using the concept of advantage introduced by Selçuk. The block cipher Serpent with a reduced number of rounds will be used as test bed. The multidimensional linear cryptanalysis will also be compared with previous methods that use biasedness of multiple linear approximations. It will be shown in the simulations that the multidimensional method is potentially more powerful. Its main theoretical advantage is that the statistical model can be given without the assumption about statistical independence of the linear approximations.

*Preliminary versions of parts of this paper appeared in Fast Software Encryption 2009 (Lecture Notes in Computer Science) and Symmetric Cryptography 2009 (Dagstuhl Seminar Proceedings).

Keywords. Linear cryptanalysis, Multidimensional linear approximation, Key recovery, Matsui's Algorithm 1, Matsui's Algorithm 2, Goodness of fit, Key ranking, Advantage.

1. Introduction

1.1. *General*

The purpose of this paper is to examine how the efficiency of key recovery attacks using linear cryptanalysis can be improved by extending the attack to multiple dimensions. One-dimensional linear cryptanalysis was introduced by Matsui [27] in 1993. The method is based on one binary linear relation involving some bits of the plaintext, ciphertext and the secret key of a block cipher. The linear relation holds for a fraction of plaintexts and therefore is called a linear approximation of the cipher. If the fraction of plaintexts that satisfy the relation deviates significantly from one half, the approximation has a large bias. In this case, the approximation is called strong, and it can be used for recovering information about the key, provided that the attacker has enough data, i.e. plaintext–ciphertext pairs.

Matsui presented two algorithms, Algorithm 1 (Alg. 1) and Algorithm 2 (Alg. 2) for iterated block ciphers. While Alg. 1 extracts one bit of information about the secret key, Alg. 2 can be used in finding several bits of the last round key of the cipher. The candidate values for the part of the last round key to be recovered are ranked according to a test statistic and the right value is expected to have the highest rank. Using the recovered part of the last round key, it is then possible to extract one more bit of information about the secret key using Alg. 1.

1.2. *Related Work*

In practice, it is often possible to find several strong linear approximations. Hence, the obvious enhancement to linear cryptanalysis is to use multiple linear approximations simultaneously. Matsui considered using two approximations already in [26]. Junod and Vaudenay analysed this approach in [22]. In an attempt to reduce the data complexities of Matsui's algorithms, Kaliski and Robshaw used several linear approximations involving the same key bits [25]. Multiple linear approximations were also used by Biryukov et al. [4] for extracting several bits of information about the key in an Alg. 1-type attack. They also extended this basic method to a combination of Alg. 1- and Alg. 2-type attacks. However, the results in both [4,25] depend on theoretical assumptions about the statistical properties of the one-dimensional linear approximations. In particular, it is assumed that they are statistically independent. Murphy noted that this assumption does not hold in general [29]. Moreover, practical experiments by Hermelin et al. [20] showed that the assumption does not always hold in the case of the block cipher Serpent.

The statistical linear distinguisher presented by Baignères et al. does not suffer from this limitation, since it works with distributions of data values instead of biases of single linear approximations [1]. This approach is applicable in case the linear approximations form a linear subspace. A linear subspace of linear approximations is called a multidimensional linear approximation. The solution presented in [1] has also another advantage over the previous approaches in [4,25]: it is based on a well-established statistical theory of log-likelihood ratio, LLR; see also [24]. Early work to this direction by Vaudenay pro-

poses to use χ^2-cryptanalysis [34]. More recently, Baignères and Vaudenay [2] studied hypothesis testing related to different distinguishing scenarios.

To realise a multidimensional linear distinguishing attack, it is necessary to calculate the probability distribution for the multidimensional approximation of the cipher. Englund and Maximov [16] and Maximov and Johansson [28] studied algorithms for computing large distributions and applied them to compute probability distributions over output domains of functions and their compositions used in building the cipher. However, this approach gets soon infeasible when the input and output sizes of the functions and their compositions exceed 32 bits. Multidimensional linear cryptanalysis allows to focus on the most essential information and control the sizes of domains over which the probability distributions are computed. Given a suitable number of strong one-dimensional linear approximations, the linear space spanned by them forms the multidimensional linear approximation for the attack. From the estimated values of biases of all the linear approximations in this linear space, one can estimate the probability distribution of the multidimensional values. The problem is then reduced to the problem of finding strong one-dimensional approximations.

1.3. *Contributions*

In this paper, the statistical theory of multidimensional linear distinguisher will be developed for extending Matsui's Alg. 1 and Alg. 2 to multiple dimensions. While in dimension one there is essentially only one statistic, the bias of a linear approximation, for realising Alg. 1 and Alg. 2 in multiple dimensions, there are several possible statistical interpretations for the key recovery problem and different statistical tests for solving them. The purpose of this work is to compare different key recovery methods for both Alg. 1 and Alg. 2.

The different methods will be compared by using the concept of advantage proposed by Selçuk [33]. Originally, the advantage was proposed to be used in measuring the success of key ranking in the one-dimensional Alg. 2. In this paper, the theory of advantage is extended to multiple dimensions and applied to study the key ranking in both Alg. 1 and Alg. 2. The advantage for different methods is determined in theory and evaluated in experiments for block cipher Serpent with a reduced number of rounds.

Both Alg. 1 and Alg. 2 can be interpreted as goodness-of-fit problems that can be solved using χ^2-test. A method based on LLR will also be studied for both algorithms. While for Alg. 1 these two methods seem to be equally efficient in practice, for Alg. 2 both theory and practice demonstrate superiority of the LLR-method over the χ^2-based method. The multidimensional methods will also be compared to the classical one-dimensional method. We also define a combination of the χ^2-based Alg. 1 and Alg. 2 and show that it is essentially identical to the combined method proposed by Biryukov et al. if in the latter, all (nonzero) linear combinations of the base approximations are included in the attack.

1.4. *Outline*

The structure of this paper is as follows: in Sect. 2, the basic statistical theory and notation is given. Results about multidimensional linear distinguishers are presented as well as the

construction of the multidimensional probability distribution using the one-dimensional linear approximations of the cipher. The advantage and the generalisation of Selçuk's theory to multiple dimensions are presented in Sect. 4. In Sect. 3, the multidimensional linear approximation of a block cipher is discussed. Different extensions of Alg. 1 to multiple dimensions are introduced in Sect. 5 with the theoretical and experimental results. The time, data and memory complexities of Alg. 1 are also considered. Similarly, Alg. 2 is studied in Sect. 6.

2. Boolean Functions and Related Statistical Concepts

2.1. *Boolean Function and Probability Distribution*

The space of n-dimensional binary vectors is denoted by $\mathbb{F}_2^n$. The sum modulo 2 is denoted by $\oplus$. The inner product for $a = (a^1, \ldots, a^n), b = (b^1, \ldots, b^n) \in \mathbb{F}_2^n$ is defined as $a \cdot b = a^1 b^1 \oplus \cdots \oplus a^n b^n$. Then the vector a is called the (linear) mask of b.

A function $f : \mathbb{F}_2^n \mapsto \mathbb{F}_2$ is called a Boolean function. A linear Boolean function is a mapping $x \mapsto u \cdot x$, where $u \in \mathbb{F}_n$. A function $f : \mathbb{F}_2^n \mapsto \mathbb{F}_2^m$ with $f = (f_1, \ldots, f_m)$, where f_i are Boolean functions, is called a vector Boolean function of dimension m. A linear Boolean function from $\mathbb{F}_2^n$ to $\mathbb{F}_2^m$ is represented by an $m \times n$ binary matrix U. The m rows of U are denoted by $u_1, \ldots, u_m$, where each u_i is a linear mask.

The correlation between a Boolean function $f : \mathbb{F}_2^n \mapsto \mathbb{F}_2$ and zero is

$$c(f) = c(f, 0) = 2^{-n} \left(\#\{\xi \in \mathbb{F}_2^n \mid f(\xi) = 0\} - \#\{\xi \in \mathbb{F}_2^n \mid f(\xi) \neq 0\} \right)$$

and it is also called the correlation of f.

We say that the vector $p = (p_0, \ldots, p_M)$ is a probability distribution (p.d.) of a random variable X taking on values in $\{0, 1, \ldots, M\}$ and denote $X \sim p$, if $\Pr(X = \eta) = p_\eta$, for all $\eta = 0, \ldots, M$. We will denote the uniform p.d. by θ. Let $f : \mathbb{F}_2^n \mapsto \mathbb{F}_2^m$ and $X \sim \theta$. We call the p.d. p of the random variable $f(X)$ the p.d. of f.

Let us recall some general properties of p.d.'s. Let $p = (p_0, \ldots, p_M)$ and $q = (q_0, \ldots, q_M)$ be p.d.'s of random variables taking on values in a set with $M+1$ elements. The Kullback–Leibler distance between p and q is defined as follows:

Definition 2.1. The *relative entropy* or *Kullback–Leibler distance* between p and q is

$$D(p \,||\, q) = \sum_{\eta=0}^{M} p_\eta \log \frac{p_\eta}{q_\eta},$$

with the conventions $0 \log 0/b = 0$, for all $b \geq 0$ and $b \log b/0 = \infty$ for $b > 0$.

The following property usually holds for p.d.'s related to any real ciphers, so it will be frequently used throughout this work.

Property 1. We say that distribution p *is close to* q if $|p_\eta - q_\eta| \ll q_\eta$, for all $\eta = 0, 1, \ldots, M$.

This is a natural property of modern ciphers with $q = \theta$ as one of their design criteria is to resist one-dimensional linear cryptanalysis, and therefore, they must have as uniform p.d.'s as possible.

If p is close to q, we can approximate their Kullback–Leibler distance using the Taylor series [1] such that

$$D(p \,||\, q) = \frac{1}{2} C(p, q) + \mathcal{O}(\varepsilon^3),$$

where $\varepsilon = \max_{\eta \in \{0,1,\ldots,M\}} |p_\eta - q_\eta|$ and the capacity $C(p, q)$ of p and q is defined as follows:

Definition 2.2. The capacity between two p.d.'s p and q is defined by

$$C(p, q) = \sum_{\eta=0}^{M} \frac{(p_\eta - q_\eta)^2}{q_\eta}.$$

If q is the uniform distribution, then $C(p, q)$ will be denoted by $C(p)$ and called the capacity of p.

The normed normal distribution with mean 0 and variance 1 is denoted by $\mathcal{N}(0, 1)$. Its probability density function (p.d.f.) is

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}$$

and the cumulative distribution function (c.d.f.) is

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)\, \mathrm{d}t.$$

The normal distribution with mean μ and variance σ^2 is denoted by $\mathcal{N}(\mu, \sigma^2)$, and its p.d.f. and c.d.f. are ϕ_{μ,σ^2} and Φ_{μ,σ^2}, respectively.

The following standard approximation of c.d.f. $\Phi(x)$ can be found in common statistical reference books, e.g. [32]. The approximation

$$\Phi(x) \approx 1 - \frac{\phi(x)}{x} \tag{1}$$

holds for large values of x and will be used in the derivations of this paper.

Let $X_1, \ldots, X_M$ be independent random variables, with $X_i \sim \mathcal{N}(\mu_i, 1)$, $i = 1, \ldots, M$ and let $\lambda = \sum_{i=1}^{M} \mu_i^2$. If all μ_i, $i = 1, \ldots, M$, are equal to zero, then the sum of the squares

$$\sum_{i=1}^{M} X_i^2 \tag{2}$$

is χ^2_M-distributed with M degrees of freedom and has mean M and variance $2M$. Otherwise the sum (2) follows the non-central $\chi^2_M(\lambda)$ distribution which has mean $\lambda + M$ and

variance $2(M+2\lambda)$. If $M > 30$, we may approximate $\chi^2_M(\lambda) \sim \mathcal{N}(\lambda+M, 2(M+2\lambda))$ [13].

Let $X_1, \ldots, X_N$ be a sequence of independent and identically distributed (i.i.d.) random variables where either $X_1, \ldots, X_N \sim p$ (corresponding to null hypothesis H_0) or $X_1, \ldots, X_N \sim q \neq p$ (corresponding to alternate hypothesis H_1). The hypothesis testing problem is then to determine whether to accept or reject H_0.

We can make two types of error in the test. In type I error, we reject H_0 when it is true. The level α of the test measures how probably this will happen: $\alpha = \Pr(H_1 \mid H_0)$. In type II error, we accept H_0, when it is not true. This is measured by the power $1-\beta$ of the test, defined as $\beta = \Pr(H_0 \mid H_1)$.

According to the Neyman–Pearson lemma [11], given empirical data $\hat{x}_1, \ldots, \hat{x}_N$, the optimal statistic for solving this problem, that is, distinguishing between p and q, is the log-likelihood ratio (LLR) defined by

$$\mathrm{LLR}(\hat{q}, p, q) = \sum_{\eta=0}^{M} N\hat{q}_\eta \log \frac{p_\eta}{q_\eta}, \tag{3}$$

where $\hat{q} = (\hat{q}_0, \ldots, \hat{q}_M)$ is the empirical p.d. calculated by

$$\hat{q}_\eta = \frac{1}{N}\#\{t = 1, \ldots, N \mid \hat{x}_t = \eta\}.$$

The distinguisher accepts H_0 and outputs p (or rejects H_0 and outputs q,) if $\mathrm{LLR}(\hat{q}, p, q) \geq \tau$ (or $\mathrm{LLR}(\hat{q}, p, q) < \tau$, respectively) where τ is the threshold. The threshold depends on the level and the power of the test. Usually $\tau = 0$.

The proof of the following result can be found in [11], see also [1].

Proposition 2.1. *The LLR-statistic calculated from i.i.d. empirical data $\hat{x}_t$, $t = 1, \ldots, N$ using (3) is asymptotically normal with mean and variance $N\mu_0$ and $N\sigma_0^2$ ($N\mu_1$ and $N\sigma_1^2$, resp.) if the data are drawn from p (q, resp.). The means and variances are given by*

$$\mu_0 = D(p \,||\, q) \quad \mu_1 = -D(q \,||\, p)$$
$$\sigma_0^2 = \sum_{\eta=0}^{M} p_\eta \log^2 \frac{p_\eta}{q_\eta} - \mu_0^2 \quad \sigma_1^2 = \sum_{\eta=0}^{M} q_\eta \log^2 \frac{p_\eta}{q_\eta} - \mu_1^2.$$

Moreover, if p is close to q, the following estimates hold

$$\mu_0 \approx -\mu_1 \approx \frac{1}{2}C(p,q) \quad \sigma_0^2 \approx \sigma_1^2 \approx C(p,q).$$

The data complexity of the test is defined as the amount of data N needed for successfully solving the hypothesis testing problem between p and q with given power and level of test. Assuming that p is close to q, the following corollary is obtained from Proposition 2.1 [1].

Corollary 2.1. *Assume that p is close to q. Then the data complexity N needed for distinguishing p from q with level α and power $1-\beta$ can be estimated as follows*

$$N \approx \frac{(z_\alpha - z_\beta)^2}{C(p,q)},$$

where $\Phi(z_\alpha) = \alpha$ and $\Phi(z_\beta) = 1-\beta$.

2.2. *Multidimensional Approximation of Boolean Functions*

Let $f : \mathbb{F}_2^n \to \mathbb{F}_2^{n'}$ be a vector Boolean function. Its one-dimensional linear approximation with input mask $u \in \mathbb{F}_2^n$ and output mask $w \in \mathbb{F}_2^{n'}$ is the Boolean function

$$x \mapsto u \cdot x \oplus w \cdot f(x), \tag{4}$$

with some (non-negligible) correlation c. For many ciphers, it is possible to find several such approximations (4) with non-negligible correlations. The aim of multidimensional linear cryptanalysis is to efficiently exploit given one-dimensional approximations with non-negligible correlations to obtain information about the cipher.

Linear approximations

$$u_i \cdot x \oplus w_i \cdot f(x),\ i = 1, \ldots, m, \tag{5}$$

are said to be linearly independent if the mask pairs (u_i, w_i), $i = 1, \ldots, m$, considered as concatenated vectors of length $n + n'$ are linearly independent.

Given a set of one-dimensional approximations of f, let m be the dimension of the linear space spanned by them. We call this set a multidimensional linear approximation of f of dimension m and it can be given by

$$Ux \oplus Wf(x), \tag{6}$$

where $\oplus$ is the component-wise modulo 2 sum and $U = (u_1, \ldots, u_m)$ and $W = (w_1, \ldots, w_m)$ are linear matrices from $\mathbb{F}_2^n \mapsto \mathbb{F}_2^m$ and $\mathbb{F}_2^{n'} \mapsto \mathbb{F}_2^m$, respectively. The linear approximations $u_i \cdot x \oplus w_i \cdot f(x)$, $i = 1, \ldots, m$, are called base approximations.

Then we need to calculate the multidimensional p.d. p of the m-dimensional approximation. We observe that it defines a vector Boolean function and recall that by the p.d. of a vector Boolean function $g : \mathbb{F}_2^n \mapsto \mathbb{F}_2^m$ we mean the p.d. of the random variable $g(X)$, where $X \sim \theta$. The following result connects the p.d. of a vector Boolean function g and its one-dimensional correlations $c(a \cdot g)$; see, for example, [20] or [18].

Proposition 2.2. *Let $g : \mathbb{F}_2^n \mapsto \mathbb{F}_2^m$ be a Boolean function with p.d. p and one-dimensional correlations $c(a \cdot g)$, $a \in \mathbb{F}_2^m$. Then*

$$p_\eta = 2^{-m} \sum_{a \in \mathbb{F}_2^m} (-1)^{a \cdot \eta} c(a \cdot g),\ \eta = 0, 1, \ldots, 2^m - 1.$$

Hence, for determining the p.d. p of approximation (6), we have to determine the correlations $c(a \cdot (Ux \oplus Wf(x)))$ of all the one-dimensional approximations of f. That is, p is determined based on one-dimensional projections of f, which is a known principle in statistics due to Cramér and Wold [12].

The following corollary is obtained from Proposition 2.2 using Parseval's theorem. An equivalent form of it can be found in [1], where the proof was based on the inverse Walsh–Hadamard transform of the deviations ε_η from the uniform distribution, $\varepsilon_\eta = p_\eta - 2^{-m}$.

Corollary 2.2. *Let $g : \mathbb{F}_2^n \mapsto \mathbb{F}_2^m$ be a Boolean function with p.d. p and one-dimensional correlations $c(a \cdot g)$. Then*

$$C(p) = 2^m \sum_\eta \varepsilon_\eta^2 = \sum_{a \neq 0} c(a \cdot g)^2.$$

We will need this equality in the next section where we study how linear distinguishing is done in multiple dimensions.

2.3. *Multidimensional Linear Distinguishers*

A linear distinguisher determines whether given data $\hat{x}_t, t = 1, \ldots, N$, are drawn from a cipher with p.d. $p \neq \theta$ or a random source with p.d. θ. In the one-dimensional case, the attacker uses one linear approximation such as (4) with correlation c. The data complexity is inversely proportional to c^2 [23,27].

When using multiple linear approximations

$$u_i \cdot x \oplus w_i \cdot f(x),\ i = 1, \ldots, m, \tag{7}$$

with non-negligible correlations $c_i = c(u_i \cdot x \oplus w_i \cdot f(x))$ and drawing data $\hat{x}_t$, $t = 1, \ldots, N$, from the cipher, the empirical correlations $\hat{c}_i, i = 1, \ldots, m$, of the approximations (7) are calculated as

$$\hat{c}_i = 2\frac{\#\{t = 1, \ldots, N \mid u_i \cdot \hat{x}_t \oplus w_i \cdot f(\hat{x}_t) = 0\}}{N} - 1.$$

The distinguisher based on the method of Biryukov et al. [4] is given as the ℓ_2-distance between the vectors $\mathbf{c} = (c_1, \ldots, c_m)$ and $\hat{\mathbf{c}} = (\hat{c}_1, \ldots, \hat{c}_m)$:

$$\left|\left|\hat{\mathbf{c}} - \mathbf{c}\right|\right|_2^2. \tag{8}$$

Under the assumption that the approximations (7) are statistically independent, it was proved in [4] that the data complexity of the distinguisher (8) is inversely proportional to

$$\vec{c}^2 = \sum_{i=1}^m c_i^2. \tag{9}$$

This notion was defined in [4] and called as capacity of the set of linear approximations. This result means a significant improvement in data complexity when compared to one-dimensional method, but relies on the assumption that the approximations are statistically independent. Later in [29], this assumption was questioned and shown that it does not hold in general. The experiments in [20] on reduced round Serpent confirmed this. Moreover, verifying statistical independence of a set of linear approximations is essentially equally hard as computing correlations of all their linear combinations. Indeed, given independence, all correlations can be computed given the correlations of a set of base approximations using the piling-up lemma. And given the correlations, statistical independence can be proved by the inverse of piling-up lemma.

A truly multidimensional approach to the distinguishing problem was given in [1] based on the LLR-statistic (3). By Corollary 2.1, the data complexity of a multidimensional linear distinguisher with p.d. p is inversely proportional to $C(p)$. By Corollary 2.2, we see that $C(p) \geq \bar{c}^2$. This indicates that data complexity for the distinguisher (8) with statistically independent linear approximations is larger than for the multidimensional LLR distinguisher. Moreover, using LLR and multidimensional p.d., we do not need to assume that the used approximations are statistically independent.

3. Linear Approximation of a Block Cipher

We will be applying the statistical methods of Sect. 2.1 in the case where $M = 2^m - 1$ from now on. Biryukov et al. proposed in Sect. 3.4. in [4] to extend the set of m linearly independent approximations to m', $m \leq m' \leq M$, approximations by including linear combinations of the base approximations with non-negligible correlations. It was argued that the same rule that the data complexity is inversely proportional applies also for the larger set of m' linearly and statistically dependent approximations. If this holds, then by Corollary 2.2 Biryukov's distinguisher should converge to the multidimensional distinguisher by adding all linear combinations of the base approximations such that $m' = M$. However, since the statistic (8) is based on the assumption of statistical independence, its data complexity cannot be determined accurately.

An optimal case would be where all the M linear approximations used for a multidimensional distinguisher have equally large correlations (in absolute value). Then using all M of them might allow reducing data complexity. On the other hand, if only a single one-dimensional approximation from a set of M approximations has a large correlation, then it is not useful to include the others in the distinguisher. In reality, all cases between these too extremes occur.

In this paper, we will present a number of methods for generalising the key recovery attacks based on Matsui's Alg. 1 and Alg. 2 to multiple dimensions using the statistical framework of multidimensional distinguishers. For each method, we will derive an explicit estimate of data complexity and show that while the data complexity decreases as the capacity of the linear approximations increases, the mere number of the linear approximations can have an opposite effect which varies depending on the method.

We have chosen the block cipher Serpent [3] as the test bed for our methods. Unlike the block cipher DES, it has been designed to resist linear cryptanalysis and therefore does not have individual strong linear approximations. Still it has many linear approx-

imations with correlations significantly stronger than the average that could potentially be combined and used in a multidimensional linear attack. Such linear approximations for Serpent have been previously found and used in experiments for Biryukov's multiple linear cryptanalysis by Collard et al. [10].

Let us study an iterated block cipher with block size n. Let x be the plaintext and y the output of the cipher after R rounds. The cipher key is expanded using a key-scheduling algorithm to a sequence of round keys. We denote by K the expanded key, that is, the vector consisting of all round key bits used in R rounds, and by h the length of K. Then a block cipher is a vector Boolean function with input $(x, K) \in \mathbb{F}_2^n \times \mathbb{F}_2^h$. By (6), an m-dimensional linear approximation of the block cipher can be considered as a vector Boolean function

$$\mathbb{F}_2^n \times \mathbb{F}_2^h \to \mathbb{F}_2^m,\ (x, K) \mapsto Ux \oplus Wy \oplus VK, \tag{10}$$

where U and W are $m \times n$ binary matrices. The matrix V has also m rows and it divides the expanded keys into 2^m equivalence classes $g = VK$, $g \in \mathbb{F}_2^m$.

Let us denote by $p(\eta|K)$ the probabilities of the m-bit values η of (10) for a fixed key K. In the analysis of this paper, it is assumed that for each $\eta \in \mathbb{F}_2^m$ the probabilities $p(\eta|K)$ are (about) equal for all K. We denote by p be the p.d. of (10) for a fixed key K, for all $K \in \mathbb{F}_2^h$. Then for each $g \in \mathbb{F}_2^m$, the data $U\hat{x}_t \oplus W\hat{y}_t$, $t = 1, \ldots, N$, are drawn from p.d. p^g, a fixed permutation of p determined by g, and all p.d. p^g, $g \in \mathbb{F}_2^m$, are each other's permutations. In particular, $p^g_{\eta \oplus k} = p^{g \oplus k}_\eta$, for all $g, \eta, k \in \mathbb{F}_2^m$.

By symmetry, the following results apply:

$$D(p \,||\, \theta) = D(p^g \,||\, \theta) \text{ and } C(p) = C(p^g), \text{ for all } g \in \mathbb{F}_2^m, \tag{11}$$

and

$$\min_{g, g \neq g'} D(p^g \,||\, p^{g'}) = \min_{g \neq 0} D(p^g \,||\, p) \text{ and } \min_{g, g \neq g'} C(p^g, p^{g'}) = \min_{g \neq 0} C(p^g, p), \text{ for all } g' \in \mathbb{F}_2^m, \tag{12}$$

where the minimum Kullback–Leibler distance and capacity will be denoted by $D_{\min}(p)$ and $C_{\min}(p)$, respectively. Moreover, if $D_{\min}(p) = 0$, we need to join the corresponding key classes to one class such that we may assume $D_{\min}(p) \neq 0$ and $C_{\min}(p) \neq 0$.

4. Advantage in Key Ranking

In a key recovery attack, a set of candidate keys is given, and the problem is to determine which candidate is the right one. Let the keys be searched from a set $\mathbb{F}_2^l$ of all 2^l strings of l bits. The algorithm consists of four phases: the *counting phase*, *analysis phase*, *sorting phase* and *searching phase* [34]. In the counting phase, data, for example plaintext–ciphertext pairs, are collected from the cipher. In the analysis phase, a real-valued statistic T is used in calculating a mark $T(\kappa)$ for all candidates $\kappa \in \mathbb{F}_2^l$.

In the sorting phase, the candidates κ are sorted, i.e. ranked, according to their marks $T(\kappa)$. Optimally, the right key, denoted by κ_0, should be at the top of the list. If this is not the case, then in the search phase the candidates in the list are tested until κ_0 is found.

The goal of this paper is to find a ranking statistic T that is easy to compute and that is also reliable and efficient in finding the right key.

The time complexity of the search phase, given amount N of data, was measured using a special purpose quantity "gain" in [4]. A similar but more generally applicable concept of "advantage" was introduced by Selçuk [33], where it was defined as follows:

Definition 4.1. Given a data sample obtained using an unknown fixed key, a key recovery attack for an l-bit key is said to achieve an advantage of a bits over exhaustive key search, if the correct key is ranked among the top $r = 2^{l-a}$ out of all 2^l key candidates.

Statistical tests for key recovery attacks are based on the wrong key hypothesis [17]. We state it as follows:

Assumption 1. (*Wrong key hypothesis*) There are two p.d.'s q and q', $q \neq q'$, such that for the right key κ_0, the data are drawn from q and for any wrong candidate key $\kappa \neq \kappa_0$ the data are drawn from $q' \neq q$.

A real-valued statistic T is computed from q and q', where one of these p.d.'s may be unknown, and the purpose of a statistic T is to distinguish between q and q'. We use D_R to denote the p.d. such that $T(\kappa_0) \sim D_R$. We assume $D_R = \mathcal{N}(\mu_R, \sigma_R^2)$, with parameters μ_R and σ_R, as this is the case with all statistics in this paper. Then μ_R and σ_R are determined with the help of linear cryptanalysis. We denote by D_W the p.d. known based on the wrong key hypothesis such that $T(\kappa) \sim D_W$ for all $\kappa \neq \kappa_0$. The p.d.f. and c.d.f. of D_W are denoted by f_W and F_W, respectively.

Ranking the candidates κ according to T means rearranging the 2^l random variables $T(\kappa)$, $\kappa \in \mathbb{F}_2^l$, in decreasing (or sometimes increasing) order of magnitude. Writing the ordered random variables as $T_{(0)} \geq T_{(1)} \geq \cdots \geq T_{(i)} \geq$, we call $T_{(i)}$ the ith order statistic. Let us fix the advantage a such that the right key should be among the $r = 2^{l-a}$ highest ranking key candidates. Hence, the right key should be at least as high as the rth wrong key with mark $T_{(r)}$. If the random variables $T(\kappa)$, $\kappa \neq \kappa_0$, are statistically independent, then by Theorem 1 in [33] the random variable $T_{(r)}$ is distributed as

$$\begin{aligned} &T_{(r)} \sim \mathcal{N}(\mu_a, \sigma_a^2), \text{ where} \\ &\mu_a = F_W^{-1}(1 - 2^{-a}) \text{ and } \sigma_a \approx \frac{2^{-(l+a)/2}}{f_W(\mu_a)}. \end{aligned} \tag{13}$$

Let us define the success probability P_S of having κ_0 among the r highest ranking candidates. If all the random variables $T(\kappa)$, $\kappa \in \mathbb{F}_2^n$ are statistically independent, we have

$$P_S = \Pr(T(\kappa_0) - T_{(r)} > 0) = \Phi\left(\frac{\mu_R - \mu_a}{\sqrt{\sigma_R^2 + \sigma_a^2}}\right) \tag{14}$$

since $T(\kappa_0) - T_{(r)} \sim \mathcal{N}\big(\mu_R - \mu_a, \sigma_R^2 + \sigma_a^2\big)$.

As the data complexity N depends on the parameters $\mu_R - \mu_a$ and $\sigma_R^2 + \sigma_a^2$, we can solve N from (14) as a function of a and vice versa. Hence, (14) describes the trade-off between the data complexity N and the time complexity of the search phase.

5. Algorithm 1

5.1. *Finding the Right Key Class*

Assume that we have found an m-dimensional approximation (10) with non-uniform p.d. p. The output y is the ciphertext obtained from the cipher by encrypting plaintext x over R rounds using the key K. We denote by g_0 the right key class to which K belongs. Our goal is to find g_0.

In the counting phase, we draw N plaintext–ciphertext pairs $(\hat{x}_t, \hat{y}_t)$, $t = 1, \ldots, N$, from the cipher. From the empirical data $U\hat{x}_t \oplus W\hat{y}_t$, $t = 1, \ldots, N$, the empirical p.d. $\hat{q}$ is computed.

To recall how Alg. 1 works for $m = 1$, let us denote by c ($\hat{c}$) the theoretical (empirical) correlation of $u \cdot x \oplus w \cdot y$. The decision in Alg. 1 in one dimension is based on the following test: the key class candidate bit g is chosen to be 0 if $c\hat{c} > 0$. Otherwise, $g = 1$. In other words, the statistical decision problem is to determine which of the two distributions $(\frac{1}{2}(1 \pm c), \frac{1}{2}(1 \mp c))$ gives the best fit with the data.

In multiple dimensions, the same data will be used for ranking the different candidate classes $g \in \mathbb{F}_2^m$. Hence, the corresponding marks $T(g)$ will be *statistically dependent*. We are not aware of any general method of calculating the c.d.f. of the order statistic $T_{(r)}$ of statistically dependent random variables. The asymptotic c.d.f. of the maximum of normal, identically distributed but dependent random variables for large $M = 2^m - 1$, $m \geq 7$ is derived in Sect. 9.3. in [14]. However, the problem still remains as the random variables $T(g_0)$ and $\max_{g \neq g_0} T(g)$ are statistically dependent.

Denote by $N(g)$ the data complexity of ranking g with advantage a, if g is the right key. We will assume $T(g)$'s to be statistically independent to determine $N(g)$. The assumption of statistical independence of $T(g)$'s could be avoided by drawing $\sum_{g \in V_m} N(g) \approx 2^m \max_g N(g)$ words of data, as then the right key class g_0 would be ranked with advantage a and each mark $T(g)$, $g \in \mathbb{F}_2^m$ could be calculated from different data. However, the resulting complexity estimate would be far too large to be of practical value.

We will study three different ways to generalise the one-dimensional Alg. 1 to multiple dimensions. Since the data are drawn i.i.d. from the p.d. p^{g_0} and not from any other p.d. p^g, $g \neq g_0$, we can interpret the problem of finding g_0 as a generalisation of the goodness-of-fit test where we determine whether given data are drawn from p.d. p^g or not. The candidate key class $g \in \mathbb{F}_2^m$ which is most strongly indicated by this test to fit the data is chosen to be the right key class. The classical goodness-of-fit tests are the χ^2-test and the G test based on the Kullback–Leibler distance. The first two methods to be presented in this paper are generalisations of these tests into the case of multiple distributions, i.e. finding one distribution from a set of distributions. The χ^2-method based on the χ^2-test and the log-likelihood method based on the G test are studied in Sects. 5.2 and 5.3, respectively.

Our third method is the LLR-method to be studied in Sect. 5.4. In [2], the problem of distinguishing one *known* p.d. from a set of other p.d.'s was studied. It was then possible to use the optimal distinguisher, the LLR-statistic, in solving the problem. However, since g_0 is unknown, we cannot apply the results of [2] in our work directly. Instead, we will use the following heuristic: since the data corresponding to $\hat{q}$ are drawn from the unknown p.d. $p^{g_0} \neq \theta$, it should be easiest to distinguish the right p.d. p^{g_0} rather than any other p.d. p^g, $g \neq g_0$ from the uniform distribution using the LLR-statistic. Hence, the candidate key class $g \in \mathbb{F}_2^m$ that gives the strongest distinguisher between the corresponding p.d p^g and θ is chosen to be the right key. This can be seen as a variant of Assumption 1.

In all our analysis, it is assumed that p^g and $p^{g'}$, $g \neq g'$, are close to each other, and all these distributions p^g are close to θ in the sense of Property 1. The assumption about closeness must be verified with practical experiments.

5.2. χ^2-*Method*

The mark for each candidate class $g \in \mathbb{F}_2^m$ based on χ^2-statistic is defined as follows:

$$s(g) = N \sum_{\eta=0}^{M} \frac{(\hat{q}_\eta - p_\eta^g)^2}{p_\eta^g}, \tag{15}$$

where N is the amount of data used in the attack, with $M = 2^m - 1$ degrees of freedom. The empirical distribution $\hat{q}$ should be near to the correct p.d. p^{g_0} while being further away from all the other p.d.'s p^g, $g \neq g_0$. Hence, the candidate class corresponding to the smallest $s(g)$ is chosen to be the right key class.

By [15], the distribution of $s(g)$ can be approximated by $\chi^2_M\,(NC(p^g, p^{g_0}))$ which can further be approximated by $s(g) \sim \mathcal{N}(\mu_g, \sigma_g^2)$, with mean $\mu_g = M + NC(p^g, p^{g_0})$ and variance $\sigma_g^2 = 2(M + 2NC(p^g, p^{g_0}))$, provided that $\mu_g > 30$ [13], i.e. m should be at least 5.

For simplicity, we have only determined the data complexity of full advantage of $a = m$ bits and assumed that $s(g)$'s are statistically independent. That is, we have determined the data complexity of ranking g_0 on the top of the list with $s(g)$. Let P_S be the probability of finding g_0 such that

$$P_S = \Pr(s(g_0) > \max_{g \neq g_0} s(g)).$$

Using Property (11), we can do calculations and approximations similar to those done in Sect. 4 in [1] or in the proof of Theorem 2 in [20] and get the following estimate of the data complexity N that would be sufficient for finding g_0 with success probability P_S

$$N \approx \frac{4m - 4\beta_S + 2\sqrt{2M(m - \beta_S)}}{C_{\min}(p)}, \tag{16}$$

where $\beta_S = \ln(\sqrt{2\pi} \ln P_S^{-1})$. Note the exponential dependence of N on m as $M = 2^m - 1$. Our experiments showed, however, that much less data are needed than what is

predicted by (16). The reason may be that the statistical dependence of the marks $s(g)$ strengthens the method. However, as noted previously, we could not find a way to do the derivations without the assumption of statistical independence. The formula for the advantage of the χ^2-method could also be derived but we have omitted it here, since it will be given for the LLR-method studied in Sect. 5.4 which performs better in terms of data complexity.

5.3. *The Log-Likelihood Method*

Another popular goodness-of-fit test is the log-likelihood test, also known as G test. The experiments on Alg. 1 done in [20] used this test. The mark based on G test uses the Kullback–Leibler distance

$$G(g) = D(\hat{q} \,||\, p^g)$$

between the empirical p.d. $\hat{q}$ and the theoretical p.d. p^g. In [15], it is shown that for each candidate class $g \in \mathbb{F}_2^m$ the random variable $G(g)$ can be approximated to be distributed as

$$\begin{aligned} G(g) &\sim \chi^2_M(\delta_g) + \xi_g, \text{ where} \\ \delta_g &= N \sum_{\eta=0}^{M} p^g_\eta \log^2 \frac{p^g_\eta}{p^{g_0}_\eta} - N D(p^g \,||\, p^{g_0})^2 \text{ and} \\ \xi_g &= 2N D(p^g \,||\, p^{g_0}) - \delta_g. \end{aligned}$$

Since p^g are near to p^{g_0}, the parameters $\delta_g \approx NC(p^g, p^{g_0})$ and $\xi_g \approx 0$ and the G test performs similarly as the χ^2-test [15].

5.4. *Log-Likelihood Ratio Method*

The log-likelihood ratio is the optimal statistic for distinguishing two distributions [11]. It is also asymptotically normal as stated in Proposition 2.1. Hence, we would like to use it for key ranking. The idea is that the empirical distribution can be used for distinguishing the p.d. p^{g_0} related to the correct key class from the uniform p.d. with large LLR value, while any wrong p.d. p^g, $g \neq g_0$ is less distinguishable from θ. For each $g \in \mathbb{F}_2^m$, we compute the mark

$$\ell(g) = \mathrm{LLR}(\hat{q}, p^g, \theta). \tag{17}$$

We choose the candidate class g with largest $\ell(g)$ to be the right key class.

We cannot apply [2] here to determine the data complexity of finding g_0 as the result would be too optimistic. The task is to distinguish an unknown p^{g_0} from a set of p.d.'s $\{p^g \mid g \in \mathbb{F}_2^m\}$. In [2], one distinguishes only p^{g_0} from $p^{g_0'}$, $g_0' \neq g_0$, the p.d. closest to p^{g_0} in Kullback–Leibler distance. We have to consider all the other candidate classes as well, which increases the data complexity. Applying the theory of key ranking described in Sect. 4, we derive the following result.

Theorem 5.1. *Assume that the random variables $\ell(g)$ are statistically independent and that $\ell(g) \sim \mathcal{N}(\mu_W, \sigma_W^2)$, $g \neq g_0$ where $\mu_W = 0$ and $\sigma_W^2 \approx \sigma_R^2$, where σ_R^2 is*

the variance of the random variable $\ell(g_0)$. If the p.d.'s p^g, $g \in \mathbb{F}_2^m$ and θ are close to each other in the sense of Property 1, the advantage a of the LLR-method with marks calculated by (17) can be approximated by

$$a \approx \left(\frac{1}{2}\sqrt{NC(p)} - \Phi^{-1}(P_S)\right)^2,$$

where $P_S(\geq 0.5)$ is the probability of success, N is the amount of data used in the attack and $C(p)$ and m are the capacity and the dimension of the linear approximation (10), respectively.

The assumption of statistical independence of $\ell(g)$'s was discussed in Sect. 5.1. The assumption about normal distribution of the wrong keys $\ell(g)$, $g \neq g_0$ is based on the law of large numbers [11]. The approximation of variance $\sigma_W^2 = \sigma_R^2$ is commonly used, for example, in [34], and the approximation of mean is based on the idea that the empirical data are not closer to any p^g, $g \neq g_0$ than θ implying that $\mu_W \leq 0$. In the worst case, with largest data complexity, $\mu_W = 0$.

Proof. Let us proceed first by finding the p.d.'s for the random variables $\ell(g)$, $g \in \mathbb{F}_2^m$. By Proposition 2.1 and property (11), random variable $\ell(g_0) \sim \mathcal{N}(N\mu_R, N\sigma_R^2)$, where $\mu_R \approx C(p)/2$ and $\sigma_R^2 \approx C(p)$.

By the assumptions, we may use (13), where $f_W = \phi_{\mu_W,\sigma_W^2}$ and $F_W = \Phi_{\mu_W,\sigma_W^2}$ to obtain

$$\mu_a = \sigma_W b, \text{ where } b = \Phi^{-1}(1 - 2^{-a}) \text{ and}$$
$$\sigma_a^2 \approx \frac{2^{-(m+a)}}{f_W^2(\mu_a)}.$$

By approximation (1),

$$\phi(b) \approx b(1 - \Phi(b)) = b2^{-a}. \tag{18}$$

On the other hand

$$\sigma_a^2 \approx \frac{2^{-(m+a)}\sigma_W^2}{\phi^2(b)}$$

and by (18) we have

$$\frac{\sigma_a^2}{\sigma_W^2} \approx \frac{2^{-(m+a)}}{b^2 2^{-2a}} = \frac{2^{-m+a}}{b^2}.$$

Since $a \leq m$, and $b > 1$, we have showed that $\sigma_a^2/\sigma_W^2 \ll 1$ and also $\sigma_a^2 \ll \sigma_R^2$. Then

$$\begin{aligned} P_S &= \Pr(\ell(g_0) > l_{(r)}) \\ &= \Phi\left(\frac{\mu_R - \mu_a}{\sigma_R}\right) \\ &\approx \Phi\left(\frac{NC(p)/2 - \sqrt{NC(p)}b}{\sqrt{NC(p)}}\right), \end{aligned}$$

Input: plaintext-ciphertext pairs $(\hat{x}_t, \hat{y}_t)$, $t = 1, \ldots, N$
Output: counter array $F(\eta)$ related to empirical p.d. $\hat{q}$ by $F(\eta) = N\hat{q}_\eta$
for $\eta = 0, \ldots, M$ **do**
    $F(\eta) = 0$;
end
for $t = 1, \ldots, N$ **do**
    for $i = 1, \ldots, m$ **do**
        $\eta_i = u_i \cdot \hat{x}_t \oplus w_i \cdot \hat{y}_t$;
    end
    $\eta = \sum_{i=1}^{m} \eta_i 2^{i-1}$; /* interpret vector $(\eta_1, \ldots, \eta_m)$ as an integer η */
    $F(\eta) = F(\eta) + 1$;
end

Fig. 1. Online phase of multidimensional Alg. 1.

from which we can solve an estimate of N as a function of a to be

$$N = \frac{4(\Phi^{-1}(P_S) + b)^2}{C(p)} \approx \frac{4(\Phi^{-1}(P_S) + \sqrt{a})^2}{C(p)}, \tag{19}$$

where the last approximation is obtained using approximation (1). It gives $\phi(b)/b \approx 2^{-a}$ and further

$$a \approx \frac{\log e}{2} b^2 + \log b + \frac{\log 2\pi}{2} \approx b^2. \tag{20}$$

By inversion, we get an estimate of a as a function of N as desired. □

The experimental advantages for the different methods are studied in the next section.

A common choice for P_S is $0.5 \leq P_S \leq 0.99$. Hence, the value of $\Phi^{-1}(P_S)$ is typically a small positive number less than 3, and if $m \geq \Phi^{-1}(P_S)^2$, the numerator of (19) is bounded above by $16m$. This shows that the dependence of the data complexity on the dimension m of the multidimensional linear approximation is linear for the LLR-method, while it is exponential in (16) for the χ^2-method. Since in practice $C(p) \approx C_{\min}(p)$, the comparison of (19) and (16) indicates that the LLR-method is more efficient than the χ^2-method or the log-likelihood method.

5.5. *Algorithms and Complexities*

For comparing the two methods, LLR and χ^2, we are interested in the complexities of the first two phases of Alg. 1 since the sorting and searching phase with fixed advantage a do not depend on the chosen method. The counting phase is done online and all the other phases can be done offline. However, we have not followed this division [34] in our implementation, as we do part of the analysis phase online. We will divide the algorithm into two phases as follows: In the *online phase*, depicted in Fig. 1, we calculate the empirical p.d. $\hat{q}$. The marks $s(g)$ for the χ^2-method and $\ell(g)$ for the LLR-method are then assigned to the keys in the *offline phase*. The offline phases for χ^2-method and LLR-method are depicted in Figs. 2 and 3, respectively.

Input: integer N, counter array $F(0),\ldots,F(M)$,
and theoretical p.d.'s $P(g,\eta)$, $\eta = 0,\ldots,M$, for each key class candidate $g = 0,\ldots,M$
Output: key class g'
for $g = 0,\ldots,M$ **do**
$s(g) = \sum_{\eta=0}^{M} \frac{(F(\eta)/N - P(g,\eta))^2}{P(g,\eta)}$;
end
$g' = \arg\min_{g=0,\ldots,M} s(g)$;

Fig. 2. Offline phase of multidimensional Alg. 1 using χ^2.

Input: counter array $F(0),\ldots,F(M)$,
and theoretical p.d.'s $P(g,\eta)$, $\eta = 0,\ldots,M$, for each key class candidate $g = 0,\ldots,M$
Output: key class g'
for $g = 0,\ldots,M$ **do**
$\ell(g) = \sum_{\eta=0}^{M} F(\eta)\log P(g,\eta)$;
end
$g' = \arg\min_{g=0,\ldots,M} \ell(g)$;

Fig. 3. Offline phase of multidimensional Alg. 1 using LLR.

The data complexities of the online phase for χ^2 and LLR are given by (16) and (19), respectively. The dependence of the data complexity of the χ^2-method on the advantage is similar to the LLR-method. The main difference is in the predicted dependence on m. The time complexity of the online phase is Nm, where N is the data complexity. The memory usage is 2^m, the size of the empirical distribution.

In the offline phase, the time and memory complexities for both methods are 2^m and 2^{2m}, respectively.

5.6. *Experiments on Four-Round Serpent*

Similarly as in previous experiments on multiple linear cryptanalysis, see [10], the Serpent block cipher was used as a test bed. Description of Serpent can be found in [3]. We tested the different methods for multidimensional Alg. 1 described in this paper on four-round Serpent, 4th–7th rounds, by selecting linearly independent one-dimensional base approximations $u_i \cdot x \oplus w \cdot y$, $i = 1,\ldots,m$, to construct a linear approximation of the form (10) with $m = 7$ and $m = 10$. The used masks w and u_i, $i = 1,\ldots,m$, can be found in [20].

We checked the assumption about closeness of the hypothetical distributions p^g and θ and saw that it holds as $|p_\eta^g - p_\eta^{g'}| < \frac{1}{150} p_\eta^g$, for all g, g' and $\eta \in \mathbb{F}_2^m$. We also checked that $C_{\min}(p) \neq 0$ and actually, $C_{\min}(p) \approx C(p)$.

The experiments showed that the empirical advantage when ranking the key classes was exactly the same for all multidimensional methods. Hence, we only depict the LLR-method. In particular, all methods were equally efficient in determining the correct key class. Equations (19) and (16) predict that the LLR-method should be the most efficient: when m increases, the data requirement of χ^2-based tests increases exponentially with m, whereas the increase is linear for the LLR-method. It is possible that the variance

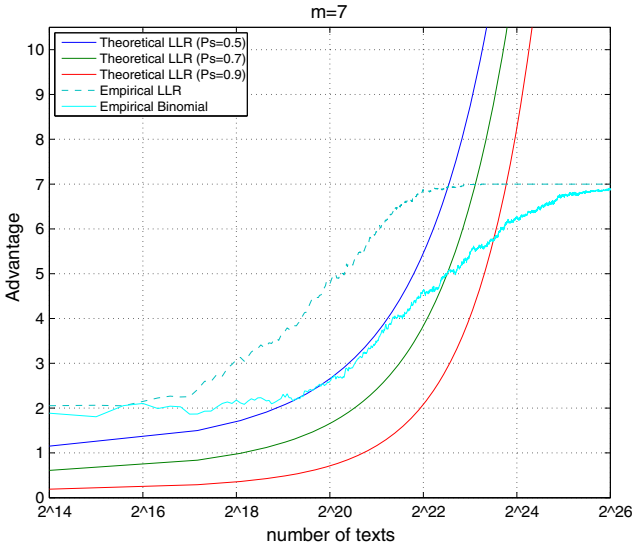

Fig. 4. Alg. 1: Theoretical and empirical advantage as a function of data complexity using LLR-method for four-round Serpent when $m = 7$.

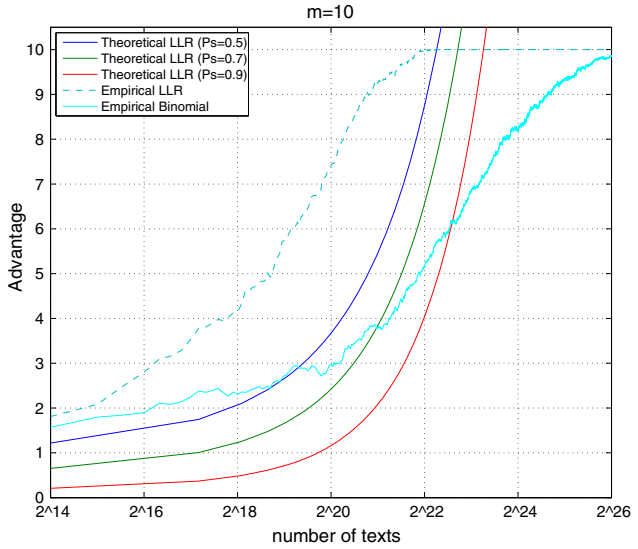

Fig. 5. Alg. 1: Theoretical and empirical advantage as a function of data complexity using LLR-method for four-round Serpent when $m = 10$.

of the χ^2-method is not as large as the theory predicts, or the statistical dependence of random variables $s(g)$ strengthens the χ^2-method more than expected.

The statistical model of the relationship between the advantage a and data complexity N derived in this paper was tested in experiments. The results are given in Figs. 4 and 5. The empirical advantage is determined by averaging the advantages obtained for

16 randomly selected keys using the LLR-method. The theoretical advantage given by Theorem 5.1 is depicted for three different values of P_S. The experiments show that the data complexity predicted this way is too high, but we believe it gives a good upper bound.

As discussed after the proof of Theorem 5.1, for full advantage $a = m$, we can in usual cases approximate $N \approx 16m/C(p)$. By increasing m, that is, using more linear approximations, the data complexity N decreases as long as the ratio $C(p)/m$ increases. This sets an upper bound for m to be used in practice. In case of four-round Serpent, the practical upper bound around $m = 12$ was found in experiments.

In both cases $m = 7$ and $m = 10$, we also show how much better the m-dimensional LLR-method is compared to the binomial method where the same set of m one-dimensional approximations and Matsui's Alg. 1 is used to determine each key class candidate bit separately and independently. The m-bit key classes are then ranked according to the product of $|\hat{c}_i|$, $i = 1, \ldots, m$. This approach is similar to the method of Biryukov et al. The enhanced method of Biryukov et al. includes in the analysis also some of the linear combinations of the base approximations with significant correlations. The experiments in [10,20] for Serpent already confirmed that this is a favourable thing to do in spite of the lack of theoretical justification. In [20], also the enhanced method of Biryukov et al. and the full multidimensional log-likelihood method were compared in experiments on Serpent and the latter was shown to be more powerful.

6. Algorithm 2

Let us study a cipher with $R + 1$ rounds. Let x be the plaintext and z be the cipher text after $R + 1$ rounds. Let the $(R + 1)$th round function and round key be f and $k \in \mathbb{F}_2^l$, respectively. Then the output after R rounds is $y = f^{-1}(z, k)$. Algorithm 2 uses a multidimensional linear approximation over R rounds given by (10) with p.d. p. The task is to find the right last round key k_0 and possibly, in addition, the right key class g_0 for the key used in the first R rounds.

6.1. *Statistical Setting for Alg. 2*

In the counting phase, we draw N data pairs $(\hat{x}_t, \hat{z}_t)$, $t = 1, \ldots, N$. In the analysis phase, for each last round key candidate k, we first calculate $\hat{y}_t^k = f^{-1}(\hat{z}_t, k)$, $t = 1, \ldots, N$. Then, for each k, we calculate the empirical p.d. $\hat{q}^k = (\hat{q}_0^k, \ldots, \hat{q}_M^k)$, where

$$\hat{q}_\eta^k = \frac{1}{N} \#\{t = 1, \ldots, N \mid U\hat{x}_t \oplus W\hat{y}_t^k = \eta\}.$$

The keys are then given mark $T(k)$ by using some statistic T that is calculated from different data $U\hat{x}_t \oplus W\hat{y}_t^k$, $t = 1, \ldots, N$, for each key candidate $k \in \mathbb{F}_2^l$. Hence, the random variables $T(k)$ are statistically independent.

If we use the wrong key $k \neq k_0$ to decrypt the ciphertext, it means we essentially encrypt over one more round and the resulting data will be more uniformly distributed. This heuristics is behind the original wrong key randomisation hypothesis [22,27], which in our case means that the data $U\hat{x}_t \oplus W\hat{y}_t^k$, $t = 1, \ldots, N$, $k \neq k_0$ are drawn i.i.d. from

the uniform distribution. On the other hand, when decrypting with the correct key k_0 the data $U\hat{x}_t \oplus W\hat{y}_t^{k_0}$, $t = 1, \ldots, N$ are drawn i.i.d. from p^{g_0}, where $g_0 \in \mathbb{F}_2^m$ is unknown.

6.2. *Key Ranking in One-dimensional Alg. 2*

Key ranking and advantage in the one-dimensional case, $m = 1$, of Alg. 2 was studied in [33]. We will present it here briefly for completeness. Let $c > 0$ be the correlation of (10) (the calculations are similar if $c < 0$) and let $\hat{c}^k$ be the empirical correlation calculated from the data. The mark used in ranking the keys is then given by $s'(k) = |\hat{c}^k|$.

The random variable $\hat{c}^{k_0}$ is binomially distributed with mean $\mu_R = c$ and variance $\sigma_R^2 = (1-c^2)/N \approx N^{-1}$. The wrong key random variables $\hat{c}^k$, $k \neq k_0$, are binomially distributed with mean $\mu_W = 0$ (following Assumption 1) and variance $\sigma_W^2 \approx \sigma_R^2$. Since N is large, we can approximate $s'(k_0) \sim \mathcal{N}(\mu_R, \sigma_R^2)$ and $s'(k) \sim \mathcal{FN}(\mu_W, \sigma_W^2)$, $k \neq k_0$, where $\mathcal{FN}$ is the folded normal distribution; see Appendix A in [33]. Then for finding g_0 with given success probability P_S and advantage a the data complexity N is estimated as follows

$$N \approx \frac{(\Phi^{-1}(P_S) + \Phi^{-1}(1 - 2^{-a-1}))^2}{c^2}.$$

6.3. *Different Scenarios in Multiple Dimensions*

When considering generalisation of Alg. 2 to the case, where multidimensional linear approximation (10) is used, basically two different standard statistical methods can be used:

- Goodness of fit (usually based on χ^2-statistic, see also [29,34]) and
- Distinguishing of an unknown p.d. from a given set of p.d.'s (usually based on LLR-statistic)

The goodness-of-fit approach is a straightforward generalisation of one-dimensional Alg. 2. It can be used in searching for the last round key. It measures whether the data are drawn from the uniform distribution, or not, by measuring the deviation from the uniform distribution. It ranks highest the key candidate whose empirical distribution is farthest away from the uniform distribution. The method does not depend on the key class candidate for the key used before the last round. Information about p.d. p is required only for measuring the strength of the test. We will study this method in Sect. 6.4. After the right last round key k_0 is found, the data derived by Alg. 2 can then be used in any form of Alg. 1 for finding the right key class g_0 of the key used in the R first rounds. In this manner, the χ^2-approach allows separating between Alg. 1 and Alg. 2.

The idea behind the goodness-of-fit approach is that the "wrong key" distributions $\hat{q}^k, k \neq k_0$ are close to the uniform distribution, whereas $\hat{q}^{k_0}$ is further away from it. Moreover, the expected value of $C(\hat{q}^{k_0}, \theta) = C(\hat{q}^{k_0})$ should be approximately $C(p)$.

The LLR-method uses the information about the p.d. related to the key class candidate g also in Alg. 2. In this sense, it is similar to the method presented in [4], where both g_0 and k_0 were searched at the same time. As we noted in Sect. 2.1, the LLR-statistic is the optimal distinguisher between two known p.d.'s. If we knew the right key class

g_0, we could simply use the empirical p.d.'s $\hat{q}^k$ for distinguishing p^{g_0} and the uniform distribution and then choose the last round candidate k for which this distinguisher is strongest [1]. In practice, the right key class g_0 is unknown when running Alg. 2 for finding the last round key.

Our approach is the following. In [2], it was described how one can use LLR to distinguish one known p.d. from a set of p.d.'s. We will use this distinguisher for distinguishing θ from the given set $p^g,\ g \in \mathbb{F}_2^m$. In the setting of Alg. 2, we can expect that for the right k_0, it should be possible to clearly conclude that the data $(\hat{x}_t, \hat{z}_t),\ t = 1, \ldots, N$, yield data $U\hat{x}_t \oplus W\hat{y}_t^{k_0},\ t = 1, \ldots, N$, which follow a p.d. p^g, for some $g \in \mathbb{F}_2^m$, rather than the uniform distribution. On the other hand, for the wrong $k \neq k_0$, the data follow the uniform distribution, rather than any $p^g,\ g \in \mathbb{F}_2^m$.

To distinguish k_0 from the wrong key candidates, we determine, for each round key candidate k, the key class candidate g, for which the LLR-statistic is the largest with the given data. The right key k_0 is expected to have g_0 such that the LLR-statistic with this pair (k_0, g_0) is larger than for any other pair $(k, g) \neq (k_0, g_0)$. In this manner, we also recover g_0 in addition to k_0. The LLR-method is studied in Sect. 6.5.

6.4. *The χ^2-Method*

This method separates Alg. 1. and Alg. 2 such that the latter does not need any information of p. Both methods are interpreted as goodness-of-fit problems, for which the natural choice of ranking statistic is χ^2. We will show first how to find the last round key k with Alg. 2.

6.4.1. *Algorithm 2 with χ^2*

Given empirical p.d. $\hat{q}^k$, the mark based on χ^2-statistic is calculated from the data by

$$S(k) = 2^m N \sum_{\eta=0}^{M} (\hat{q}_\eta^k - 2^{-m})^2, \tag{21}$$

where $M = 2^m - 1$ is the number of degrees of freedom. Formula (21) can be interpreted as the ℓ_2-distance between the empirical p.d. and the uniform p.d.. By Assumption 1, the right round key should produce data that are farthest away from the uniform p.d. and the last round key candidate k for which the mark $S(k)$ is largest is chosen. Obviously, (21) simplifies to $(\hat{c}^k)^2$, if $m = 1$.

The following theorem describes the relationship between the data complexity and the time complexity of the search phase.

Theorem 6.1. *Suppose the cipher satisfies Assumption 1 where $q' = \theta$ and the p.d.'s $p^g,\ g \in \mathbb{F}_2^m$ and θ are close to each other. Then the advantage of the χ^2-method using marks (21) is given by*

$$a \approx \frac{\left(NC(p) - 2\sqrt{M}\Phi^{-1}(P_S)\right)^2}{2M}, \tag{22}$$

where $M = 2^m - 1$, P_S (> 0.5) is the probability of success, N is the amount of data used in the attack and $C(p)$ and m (≥ 5) are the capacity and the dimension of the linear approximation (10)*, respectively.*

Proof. According to [15], the random variable $S(k_0)$ is distributed approximately as

$$S(k_0) \sim \chi^2_M(NC(p^{g_0})) = \chi^2_M(NC(p)),$$

because of the symmetry property (11). Since $m \geq 5$, we may approximate the distribution by a normal distribution with $\mu_R = M + NC(p)$ and $\sigma_R^2 = 2(M + 2NC(p))$. The parameters do not depend on g_0 or k_0. For the wrong keys $k \neq k_0$, we obtain by [15] that

$$S(k) \sim \chi^2_M(0) = \chi^2_M,$$

so that $\mu_W = M$ and $\sigma_W^2 = 2M$. The mean and variance in (13) are

$$\mu_a = b\sqrt{2M} + M$$
$$\sigma_a^2 = \frac{2M}{2^{l+a}\phi(b)^2},$$

where $b = \Phi^{-1}(1 - 2^{-a})$.

When $a < l$ is large, we use approximation (20) to get $a \approx b^2$ by which $\mu_a = \sqrt{2aM} + M$. Further by approximation (1), we get $\phi(b) \approx b2^{-a}$. Then

$$\sigma_a^2 \approx \frac{2M}{2^{l+a}2^{-2a}b^2} \approx \frac{M}{a2^{l-a}} < M.$$

For small a, this estimate holds trivially. Then to proceed, we restrict to the case $NC(p) < M/4$. This is not essential restriction, since finally $NC(p)$ will be close to a small constant multiple of $\sqrt{M}$. Then we have

$$\sqrt{2M} < \sqrt{\sigma_a^2 + \sigma_R^2} < 2\sqrt{M},$$

and we use the upper bound as an estimate for $\sqrt{\sigma_a^2 + \sigma_R^2}$. Then we can solve data complexity from the formula (14) of the success probability to obtain the following estimate of the data complexity

$$N \approx \frac{\left(\sqrt{2a} + 2\Phi^{-1}(P_S)\right)\sqrt{M}}{C(p)}. \tag{23}$$

By solving for a, we get the estimated advantage as claimed. □

Note that the normal approximation of the wrong key distribution is valid only when $m > 5$, that is, when the approximation of χ^2-distribution by a normal distribution is

valid. It is not possible to perform the theoretical calculations for small m as the χ^2-distribution does not have a simple asymptotic form in that case and we cannot determine f_W and F_W in (13). Since for $m = 1$ our χ^2-method reduces to the square of $s(k)$ that was used by Selçuk, the theoretical distributions differ from our calculations and we get a slightly different formula for the advantage. Despite this difference, the methods are essentially equivalent for $m = 1$.

Keeping the capacity and advantage constant, we see by (23) that the data complexity increases exponentially as $2^{m/2}$ as the dimension m of the linear approximation increases and is sufficiently large. Hence, in order to strengthen the attack by increasing m, the capacity should increase faster than $2^{m/2}$. This is a very strong condition and it suggests that in applications, only approximations with small m should be used with χ^2-attack. The experimental results for different m presented in Sect. 6.7 as well as the theoretical curves depicted in Fig. 9 suggest that increasing m in the χ^2-method does not necessarily mean improved performance for Alg. 2.

While (22) and (23) depend on the theoretical distribution p, the actual χ^2-mark (21) is independent of p. Therefore, to realise the attack, we do not need to know p accurately. It suffices to find an approximation (10) that deviates as much as possible from the uniform distribution. On the other hand, if we use time and effort for computing an approximation of the theoretical p.d. p and if we may assume that the approximation is accurate, we would also like to exploit this knowledge for finding the right class of the key used in R first rounds using Alg. 1. As noted in Sect. 5, there are several ways to realising a multidimensional Alg. 1. Considering Alg. 1 as a χ^2-based goodness-of-fit problem, we may combine Alg. 1 and Alg. 2 to be described next.

6.4.2. *Combined Method and Discussion*

The sums of squares of correlations used in the method of Biryukov et al. [4] are closely related to the sums of squares (15) and (21). Indeed, we could define a combined χ^2-mark S' by considering the sum of (15) and (21) and setting

$$S'(k, g) = \sum_{k' \neq k} S(k') + s(k, g),$$

where $s(k, g)$ is calculated from the empirical p.d. $\hat{q}^k$, $k \in \mathbb{F}_2^l$ by (15). The right key (k_0, g_0) should give the smallest mark. If we approximate the denominators in (15) by 2^{-m} and scale by $2^m N$, we obtain from $S'(k, g)$ the mark

$$B(k, g) = \sum_{k' \neq k} ||\hat{q}^{k'} - \theta||_2^2 + ||\hat{q}^k - p^g||_2^2, \tag{24}$$

which is closely related to the one used in [4]:

$$\sum_{k' \neq k} ||\hat{\mathbf{c}}^{k'}||_2^2 + ||\hat{\mathbf{c}}^k - \mathbf{c}^g||_2^2, \tag{25}$$

where $\hat{\mathbf{c}}^k$ is the empirical correlation vector corresponding to the last round key candidate k and $\mathbf{c}^g$ is the theoretical correlation vector corresponding to the key class candidate g.

Indeed, if in (25) all correlation vectors $\hat{\mathbf{c}}^k$ and $\mathbf{c}^g$ contain correlations from all linear approximations then (25) becomes the same as $2^m B(k, g)$ as can be seen by Proposition 2.2 and Parseval's theorem. Initially, in the theoretical derivation of (25) only linearly and statistically independent approximations were included in the correlation vectors. However, in Sect. 3.4 of [4] it was proposed to take into account all linear approximations with strong correlations when forming the mark (25) in practice. In practical experiments by Collard et al. [10], this heuristic enhancement was demonstrated to improve the results for Serpent. In this paper, we have shown how to remove the assumption about independence of the linear approximations and offer the option of including all linear approximations that have sufficient contribution to the capacity.

Other possibilities for combining Alg. 1 and Alg. 2 based on χ^2 or its variants are also possible, with different weights on the terms of the sum in (24), for instance. However, the mathematically more straightforward way is to use the pure χ^2-method defined by (21) and (15), as its statistical behaviour is well known. An even more efficient method can be developed based on LLR as will be shown next.

6.5. *The LLR-Method*

This method is also based on the same heuristic as the wrong key hypothesis: For $k \neq k_0$, the distribution of the data should look uniform and for k_0 it should look like p^{g_0}, for some g_0. Hence, for each k, the problem is to distinguish the uniform distribution from the discrete and known set $p^g,\ g \in \mathbb{F}_2^m$. Let us use the notation $L(k, g) = \mathrm{LLR}(\hat{q}^k, p^g, \theta)$. We propose to use the following mark

$$L(k) = \max_{g \in \mathbb{F}_2^m} L(k, g). \tag{26}$$

Now k_0 should be the key for which this maximum over g is the largest, and ideally, then the maximum is achieved for $g = g_0$. While the symmetry property (11) allows one to develop statistical theory without knowing g_0, in practice one must search through $\mathbb{F}_2^l$ for k_0 and $\mathbb{F}_2^m$ for g_0 even if we are only interested in determining k_0.

Theorem 6.2 gives the trade-off between the time complexity of the search phase and the data complexity of the algorithm.

Theorem 6.2. *Suppose the cipher satisfies Assumption 1 where $q' = \theta$ and the p.d.'s $p^g,\ g \in \mathbb{F}_2^m$ and θ are close to each other. Suppose further that for all the wrong last round key candidates $k \neq k_0$, the 2^m random variables $L(k, g),\ g \in \mathbb{F}_2^m$ are statistically independent. Then the advantage of the LLR-method for finding the last round key k_0, assuming that it is paired with the right key class g_0, is given by*

$$a \approx (\sqrt{NC(p)} - \Phi^{-1}(P_{12}))^2 - m \approx NC(p) - m. \tag{27}$$

Here N is the amount of data used in the attack, P_{12} (> 0.5) is the probability of success and $C(p)$ and m are the capacity and the dimension of the linear approximation (10), respectively.

For fixed $k \neq k_0$, the random variables $L(k, g)$, $g \in \mathbb{F}_2^m$ are calculated using the same data $\hat{q}^k$. Hence, they are actually statistically dependent. Similarly as in Sect. 5, we will assume that they are statistically independent to simplify calculations. The practical experiments concentrated on finding the right last round key k_0 such that we did not actually need the knowledge of g_0 or its data complexity. The following calculations give a theoretical model that can be used in describing how the LLR-method behaves especially when compared to other methods.

Proof. Using Proposition 2.1 and property (11), we can state Assumption 1 as follows: For the right key (k_0, g_0),

$$L(k_0, g_0) \sim \mathcal{N}(\mu_R, \sigma_R^2), \text{ where } \mu_R = \frac{1}{2}NC(p) \text{ and } \sigma_R^2 = NC(p), \tag{28}$$

and for $k \neq k_0$ and any $g \in \mathbb{F}_2^m$,

$$L(k, g) \sim \mathcal{N}(\mu_W, \sigma_W^2), \text{ where } \mu_W = -\frac{1}{2}NC(p) \text{ and } \sigma_W^2 = NC(p). \tag{29}$$

Hence, μ_R, σ_R^2, μ_W and σ_W^2 do not depend on $g \in \mathbb{F}_2^m$.

For fixed $k \neq k_0$, the random variables $L(k, g)$ for $k \neq k_0$ are identically normally distributed with mean μ_W and variance σ_W^2. Assuming that for each $k \neq k_0$, the random variables $L(k, g)$'s are independent, we obtain that the c.d.f. of their maximum is given by [14]

$$F_W(x) = \Phi_{\mu_W, \sigma_W^2}(x)^{M+1}$$

and p.d.f. is

$$f_w(x) = (M+1)\Phi_{\mu_W, \sigma_W^2}(x)^M \phi_{\mu_W, \sigma_W^2}(x).$$

Let us fix the advantage a such that $r = 2^{l-a}$. The mean μ_a of the rth wrong key statistic $L_{(r)}$ can now be calculated from (13) to be

$$\begin{aligned} \mu_a &= \mu_W + \sigma_W b = -1/2NC(p) + \sqrt{NC(p)}b, \\ b &= \Phi^{-1}(\sqrt[M+1]{1 - 2^{-a}}), \end{aligned} \tag{30}$$

and the variance is

$$\sigma_a^2 = \frac{2^{-l-a}\sigma_W^2}{(M+1)^2(1-2^{-a})^{2(1-1/(M+1))}\phi^2(b)} \ll \sigma_R^2. \tag{31}$$

Let

$$P_{12} = \Pr(L(k_0) > L_{(r)} \mid L(k_0, g_0) > \max_{g \neq g_0} L(k_0, g))$$

be the probability that we rank k_0 among the r highest ranking keys provided that we pair g_0 with k_0.

We can calculate P_{12} using (14), (30) and (31) to obtain

$$P_{12} = \Phi\left(\frac{\mu_R - \mu_W - \sigma_w b}{\sigma_R}\right) \approx \Phi(\sqrt{NC(p)} - b).$$

Hence, the amount of data

$$N \approx \left(\Phi^{-1}(P_{12}) + b\right)^2 / C(p) \tag{32}$$

is estimated to be sufficient. We approximate $\Phi(b) = \sqrt[M+1]{1-2^{-a}} \approx 1 - 2^{-m-a}$. Then by replacing a by $a + m$ in (20) we have $a + m \approx b^2$ and can solve an estimate of advantage a as a function of N from (32) to get the claim. □

Let

$$P_1 = \Pr(L(k_0, g_0) > \max_{g \neq g_0} L(k_0, g))$$

be the probability that given k_0, we choose g_0, i.e. the probability of success of Alg. 1. with full advantage $a = m$. Let

$$P_2 = \Pr(L(k_0) > L_{(r)})$$

be the probability that we rank k_0, paired with *any* $g \in \mathbb{F}_2^m$, among the r highest ranking keys. Then

$$\begin{aligned} P_2 &= P_{12}P_1 + \Pr(L(k_0) > L_{(r)} \mid L(k_0) = \mathrm{LLR}(k_0, p^g, \theta), g \neq g_0)(1 - P_1) \\ &\geq P_{12}P_1. \end{aligned}$$

Denote by N_1 and N_2 the data complexities needed to achieve success probabilities P_1, and P_2, respectively. The data complexity N_1 is given in (19). If we pair k_0 with $g \neq g_0$, then $L(k_0) \geq L(k_0, g_0)$ for a fixed empirical p.d. $\hat{q}^{k_0}$, so that k_0 gets ranked *higher* than by using the correct g_0. Hence, assuming that k_0 gets paired with g_0 only decreases P_2 so the theory predicts N_2 to be (slightly) larger than in reality. Then we can approximate N_2 from above by (32) and the corresponding advantage with success probability P_2 is approximately given by (27).

The data complexity N_1 is an overestimate for the actual data complexity of Alg. 1 so in practice, $N_2 > N_1$. Then if k_0 is ranked with advantage a and success probability $P_2 > 0.5$ among the 2^{l-a} highest ranking keys, it is also paired with the right key class g_0. We have the following corollary:

Corollary 6.1. *Under the assumptions of Theorem 6.2, the data complexity of the LLR-method for ranking the last round key k_0 among the $r = 2^{l-a}$ highest ranking keys can be estimated as*

$$N = \max(N_1, N_2) \approx \frac{a + m}{C(p)}, \tag{33}$$

Input: plaintext-ciphertext pairs $(\hat{x}_t, \hat{z}_t)$, $t = 1, \ldots, N$
Output: array $F(k, \eta)$ related to empirical p.d. $\hat{q}^k$ by $F(k, \eta) = N\hat{q}^k(\eta)$, $k = 0, \ldots, 2^l$, $\eta = 0, \ldots, M$
for $k = 0, \ldots, 2^l - 1$, $\eta = 0, \ldots, M$ **do**
    $F(k, \eta) = 0$;
end
for $t = 1, \ldots, N$ **do**
    for $k = 0, \ldots, 2^l - 1$ **do**
        $\hat{y}_t^k = f^{-1}(\hat{z}_t, k)$; /* decrypt the ciphertext partially */
        for $i = 1, \ldots, m$ **do**
            $\eta_i = u_i \cdot \hat{x}_t \oplus w_i \cdot \hat{y}_t^k$;
        end
        $\eta = \sum_{i=1}^{m} \eta_i 2^{i-1}$; /* interpret vector $(\eta_1, \ldots, \eta_m)$ as an integer η */
        $F(k, \eta) = F(k, \eta) + 1$;
    end
end

Fig. 6. Online phase of multidimensional Alg. 2.

Input: integer N, array $F(k, \eta)$, $k = 0, \ldots, 2^l - 1$, $\eta = 0, \ldots, M$
Output: marks $S(k)$, $k = 0, \ldots, 2^l - 1$
for $k = 0, \ldots, 2^l - 1$ **do**
    $S(k) = \sum_{\eta=0}^{M} (F(k, \eta)/N - 2^{-m})^2$;
end

Fig. 7. Offline phase of multidimensional Alg. 2 using χ^2-method.

and with this data complexity k_0 is paired with the right key class g_0 with a high success probability. On the other hand, given an amount N of data, the advantage of the LLR-method is

$$a \approx (\sqrt{NC(p)} - \Phi^{-1}(P_2))^2 - m \approx NC(p) - m, \tag{34}$$

where P_2 (> 0.5) is the probability of success and $C(p)$ and m are the capacity and the dimensions of the linear approximation (10)*, respectively.*

With fixed N and capacity $C(p)$, the advantage decreases linearly with m, whereas in (22), the logarithm of advantage decreases linearly with m. For fixed m and p, the advantage of the LLR-method seems to be larger than the advantage of the χ^2-method. The experimental comparison of the methods is presented Sect. 6.7.

6.6. *Algorithms and Complexities*

Similarly as in Sect. 5.5, we divide Alg. 2 into online and offline phases. In the online phase, depicted in Fig. 6 we calculate the empirical p.d.'s for the round key candidates. The marks $S(k)$ for the χ^2-method and $L(k)$ for the LLR-method are then assigned to the keys in the offline phase. The offline phases for χ^2-method and LLR-method are depicted in Figs. 7 and 8, respectively. After the keys k are each given the mark, they can be ranked according to the mark. Given $\hat{q}^{k_0}$, the multidimensional Alg. 1 can be

Input: integer N, array $F(k,\eta)$, $k=0,\ldots,2^l-1$, $\eta=0,\ldots,M$,
and for each key class candidate $g=0,\ldots,M$ the theoretical p.d.'s $P(g,\eta)$, $\eta=0,\ldots,M$
Output: marks $L(k)$, $k=0,\ldots,2^l-1$
for $k=0,\ldots,2^l-1$ **do**
  for $g=0,\ldots,M$ **do**
    $L'(g)=\mathrm{LLR}(\hat{q}^k,p^g,\theta)$, where $\hat{q}^k_\eta=F(k,\eta)/N$;
  end
  $L(k)=\max_{g\in 0,1,\ldots,M} L'(g)$;
end

Fig. 8. Offline phase of multidimensional Alg. 2 using LLR-method.

used for finding g_0 offline. The method used in Alg. 1 does not depend on the one used in Alg. 2, so, for example, the LLR ranking, see Fig. 3, can be used for finding g_0.

For fixed advantage a, the data complexities of the online phase for the χ^2 and LLR are given in (23) and (33), respectively. Theoretical and practical results imply that the data complexities of different methods in Alg. 2 dominate the data complexity of Alg. 1 given in (19). Hence, the total data complexities are given by (23) and (33), even for deriving both k_0 and g_0. The memory and time complexities for online phase are 2^{l+m} and $2^l mN$, where N is the data complexity.

For the offline phase of LLR-based Alg. 2, the time and memory complexities are 2^{l+m} and $2^m \max(2^l, 2^m)$, respectively. The method obtained by a combination of χ^2-based Alg. 2 and LLR-based Alg. 1 has the same complexities. Thus, we recommend using the LLR-method rather than χ^2-method unless the accuracy of the p.d. p of the linear approximation (10) is a concern. If we use χ^2 for finding only the last round key k_0, we have the same time complexity as for LLR but a reduced memory complexity 2^l, since we do not have to store the theoretical distributions p^g. In some situations, it may even be advantageous to combine different methods. For example, we may first find, say, r best last round keys by χ^2-based Alg. 2, such that the data complexity is given by (23), with advantage $a = l - \log r$. Then we can proceed by applying Alg. 2 based on LLR-method in a reduced search space of size $r < 2^l$. Other similar variants are possible. Their usefulness depends on the cipher that is being analysed.

6.7. *Experiments on Five-Round Serpent for Alg. 2*

The purpose of the experiments was to test the accuracy of the derived statistical models and to demonstrate the better performance of the LLR-based method in practice. We take 5 rounds of Serpent, from the 4th to 8th round, m-dimensional linear approximation over four rounds, 4th to 7th, and searched for a 12-bit part of the round key used in the 8th round. Different m were used. Each experiment was performed for 16 different, randomly selected keys. Given a data sample of size N, we measured the experimental ranking of k_0 among all 2^{12} candidates and computed the corresponding advantage. In the figures, the average advantage taken over 16 keys is depicted for each sample sizes N that are integer multiples of 2^{21}. Comparing the results to the experiments done with Alg. 1 in Sect. 5.6, it is noted that for large advantages the data complexities of Alg. 2 dominate the ones of Alg. 1.

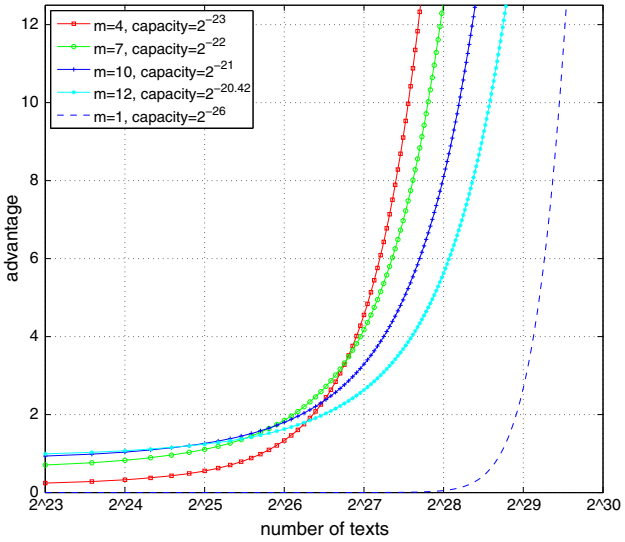

Fig. 9. Alg. 2: Theoretical advantage for χ^2-method for different m and $P_S = 0.95$.

We calculated the capacities for the approximation (10) over four-round Serpent for different m. The results in Sect. 5.6 showed that p^g's can be considered to be close to each other and θ.

The theoretical advantage of the χ^2-method predicted in (22) has been plotted as a function of data complexity in Fig. 9. The figure shows that increasing m larger than 4, the attack is weakened. This suggests using $m = 4$ base approximations in the χ^2-attack. Since we should have m at least 5 for the normal approximation of χ^2_M to hold, the theoretical calculations do not necessarily hold for small m. However, the experiments, presented in Fig. 10, are in accordance with the theory also for $m = 1$ and $m = 4$. The most efficient attack is obtained by using $m = 4$ equations. Increasing m (and hence the time and memory complexities of the attack) actually weakens the attack. The optimal choice of m depends on the cipher and the size of the correlations of available linear approximations.

The reason is the χ^2-statistic itself: it only measures if the data follow a certain distribution, the uniform distribution in this case. The more the approximations we use, the larger the distributions become and the more the uncertainty we have about the "fitting" of the data. Small errors in experiments generate large errors in χ^2 as the fluctuations from the relative frequency 2^{-m} become more significant.

The theoretical advantage of the LLR-method (34) is plotted against the data complexity in Fig. 11 for different m. The empirical advantages for several different m are shown in Fig. 12 in the same experimental setting as was used for χ^2. Unlike for χ^2, we see that the method can be strengthened by increasing m, until the increase in the capacity $C(p)$ becomes negligible compared to increase in m. For four-round Serpent, this happens when $m \approx 12$. Experimental results presented in Figs. 10 an 12 confirm the theoretical prediction that the LLR-method is more powerful than the χ^2-method.

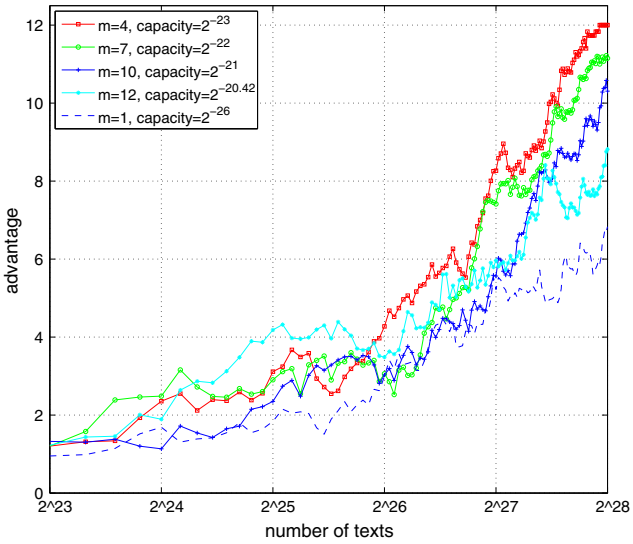

Fig. 10. Alg. 2: Empirical advantage for χ^2-method for different m.

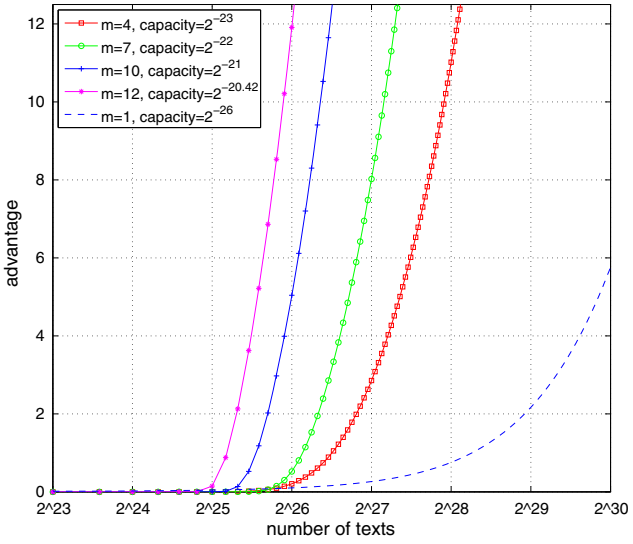

Fig. 11. Alg. 2: Theoretical advantage for LLR-method for different m and $P_{12} = 0.95$.

Also the theoretical and empirical curves agree nicely. For example, the full advantage of 12 bits with $m = 7$ is achieved at $\log N = 27.5$ for LLR, whereas for χ^2-method $\log N = 28$. Moreover, the LLR can be strengthened by increasing m. For $m = 12$, the empirical and theoretical data complexities are close to 2^{26}.

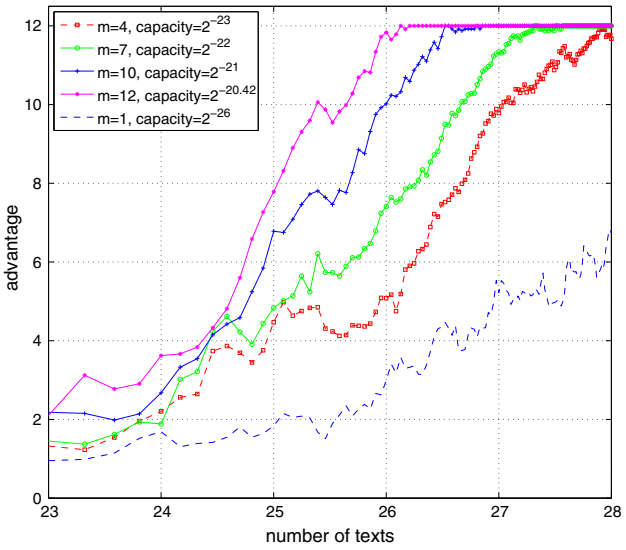

Fig. 12. Alg. 2: Empirical advantage for LLR-method for different m.

7. Conclusions

We studied several ways of extending Matsui's Algorithms 1 and 2 to multiple dimensions. Using the advantage, we could compare the efficiency of the different methods in theory and in practice. The theory predicted that for both Alg. 1 and Alg. 2 the key ranking based on LLR is more efficient than goodness-of-fit tests using χ^2 (or G test). However, in practical experiments for Alg. 1, the methods seemed to perform equally well. It remains an open question how to avoid the assumption about statistical independence of the ranking values and make the theoretical prediction more accurate for Alg. 1.

For Alg. 2, both theoretical and practical results are in agreement. We showed that the χ^2-based method is weaker than the LLR-based method. Hence, we recommend to use the LLR-method proposed in this paper rather than the χ^2-method as long as the estimated p.d. of the multidimensional linear approximation can be assumed to be accurate for all keys. If the theoretical p.d. varies a lot with keys or otherwise cannot be determined accurately, then the χ^2-method is preferred.

8. Later Developments

Let us conclude this revised version of the paper by giving a brief overview of the most important advances in multidimensional linear cryptanalysis during the decade that passed since the submission of this paper. At CT-RSA 2010, the authors presented an improved method for Alg. 1 called as the convolution method [19] and an attack on block cipher PRESENT [9]. The dependence of p.d. p of the key we observed in this paper has been considered, and more advanced statistical models have been developed. The variance due to random key was first integrated in the wrong key model

of one-dimensional linear cryptanalysis [8], then to the model of the cipher [21] and finally for both [6] considering also explicitly the case of sampling without replacement. Since 2012, the hypothesis testing model has been adopted in many works on statistical cryptanalysis to determine distinguishing complexity [7] and data complexity for key recovery. Instead of relating the advantage a to ranking of the candidate keys, the quantity 2^{-a} is interpreted as the probability of the error of accepting a wrong candidate key. With the hypothesis testing model, the problem with statistically dependent order marks we had in this paper disappears, while the data complexity estimates remain essentially the same. Zero-correlation multidimensional linear cryptanalysis was invented in [7], where it was also shown to be linked with integral attacks. The corresponding mathematical link between general multidimensional linear cryptanalysis and truncated differential cryptanalysis was proved in [5]. It allows to transform differential-type attacks to linear-type attacks and vice versa. For example, the multidimensional linear attack presented in [9] gives also the best differential-type attack on this cipher. Finally, let us mention that the recently presented affine multidimensional linear cryptanalysis allows to remove all trivial linear approximations, e.g. where either the input mask or the output mask is equal to zero without introducing artificial independence assumptions [30]. The $2^m - 1$ linear approximations over four rounds of SERPENT we used in this paper share the same output mask. Using the affine method, the trivial linear approximations (linear combinations of even number of base approximations) can be removed. The set of the remaining 2^{m-1} linear approximations has the same capacity as the original set, but the variance of the related χ^2-statistic is significantly reduced, thus potentially improving the power of the attack.

Acknowledgements

Open access funding provided by Aalto University. The research for this paper was done in 2008–2009 when all three authors were affiliated with the Department of Information and Computer Science of Helsinki University of Technology, Finland. The work of the first two authors was funded by the Academy of Finland, Project #122736. The authors wish to thank the anonymous reviewers of Journal of Cryptology for their invaluable comments that were helpful for improving the presentation of this paper.

References

[1] T. Baignères, P. Junod, S. Vaudenay, How far can we go beyond linear cryptanalysis? in P.J. Lee, editor, *Advances in Cryptology—ASIACRYPT'04*, LNCS, vol. 3329 (Springer, Berlin, 2004), pp. 432–450

[2] T. Baignères, S. Vaudenay, The complexity of distinguishing distributions (invited talk), in R. Safavi-Naini, editor, *Information Theoretic Security*. LNCS, vol. 5155 (Springer, Berlin, 2008), pp. 210–222

[3] E. Biham, R. Anderson, L. Knudsen, Serpent: a new block cipher proposal, in S. Vaudenay, editor, *Fast Software Encryption*. LNCS, vol. 1372 (Springer, Berlin, 1998), pp. 222–238

[4] A. Biryukov, C. De Cannière, M. Quisquater, On multiple linear approximations, in M. Franklin, editor, *Advances in Cryptology—CRYPTO'04*. LNCS, vol. 3152 (Springer, Berlin, 2004), pp. 1–22
[5] C. Blondeau, K. Nyberg, Links between truncated differential and multidimensional linear properties of block ciphers and underlying attack complexities, in P.Q. Nguyen, E. Oswald, editors., *Advances in Cryptology—EUROCRYPT 2014*. LNCS, vol. 8441 (Springer, Berlin 2014), pp. 165–182
[6] C. Blondeau, K. Nyberg, Joint data and key distribution of simple, multiple, and multidimensional linear cryptanalysis test statistic and its impact to data complexity. *Des. Codes Cryptogr.*, 82(1–2):319–349, 2017.
[7] A. Bogdanov, G. Leander, K. Nyberg, M. Wang, Integral and multidimensional linear distinguishers with correlation zero, in X. Wang, K. Sako, editors, *Advances in Cryptology—ASIACRYPT 2012*. LNCS, vol. 7658 (Springer, Berlin, 2012), pp. 244–261
[8] A. Bogdanov, E. Tischhauser, On the wrong key randomisation and key equivalence hypotheses in Matsui's Algorithm 2, in S. Moriai, editor, *Fast Software Encryption—20th International Workshop, FSE 2013*, LNCS, vol. 8424 (Springer, Berlin, 2013), pp. 19–38
[9] J.Y. Cho. Linear cryptanalysis of reduced-round PRESENT, in Pieprzyk [31] (pp. 302–317).
[10] B. Collard, F.-X. Standaert, J.-J. Quisquater, Experiments on the multiple linear cryptanalysis of reduced round Serpent, in K. Nyberg, editor, *Fast Software Encryption*. LNCS, vol. 5086 (Springer, Berlin, 2008), pp. 382–397.
[11] T.M. Cover, J.A. Thomas, *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing, 2nd edn (Wiley-Interscience, 2006).
[12] H. Cramér and H. Wold. Some theorems on distribution functions. *J. Lond. Math. Soc.*, s1–11(4):290–295, 1936.
[13] H. Cramér. *Mathematical Methods of Statistics*. Princeton Mathematical Series, 7th edn (Princeton University Press, 1957).
[14] H.A. David, *Order Statistics*. A Wiley Publication in Applied Statistics. 1 edn, (Wiley, New York, 1970).
[15] F.C. Drost, W.C.M. Kallenberg, D.S.Moore, J.Oosterhoff, Power approximations to multinomial tests of fit. *J. the Am. Stat. Assoc.*, 84(405):130–141 (1989).
[16] H. Englund, A. Maximov, Attack the Dragon, in S. Maitra, C.E. Veni Madhavan, editors, *Progress in Cryptology—INDOCRYPT'05*. LNCS, vol. 3797 (Springer, Berlin, 2005), pp. 130–142
[17] C. Harpes, G.G. Kramer, J.L. Massey, A generalization of linear cryptanalysis and the applicability of Matsui's Piling-up lemma, in L.C. Guillou, J.-J. Quisquater, editors, *Advances in Cryptology—EUROCRYPT'95*, LNCS, vol. 921 (Springer, Berlin, 1995), pp. 24–38
[18] M. Hermelin, K. Nyberg, Multidimensional linear distinguishing attacks and Boolean functions, in *Fourth International Workshop on Boolean Functions: Cryptography and Applications* (2008).
[19] M. Hermelin, K. Nyberg, Dependent linear approximations: the algorithm of Biryukov and others revisited, in Pieprzyk [31], pp. 318–333.
[20] M. Hermelin, K. Nyberg, J.Y. Cho, Multidimensional linear cryptanalysis of reduced round Serpent. in J. Seberry Y. Mu, W. Susilo, editor, *Information Security and Privacy*, LNCS, vol. 5107 (Springer, Berlin, 2008), pp. 203–215
[21] J. Huang, S. Vaudenay, X. Lai, K. Nyberg, Capacity and data complexity in multidimensional linear attack, in R. Gennaro, M. Robshaw, editors, *Advances in Cryptology—CRYPTO 2015—Part I*. LNCS, vol. 9215 (Springer, Berlin, 2015), pp. 141–160
[22] P. Junod, S. Vaudenay, Optimal key ranking procedures in a statistical cryptanalysis, in T. Johansson, editor, *Fast Software Encryption*. LNCS, vol. 2887 (Springer, Berlin, 2003), pp. 235–246
[23] P. Junod, On the complexity of Matsui's attack, in S. Vaudenay, A.M. Youssef, editors, *Selected Areas in Cryptography*. LNCS, vol. 2259 (Springer, Berlin, 2001), pp. 199–211
[24] P. Junod, On the optimality of linear, differential and sequential distingishers, in E. Biham, editor, *Advances in Cryptology—EUROCRYPT 2003*. LNCS, vol. 2656 (Springer, Berlin, 2003), pp. 17–32
[25] B.S. Kaliski Jr., M.J.B. Robshaw, Linear cryptanalysis using multiple approximations, in Y.G. Desmedt, editor, *Advances in Cryptology—CRYPTO'94*. LNCS, vol. 839 (Springer, Berlin, 1994), pp. 26–39
[26] M. Matsui, The first experimental cryptanalysis of the Data Encryption Standard, in Y.G. Desmedt, editor, *Advances in Cryptology—CRYPTO'94*. LNCS, vol. 839 (Springer, Berlin, 1994), pp. 1–11
[27] M. Matsui, Linear cryptanalysis method for DES cipher. in T. Helleseth, editor, *Advances in Cryptology—EUROCRYPT'93*. LNCS, vol. 765 (Springer, Berlin, 1994), pp. 386–397

[28] A. Maximov, T. Johansson, Fast computation of large distributions and its cryptographic applications, in B. Roy, editor *Advances in Cryptology—ASIACRYPT*. LNCS, vol. 3788 (Springer, Berlin, 2005), pp. 313–332.
[29] S. Murphy, The independence of linear approximations in symmetric cryptology. *IEEE Trans. Inf. Theory*, **52**(12):5510–5518 (2006)
[30] K. Nyberg, Affine linear cryptanalysis, in *Cryptography and Communications*, 8 (2018), pp. 1–11.
[31] J. Pieprzyk, (ed), *Topics in Cryptology—CT-RSA 2010*, LNCS. vol. 5985 (Springer, Berlin, 2010).
[32] L. Råde, B. Westergren, *Beta Mathematics Handbook*, 2nd edn. (CRC Press, Boca Raton, 1992)
[33] A.A. Selçuk, On probability of success in linear and differential cryptanalysis. *J. Cryptol.*, **21**(1):131–147 (2008)
[34] S. Vaudenay, An experiment on DES statistical cryptanalysis, in *CCS'96: Proceedings of the 3rd ACM Conference on Computer and Communications Security, New York, NY, USA* (1996), pp. 139–147 ACM.