Linear Approximations of Random Functions and Permutations

Mohsin Khan and Kaisa Nyberg

mohsin.khan@helsinki.fi,kaisa.nyberg@aalto.fi

September 1, 2019

Abstract. The goal of this paper is to investigate linear cryptanalysis of random functions and permutations. Our motivation is twofold. First, before a practical cipher can be distinguished from an ideal one, the cryptanalyst must have an accurate understanding of the statistical behavior of the ideal cipher. Secondly, this issue has been neglected both in old and in more recent studies, particularly when multiple linear approximations are being used simultaneously. Traditionally, the models have been based on the average behavior and simplified using other artificial assumptions such as independence of the linear approximations. The new models given in this paper are realistic, accurate and easy to use. They are backed up by standard statistical tools such as Pearson's χ^2 test and finite population correction and shown to work well in small practical examples.

Keywords: random function · random permutation · multinomial distribution · block cipher · multidimensional linear cryptanalysis · correlation · capacity · wrong-key hypothesis · ideal cipher

1 Introduction

Modelling linear key-recovery attacks. Linear cryptanalysis is a statistical method used for distinguishing a block cipher from a random permutation and can be extended to key recovery attacks for practical block ciphers. It makes use of the nonrandom behavior of linear approximations of the cipher. Linear approximations are single-bit values obtained by exclusive-or summation of certain input bits and output bits of the block cipher, over some rounds of the block cipher.

In the setting of linear key-recovery attacks, the traditional heuristic assumption is that the cipher $E(K, x)$ becomes a pseudorandom function or permutation if some of its rounds are replaced by encryption using a wrong key. On the other hand, if the key is correct, then the data is computed from the cipher. Distinguishing between these two cases using statistical tests requires statistical models of the test statistic for both cases. For a recent overview of the existing models, we refer to [BTV18]. Such statistical models are always based on trade-offs between accuracy and feasibility. The traditional approach has been to state some unproven assumptions, called as *wrong-key hypothesis* and *right-key hypothesis*, which are desired to capture the statistical behavior, but still simple enough to allow feasible computation of the model.

In all existing studies the wrong-key hypothesis in linear cryptanalysis, as well as in other statistical attacks, is based on some understanding of the behavior of the family of random permutations, when the target cipher is a block cipher, or the behavior of the family of random functions in some other cases such as stream ciphers.

Then the main effort in the cryptanalytic attack is focused on identifying and demonstrating evidence of nonrandom behavior in the target cipher. In linear cryptanalysis

the problem is to find bit combinations that exhibit nonrandom behavior. The known search algorithms for finding good linear approximations are based on Matsui's seminal work [Mat94], where strong linear approximations are found by identifying one ore more strong linear approximation trails that the linear approximations is composed of. The right-key hypothesis is then derived from a statistical model that captures the probability distributions and their parameters of the linear approximations in the case of the cipher.

The success probability and the data complexity of the attack are then estimated based on statistical distinguishing between the probability distributions in the right-key case and the wrong-key case. It is clear that a proper understanding of random behavior is in an essential role in statistical cryptanalysis.

Along the history of the linear cryptanalysis method the wrong-key hypothesis has taken different forms and the main contributions are rather scattered in the literature. The first goal of this paper is to give a concise presentation of random behavior under the linear cryptanalysis. Our second goal is to present a new and more realistic model of the wrong-key hypothesis for the multidimensional linear cryptanalysis. The statistical behavior turns out to depend significantly on the structure of a multidimensional linear approximation.

Existing wrong-key models in linear cryptanalysis. The understanding about the statistical behavior of linear approximations of random functions and permutations has developed a lot during the times. In early works, correlations of linear approximations of random permutations were estimated to be negligible and equal to their expected value, zero. While it was understood already in 1994 by O'Connor [O'C95] that the correlations of linear approximations vary within the family of functions or permutations, it was not until in 2006 this fact was examined in more detail by Daemen and Rijmen [DR07]. They considered the probability distribution of correlations of linear approximations of random functions and permutations and showed that they behave similarly and can be approximated using normal distributions with the same parameters with the distinction that the interval of the discrete distribution for permutations can have only even values.

These advanced models of linear approximations of random functions and permutations led to the observation that if a linear approximation of a cipher has correlation equal to zero for all keys then it is not random and can be distinguished from random [BR14]. Conversely, this means that, under the traditional hypothesis according to which correlations of linear approximations of random permutations are equal to zero, even random permutations will be identified as nonrandom given sufficient amount of data. This example illustrates how important it is to state the wrong-key assumption accurately.

The model of [DR07] was extended to a wrong-key model by Bogdanov and Tischhauser [BT13] by integrating data sampling to it. While being an important opening to key-dependent models, it had two main drawbacks. First, the right-key model was still based on the assumption that the correlation of the linear approximation is equally large (in absolute value) for all keys. Secondly, the plaintexts were assumed to be drawn with replacement. While giving realistic estimates for small sample sizes, this approach leads to significant deviations from the true behavior when the sample size approaches the full codebook. These two drawbacks lead to the unintuitive observation that in the context of this model the success probability is not an increasing function of the data-complexity. These problems were corrected in the model given in [BN17b] in the case of a single linear approximation based on a single dominant trail.

With the goal of making the linear distinguishers more powerful, several authors have proposed to use multiple linear distinguishers simultaneously. Before key-dependency of the probability distributions was discovered to be a relevant issue, the wrong-key hypothesis was always based on the assumption that in the wrong-key case, the expected correlations, that is, the correlations of linear approximations computed for the full codebook of

the cipher behave as on average, that is, are equal to zero [HCN19]. In the process of integrating key-dependency into the models all works so far, see e.g. [BTV16, BN17a, BN17b], adopted the simplifying assumption that the correlations of any set of multiple linear approximations of a random permutation (that is, considered over the set of all permutations) are independent. A subsequent version [BTV18] of [BTV16], this assumption was stated only for correlations of linearly independent linear approximations of a random permutation. Whether this means a true theoretical improvement is not known.

In general, not much is known about the statistical independence of correlations considered as random variables over the key space. Only correlations of components of permutations are known to be independent as they are always constants, that is, equal to zero. A multidimensional approximation computed for a permutation is not in general a balanced function. Hence the correlations of its components may not be equal to zero and may have statistical dependencies.

The assumption about independence of correlations was needed to derive statistical distributions of the sum of squares of the correlations, also called as capacity, of the individual linear approximations. More specifically, the independence assumption has been used for expressing the variance of the sum of squared correlations as the sum of the variances of the squared correlations of the individual linear approximations. In this paper, it will be shown that this result can be achieved without the independence assumption for certain sets of linear approximations.

Our contributions. We start by deriving exact formulas for the mean and variance of the capacity of the value distribution of a multinomially distributed variable and make the observation that the variance of the capacity is additive, that is, it can be expressed as the sum of the variances of the capacities of the individual variables in the case when the expected distribution is uniform. This corresponds to the case of the expected value distribution of a random function.

We continue by revisiting the distributions of correlations of single linear approximations of random functions and random permutations. As an addition to [DR07] we observe that a linear approximation of a random Boolean function is also random, while the exact distribution of a linear approximation of a random permutation can be given in terms of a hypergeometric distribution.

While multidimensional linear approximations of some functions can be modeled using the multinomial distribution, this is never the case for a multidimensional linear approximation of permutations. Even in case of a single variable, the hypergeometric distribution must be used instead of the binomial distibution. We leave it an open question whether the multivariate hypergeometric distribution might give a feasible approach in this case, and instead, use continuous approximations of the probability distributions to model the statistical behavior of the capacity of a multidimensional linear approximation of a random permutation. This leads us to the study of the χ^2 distribution.

In many practical applications of multidimensional linear cryptananlysis, the linear space of linear approximations contains many trivial approximations that have a constant correlation zero. Their impact has been ignored in previous works. We show that they may create a significant error factor if not treated properly. We conjecture a model of the behavior of capacity under existence of trivial linear approximations and present experimental evidence to support this conjecture. We also identify a type of multidimensional linear approximation, which we call the Davies-Mayer approximation and prove that a multidimensional linear approximation is a Davies-Mayer approximation if and only if it does not contain trivial linear approximations.

Having found a realistic solution to the problem of how to model wrong-key behaviour for multidimensional linear cryptanalysis, we apply the same approach for the recently presented variant of linear cryptanalysis, named as affine multidimensional cryptanaly-

sis [Nyb19].

Outline. The standard definitions of linear cryptanalysis are recalled in Section 2. Then the mean and variance of capacity are computed for a general multinomial distribution in Section 3 and then we recall the related discrete probability distributions and their continuous approximations. In Section 4, the distributions of correlations of single linear approximations are revisited. The new contributions of the structure and probability distributions of multidiemsional linear approximations are presented in Section 5 and applied to an affine set of approximations in Section 6. The conclusions are drawn in Section 8.

2 Preliminaries and Notation

2.1 Correlations and Capacity

Let F be a function from S to $\mathbb{F}_2^s$, where S is an arbitrary set and $\mathbb{F}_2^s$ is a vector space over $\mathbb{F}_2$ of dimension s. We focus on two ways of defining F. First, we can define it by giving s Boolean functions $f_1, \ldots, f_s$, that is, s coordinate functions of F, and their values $f_i(x)$, $x \in S$, $i = 1, \ldots, s$. Given $\alpha = (\alpha_1, \ldots, \alpha_s) \in \mathbb{F}_2^s$, we denote by $\alpha \cdot F$ the linear combination of the coordinate functions of $F = (f_1, \ldots, f_s)$ determined by α, that is,

$$\alpha \cdot F = \alpha_1 f_1 + \ldots + \alpha_s f_s,$$

and call $\alpha \cdot F$ a component of F. We say that t components $\alpha_i \cdot F$, $i = 1, \ldots, t$, are linearly independent if the vectors α_i, $i = 1, \ldots, t$, are linearly independent. The second way of defining F is just by giving the (indexed) set of its values $F(x)$, $x \in S$.

Functions are in general imbalanced, that is, all values in the image space are not taken equally often. Related to the two ways of defining F, we have two ways of measuring the imbalance of F. First, we can determine the imbalance of its components using correlations. Let f be a $\mathbb{F}_2$-valued function in S. Then its correlation $\mathrm{cor}(f)$ is given by

$$\mathrm{cor}(f) = 2^{-n}(\#\{x \in S \,|\, f(x) = 0\} - \#\{x \in S \,|\, f(x) = 1\}). \tag{1}$$

Secondly, we can determine the imbalance of F by measuring the uniformity of its value distribution. Given $\eta \in \mathbb{F}_2^s$ let us denote by p_η its probability, that is,

$$p_\eta = 2^{-n}\#\{x \in S \,|\, F(x) = \eta\}.$$

Then the imbalance of this distribution is measured using capacity

$$\mathrm{Cap}(F) = 2^s \sum_{\eta \in \mathbb{F}_2^s} (p_\eta - 2^{-s})^2. \tag{2}$$

It is well-known that these two approaches to measuring imbalance are related due to the following equality,

$$p_\eta = 2^{-s} \sum_{\alpha \in \mathbb{F}_2^s} (-1)^{\alpha \cdot \eta} \mathrm{cor}(\alpha \cdot F), \text{ for all } \eta \in \mathbb{F}_2^s, \tag{3}$$

or equivalently,

$$\mathrm{cor}(\alpha \cdot F) = \sum_{\eta \in \mathbb{F}_2^s} (-1)^{\alpha \cdot \eta} p_\eta, \text{ for all } \alpha \in \mathbb{F}_2^s. \tag{4}$$

Then we can express $\mathrm{Cap}(F)$ also as

$$\mathrm{Cap}(F) = \sum_{\alpha \in \mathbb{F}_2^s, \alpha \neq 0} \mathrm{cor}(\alpha \cdot F)^2. \tag{5}$$

3 Capacity of Random Function

In particular, a random function $F : S \to \mathbb{F}_2^s$ can be generated either by selecting s Boolean functions randomly and independently, or by picking its values $F(x)$ randomly and independently from $\mathbb{F}_2^s$. The value distribution of a random function F is a multinomial distribution. By Equation 5 the expected value of the capacity of the value distribution of a random function is the sum of the expected values of the squared correlations taken over the non-trivial components of F. We are also interested to determine the variance of the capacity for random functions. The problem is not trivial, since we can neither assume all non-trivial components of F to be independent, nor to have independent correlations. Nevertheless, it will be shown that, based solely on the properties of the multinomial value distribution of a random function F, the variance of its capacity is obtained as the sum of the variances of the squared correlations of its non-trivial components.

3.1 Capacity of General Multinomial Distribution

We first give the mean and variance of capacity for a general multinomial distribution and then obtain these parameters for the uniform expected distribution as a special case.

Let $x_1, \ldots, x_k$ be a set of k, $k \geq 2$, stochastic variables that follow a multinomial distribution with expected probabilities $p_i, \ldots, p_k$ and number of trials m. The density function of this distribution is given by

$$f(x_1, \ldots, x_k) = \frac{m!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}.$$

Then $x_1 + \cdots + x_k = m$ and the capacity of this distribution is given by

$$C = \frac{k}{m^2} \sum_{i=1}^{k} (x_i - \frac{m}{k})^2.$$

Hence the capacity is also a stochastic variable. The proof of the following result is given in Appendix A.

Theorem 1. *Let C be the capacity of a multinomial distribution with parameters $p_1, \ldots, p_k$ and m. Then*

$$\begin{aligned} \operatorname{Exp} C &= \frac{k-1}{m} + \frac{(m-1)k}{m} \sum_{i=1}^{k} (p_i - \frac{1}{k})^2 \\ \operatorname{Var} C &= \frac{(m-1)k^2}{m^3} \left((4m-8)P_3 - (4m-6)P_2^2 + 2P_2 \right), \end{aligned}$$

where

$$P_2 = \sum_{i=1}^{k} p_i^2 \quad \textit{and} \quad P_3 = \sum_{i=1}^{k} p_i^3.$$

Note that in the expression of the expected capacity we have

$$k \sum_{\eta \in \mathbb{F}_2^s} (p_i - \frac{1}{k})^2 = kP_2 - 1,$$

which is the capacity of the expected distribution p_i, $i = 1, \ldots, k$.

If the expected distribution is uniform, that is, $p_i = \frac{1}{k}$, for all $i = 1, \ldots, k$, then $P_2 = 1/k$ and $P_2^2 = P_3 = 1/k^2$, and the capacity parameters are as follows.

Corollary 1. *Let C be the capacity of a multinomial distribution with parameters $p_i = \frac{1}{k}$, for all $i = 1, \ldots, k$, and m. Then*

$$\begin{aligned} \operatorname{Exp} C &= \frac{k-1}{m} \\ \operatorname{Var} C &= \frac{2(k-1)(m-1)}{m^3}. \end{aligned}$$

3.2 Standard probability distributions

The multinomial distribution with $k = 2$ is the binomial distribution which we will denote by $\mathcal{B}(m, p)$, where $p = p_1$ and $1 - p = p_2$. The mean and variance of this distribution are mp and $mp(1-p)$, respectively. The binomial distribution corresponds to random sampling with replacement from a set S where we have two types of elements, denoted by 0 and 1. If the sampling is without replacement then the number of outcomes of type 0 follows the hypergeometric distribution $\mathcal{H}G(|S|, K, m)$, where K is the number of elements of kind 0 in the entire S. The mean and variance of the hypergeometric distribution are

$$m\frac{K}{|S|} = mp \text{ and } m\frac{K}{|S|}\frac{|S|-K}{|S|}\frac{|S|-m}{|S|-1} = mp(1-p)\frac{|S|-m}{|S|-1},$$

where we denoted by p the probability of outcomes of type 0 in the entire set S, that is, $p = \frac{K}{|S|}$. The variances of the binomial and hypergeometric distributions differ by a factor, whose close estimate

$$B = \frac{|S|-m}{|S|}, \tag{6}$$

is called the finite population correction coefficient [RT89]. For sufficiently large S, both distributions can be approximated by the normal distribution $\mathcal{N}(\mu, \sigma^2)$, where μ is the mean and σ^2 is the variance, as follows:

$$\mathcal{B}(m, p) \approx \mathcal{N}(mp, mp(1-p)) \text{ and } \mathcal{H}G(|S|, K, m) \approx \mathcal{N}(mp, mp(1-p)B). \tag{7}$$

The distribution of the sum of squares of ℓ independent standard normal deviates is called the chi-squared distribution and denoted by $\chi^2_\ell(\delta)$, where ℓ is the degree of freedom and δ is the non-centrality parameter computed as the sum of squares of the means of the said normally distributed variables. The mean of the $\chi^2_\ell(\delta)$ distribution is $\ell + \delta$ and its variance is $2(\ell + 2\delta)$.

The well-known Pearson's chi-squared test is defined in the same setting as the multinomial distribution Subsection 3.1. The test statistic defined as

$$T = \sum_{i=1}^{k} \frac{(x_i - mp_i)^2}{mp_i} \tag{8}$$

follows the $\chi^2_{k-1}(\delta)$ distribution, where

$$\delta = \sum_{i=1}^{k} (mp_i)^2. \tag{9}$$

Then $T = mC$, and the chi-squared distribution of T gives a continuous approximation of the discrete probability distribution of C. In the case where $p_i = 1/k$ for all $i = 1 \ldots, k$ the mean of C obtained from the χ^2 distribution of T is $(k-1)/m$ which is the same as the mean given by Corollary 1, while the variances differ by a negligble term $2(k-1)/m^3$.

The multinomial distribution and the related Pearson's χ^2 distribution apply to the case when the values x_i are obtained by drawing samples of m elements from S with replacement. If sampling is without replacement then the multivariate hypergeometric distribution shall be used instead of the multinomial distribution. Then the statistic T given in Equation 8 must be multiplied by the inverse of the finite population correction coeffifient to get a χ^2-distributed variable [RT89]. We state this result for further reference as follows.

Lemma 1. *Let T be given by Equation 8 where the values of variables x_i, $i = 1, \ldots, k$ are obtained by sampling m elements from S without replacement and the initial probabilities p_i are as defined in the setting of the multinomial distribution. Then the variable*

$$B^{-1}T,$$

where B is given by Equation 6, approximately follows $\chi^2_{k-1}(\delta)$ distribution, where δ is given by Equation 9.

4 Probability Distribution of Single Linear Approximation

We first derive the distributions of linear approximations of random Boolean functions and random balanced Boolean functions.

4.1 Zeroes of Boolean functions

Let f be a Boolean function in $\mathbb{F}_2^n$. We say that $x \in \mathbb{F}_2^n$ is a zero of f if $f(x) = 0$ and denote by N_0 the set of the zeroes of f. Given a vector $a \in \mathbb{F}_2^n$ the Boolean function $g(x) = f(x) + a \cdot x$ is called a linear approximation of f. To compute the correlation of $g(x) = f(x) + a \cdot x$ let us first determine the number of its zeroes.

Lemma 2. *Let $a \cdot x$ be a linear function in $\mathbb{F}_2^n$. Then the number of zeros of the linear approximation $g(x) = f(x) + a \cdot x$ is equal to*

$$\#\{x \in \mathbb{F}_2^n \mid f(x) = 0,\, a \cdot x = 0\} + \#\{x \in \mathbb{F}_2^n \mid f(x) = 1,\, a \cdot x = 1\} = 2^{n-1} - N_0 + 2z,$$

where we denoted

$$z = \#\{x \in \mathbb{F}_2^n \mid f(x) = 0,\, a \cdot x = 0\} \tag{10}$$

Proof. Clearly

$$\#\{x \in \mathbb{F}_2^n \mid f(x) = 1,\, a \cdot x = 1\} = 2^{n-1} - (N_0 - z).$$

□

The following lemma gives the distribution z.

Lemma 3. *Let Boolean function f in $\mathbb{F}_2^n$ be chosen randomly and equiprobably from the set of all Boolean functions with a fixed number N_0 of zeroes and $a \cdot x$ be a fixed linear function. Then z defined by Equation 10 follows the hypergeometric distribution $\mathcal{HG}(2^n, 2^{n-1}, N_0)$.*

Proof. Given a fixed linear function $a \cdot x$, the N_0 zeros of f are chosen by choosing z zeros among the 2^{n-1} zeros of $a \cdot x$ and $N_0 - z$ zeros among the 2^{n-1} inputs x such that $a \cdot x = 1$. □

4.2 Random function

Let us first consider random Boolean functions. For a random Boolean function the number of its zeros follow the binomial distribution $\mathcal{B}(2^n, \frac{1}{2})$.

Theorem 2. *Let f be selected randomly and equiprobably from the set of all Boolean functions of n variables. Then for any fixed $a \in \mathbb{F}_2^n$ the number of zeros of the linear approximation $a \cdot x + f(x)$ follows a binomial distribution $\mathcal{B}(2^n, \frac{1}{2})$. In other words, any linear approximation of a random Boolean function is a random Boolean function.*

Proof. For any fixed linear Boolean function $g(x) = a \cdot x$, the mapping which maps a Boolean function f to the function $f + g$ is a bijection in the set of all Boolean functions. Hence the distribution of the number of zeroes of $f + g$ follows the same distribution as the number of zeroes of f when f is drawn uniformly at random from the set of Boolean functions. □

For an alternative proof that computes the distribution of the zeroes of the linear approximation, see Appendix B.

By application of Corollary 1 for $k = 2$ we get the following result.

Corollary 2. *The distribution of a correlation c of a linear approximation of a random Boolean function of n variables has the following parameters:*

$$\begin{aligned} \mathrm{Exp}(c) &= 0 \\ \mathrm{Var}(c) &= \mathrm{Exp}(c^2) = 2^{-n} \\ \mathrm{Var}(c^2) &= 2^{1-2n} - 2^{1-3n}. \end{aligned}$$

The distribution of $2^{n/2}c$ can be approximated by the standard normal distribution, and the distribution of its square $2^n c^2$ by the χ^2 distribution with one degree of freedom.

4.3 Balanced random function

From Lemma 2 and Lemma 3 we get the following result.

Theorem 3. *Let f be selected randomly and equiprobably from the set of all balanced Boolean functions of n variables. Then for any fixed $a \in \mathbb{F}_2^n$ the number of zeros of the linear approximation $f(x) + a \cdot x$ is an even integer $2z$ where $z \sim \mathcal{HG}(2^n, 2^{n-1}, 2^{n-1})$.*

Corollary 3. *The distribution of a correlation c of a linear approximation of a random balanced Boolean function of n variables has the following parameters:*

$$\mathrm{Exp}(c) = 0 \ \textit{and}\ \mathrm{Var}(c) = \mathrm{Exp}(c^2) = \frac{1}{2^n - 1}\,.$$

Proof. We have $c = 2^{2-n}z - 1$, where

$$\mathrm{Exp}(z) = 2^{n-2} \ \text{ and } \ \mathrm{Var}(z) = \frac{(2^{n-2})^2}{2^n - 1}\,.$$

□

For the values of n typically used in block ciphers, the distribution of $2^{n/2}c$ can be approximated using the standard normal distribution, and the distribution of its square $2^n c^2$ using χ^2 distribution with one degree of freedom. By [DR07], it suffices to have $n \geq 5$.

Our derivations of the distributions of correlations of linear approximations of random and random balanced Boolean functions are essentially the same as those given by Daemen and Rijmen in [DR07]. In this section, we completed their work by giving the exact distributions in both cases and the observation that a linear approximation of a random Boolean function is itself a random Boolean function.

4.4 Random vectorial function and random permutation

A single linear approximation of a random vectorial Boolean function is a random Boolean function and hence the number of its zeros is binomially distributed as given by Theorem 2. Since a single component of a random permutation is a balanced random Boolean function, the distribution of the zeros of a single linear approximation is given by Theorem 3.

5 Multidimensional Linear Approximation

5.1 Multidimensional linear approximation as a vectorial Boolean function

In this section we give a description of a multidimensional linear approximation as a vectorial Boolean function.

In the context of linear cryptanalysis, a linear approximation (a, b) of an n-bit permutation P can also be identified with an n-variable Boolean function defined as

$$x \mapsto a \cdot x + b \cdot P(x).$$

The multidimensional linear cryptanalysis method considers a number of linear approximations (a, b) that form a linear subspace in the vector space of $2n$-bit vectors. We denote this subspace by L and its dimension by t. Then a multidimensional linear approximation can be identified with a vector-valued Boolean function, where the inputs are n-bit vectors and outputs are t-bit vectors. We denote this vector-valued Boolean function by Λ. Then the components of Λ are the linear approximations (considered as Boolean functions) with masks in L. It means that for each $(a, b) \in L$ there is a unique t-bit vector β such that the equality

$$a \cdot x \oplus b \cdot P(x) = \beta \cdot \Lambda(x), \tag{11}$$

holds for all n-bit inputs x.

By Equation 2 and Equation 5 the capacity of Λ is given as

$$\mathrm{Cap}(\Lambda) = \sum_{(a,b)\in L,(a,b)\neq 0} \mathrm{cor}(a, b)^2 = \sum_{\beta\neq 0} \mathrm{cor}(\beta \cdot \Lambda)^2 = 2^t \sum_{\eta} (p_\eta - 2^{-t})^2, \tag{12}$$

where the second sum is taken over all t-bit vectors η of the value space of Λ. One known consequence of this result is that the value distribution of a multidimensional linear approximation is uniform if and only if the correlations of all its non-zero linear approximations are equal to zero.

5.2 Structure of a multidimensional linear approximation

A multidimensional linear approximation L may contain mask pairs of the form $(a, 0)$ and $(0, b)$. We call the corresponding linear approximations trivial. Let us now examine their effect on the distribution of the capacity $\mathrm{Cap}(\Lambda)$. Let us denote by U the linear subspace of the multidimensional approximation consisting of the approximations of the form $(a, 0)$ and let u be its dimension. Similarly, we denote by V the subspace of the masks of the form $(0, b)$ and by v its dimension. Often such spaces span the whole multidimensional approximation, that is, all masks are of the form (a, b), where $(a, 0) \in U$ and $(0, b) \in V$, but in general, there may exist a linear subspace W such that the entire mask space of the multidimensional approximation L can be written as a direct sum

$$L = U \oplus V \oplus W, \tag{13}$$

where W is such that a and b are nonzero for all $(a,b) \in W$, $(a,b) \neq (0,0)$, which implies that $U \cap W = V \cap W = \{(0,0)\}$.

Let us denote by Λ_1, Λ_2 and Λ_3 the multidimensional linear approximations determined by the mask sets U, V and W, respectively. Then the values of Λ_1 are u-bit vectors, the values of Λ_2 are v-bit vectors, and the values of Λ_3 are $(t-u-v)$-bit vectors, and $\Lambda = (\Lambda_1, \Lambda_2, \Lambda_3)$. Since all linear approximations in U and V are balanced, the value distributions of Λ_1 and Λ_2 are uniform. Considering this property for Λ_1 we get 2^u conditions for the value distribution of Λ as follows

$$\sum_{\eta,\nu} \Pr(\Lambda(x) = (\xi,\eta,\nu)) = \Pr(\Lambda_1(x) = \xi) = 2^{-u}, \text{ for all } u\text{-bit vectors } \xi.$$

From these conditions $2^u - 1$ are independent, since

$$\sum_{\xi} \Pr(\Lambda_1(x) = \xi) = 1.$$

Similarly, by the uniformity of the value distribution of Λ_2, we get the following conditions from which $2^v - 1$ are independent.

$$\sum_{\xi,\nu} \Pr(\Lambda(x) = (\xi,\eta,\nu)) = \Pr(\Lambda_2(x) = \eta) = 2^{-v}, \text{ for all } v\text{-bit vectors } \eta.$$

We conclude that the degree of freedom of the full-codebook value distribution of a multidimensional linear approximation Λ of a permutation, as considered above, is bounded from above by

$$2^t - 1 - (2^u - 1) - (2^v - 1) = 2^t - 2^u - 2^v + 1.$$

Let us now consider Λ and the probabilities p_ζ of its t-bit values as statistical variables over the space of all permutations. We apply Pearson's χ^2 test and compute the test variable as

$$T(\Lambda) = 2^n \sum_{\nu} \frac{(p_\zeta - 2^{-t})^2}{2^{-t}} = 2^n \mathrm{Cap}(\Lambda). \tag{14}$$

Then $T(\Lambda)$ follows the χ^2 distribution. For linear approximations of random permutations the expected value for each correlation $\mathrm{cor}(a,b)$, with $(a,b) \neq 0$, is equal to zero, also for those where $a \neq 0$ and $b \neq 0$. Hence the expected value of each p_ζ is equal to 2^{-t}. Thus we have proved the following result.

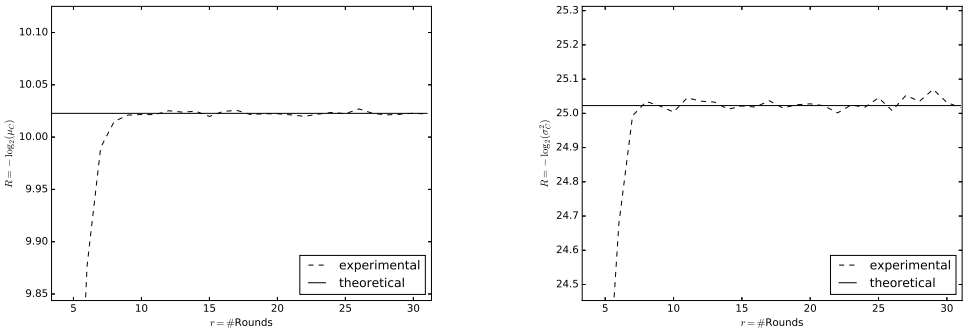
Theorem 4. *Let the multidimensional linear approximation have dimension t and the linear subspaces of trivial masks of the form $(a,0)$ and $(0,b)$ have dimensions u and v, respectively. Then for a random permutation of n-bit vectors the capacity of this multidimensional linear approximation follows, when multiplied by the factor 2^n, the central χ^2 distribution with at most $2^t - 2^u - 2^v + 1$ degrees of freedom.*

Based on this result, we conjecture that the value distribution of a multidimensional linear approximation of a random permutation with mask subspaces U and V of dimensions u and v, respectively, has the maximum degree of freedom, that is, $2^t - 2^u - 2^v + 1$.

Conjecture 1. *For permutations in $\{0,1\}^n$ drawn uniformly at random the capacity of a multidimensional linear approximation with dimension t and the linear subspaces of trivial masks with dimensions u and v follows, when multiplied by 2^n, the χ^2 distribution with $2^t - 2^u - 2^v + 1$ degrees of freedom.*

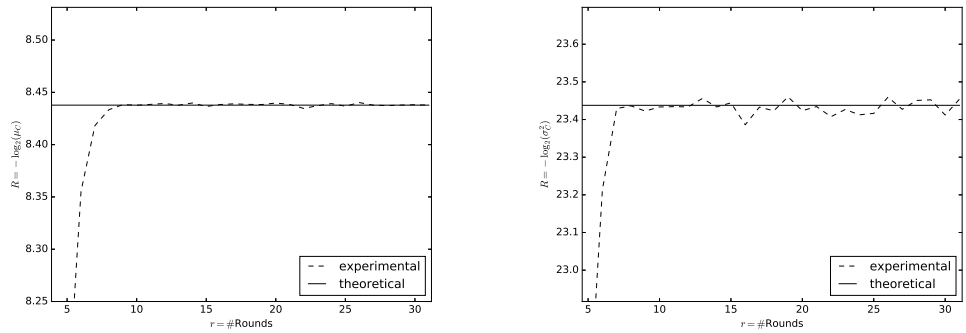
5.3 Experiments

We performed some experiments to check the validity of Conjecture 1. The first set of experiments is performed to illustrate the mean and variance of a multidimensional approximation defined by a mask space of the form $U \times V$ where the dimension of U is 6 and the dimension of V varies from 1 to 6. We used multiple rounds of SMALLPRESENT-[4] with 2^{14} random keys to simulate the family of random permutation. The test cipher SMALLPRESENT-[4] is an iterated block cipher with 31 rounds and block size 16 bits [Lea10]. The state bits at input and output to each round are numbered from 0 to 15 from right to left. For each key and for a fixed number of rounds the distribution of the capacity is computed. Then the mean and the variance over the 2^{14} keys is computed. In the Figures 1 – 6 the negatives of the base 2 exponents, that is $-\log_2$, of the mean and variance are plotted as the number of rounds increases. In each case the hypothetical value given by Conjecture 1 is depicted using a horisontal line.



(a) Mean (b) Variance

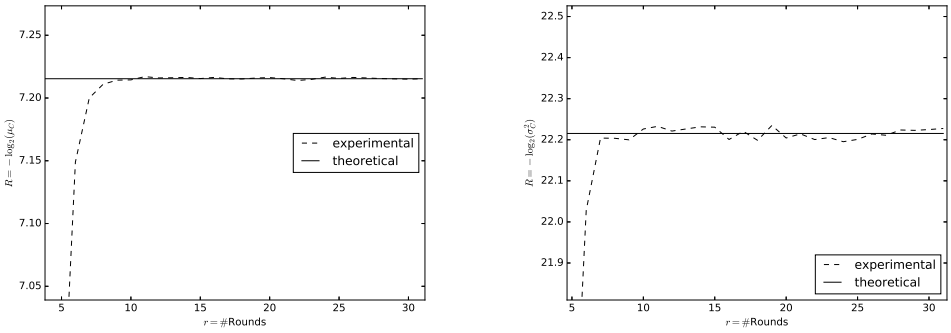
Figure 1: Mean and variance of capacity of multidimensional linear approximation of dimension 7. Input masks spanned by bits $9, 10, 11, 13, 14, 15$. Output masks spanned by bit: 9.



(a) Mean (b) Variance

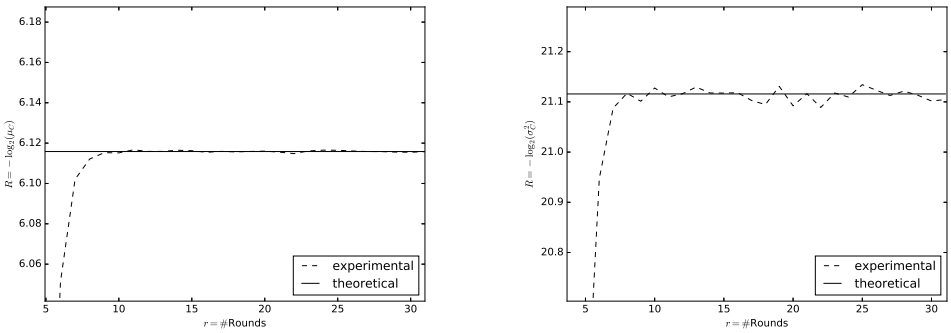
Figure 2: Mean and variance of capacity of multidimensional linear approximation of dimension 8. Input masks spanned by bits $9, 10, 11, 13, 14, 15$. Output masks spanned by bits: $9, 10$.

We see that the results of the experiments support Conjecture 1 perfectly. Finally we also show an example of the full experimental probability distribution of the capacity



(a) Mean (b) Variance

Figure 3: Mean and variance of capacity of multidimensional linear approximation of dimension 9. Input masks spanned by bits $9, 10, 11, 13, 14, 15$. Output masks spanned by bits: $9, 10, 11$.



(a) Mean (b) Variance

Figure 4: Mean and variance of capacity of multidimensional linear approximation of dimension 10. Input masks spanned by bits $9, 10, 11, 13, 14, 15$. Output masks spanned by bits: $9, 10, 11, 13$.

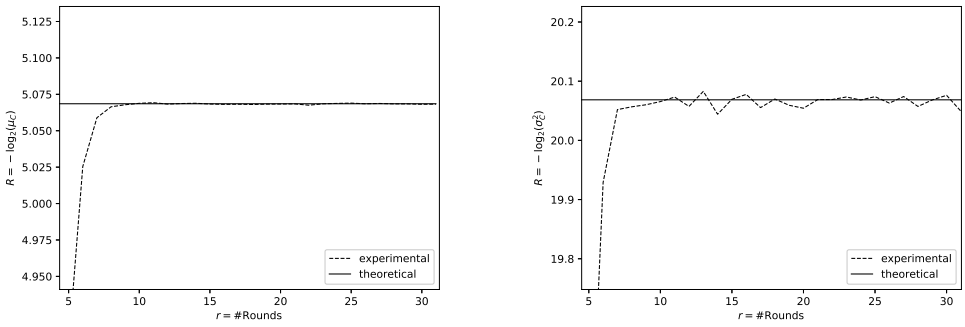
plotted in Figure 7. It is computed for 20 rounds of SMALLPRESENT-[4] for subspaces U and V of dimension 4. Both U and V are spanned by bits in positions 5, 6, 9, and 10, where we denoted by e_j the bit vector with a single 1-bit in position j. It is compared with the χ^2 distribution with $2^8 - 2^4 - 2^4 + 1 = 225$ degrees of freedom plotted as a solid curve.

5.4 Special case $u = v = 0$

Let us start by defining a special type of multidimensional linear approximation, which we call a Davies-Mayer approximation for reasons to be explained in this section.

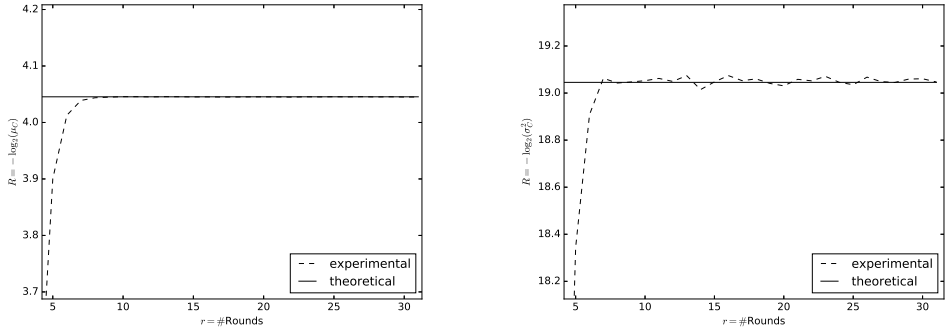
Definition 1. A multidimensional linear approximation L is called a Davies-Mayer approximation if given any linearly independent set of mask pairs (a_i, b_i), $i = 1, \ldots, t$, which span L, the set of input masks a_i, $i = 1, \ldots, t$, and the set of output masks b_i, $i = 1, \ldots, t$, are linearly independent.

Given a Davies-Mayer approximation, we can define a linear bijection A from the set



(a) Mean (b) Variance

Figure 5: Mean and variance of capacity of multidimensional linear approximation of dimension 11. Input masks spanned by bits $9, 10, 11, 13, 14, 15$. Output masks spanned by bits: $9, 10, 11, 13, 14$.



(a) Mean (b) Variance

Figure 6: Mean and variance of capacity of multidimensional linear approximation of dimension 12. Input masks spanned by bits $9, 10, 11, 13, 14, 15$. Output masks spanned by bits: $9, 10, 11, 13, 14, 15$.

of input masks of L to the set of output masks of L by setting

$$A(a_i) = b_i, \, , \, i = 1, \ldots, t.$$

Then a linear approximation $(a, b) \in L$ can be expressed as

$$a \cdot x + b \cdot P(x) = a \cdot x + a \cdot (A^t \circ P)(x) = a \cdot (x + (A^t \circ P)(x)). \tag{15}$$

If P is a random permutation, that is, picked uniformly at random for the set of all permutations, then the same holds for the permutation $(A^t \circ P)$. The function

$$x \mapsto x + (A^t \circ P)(x)$$

is the well-known Davies-Mayer construction which is indistinguishable from random functions. By Equation 15 the linear approximations in L form a linear subspace of the components of the Davies-Mayer function, and hence is also a random function.

Theorem 5. *If a multidimensional linear approximation of a permutation P does not contain any mask pair of the form $(a, 0)$ or $(0, b)$, where $a \neq 0$, $b \neq 0$, then it is a Davies-*

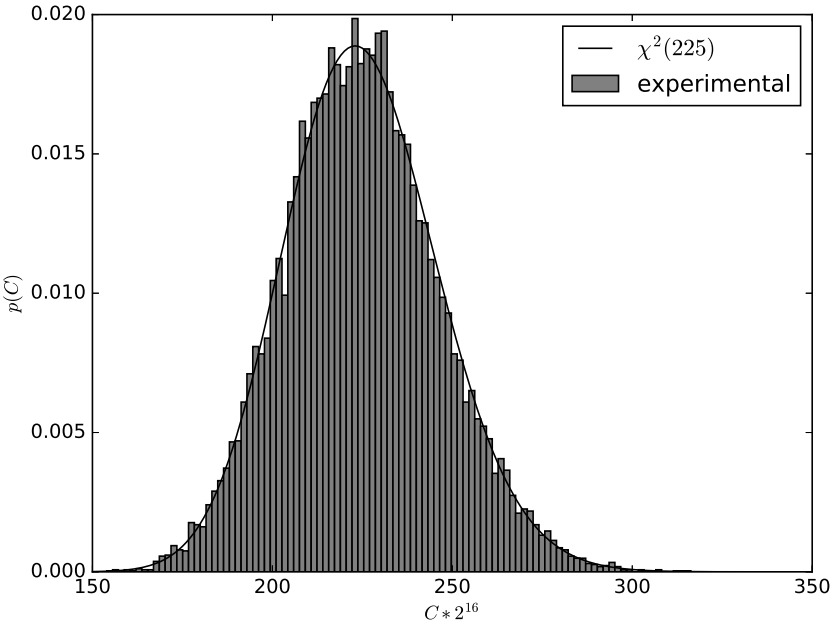


Figure 7: Probability distribution of capacity C of the distribution of 8 bits multiplied by 2^{16}. Input masks spanned by bits $5, 6, 9, 10$. Output masks spanned by bits $5, 6, 9, 10$.

Mayer approximation and is indistinguishable from a random function when P is selected uniformly at random.

Proof. It remains to show that if a multidimensional linear approximation L does not contain any trivial approximations, that is, mask pairs of the form $(a, 0)$ or $(0, b)$, where $a \neq 0$, $b \neq 0$, then L is a Davies-Mayer approximation.

Let us suppose that L is not a Davies-Mayer approximation. Then L has a basis (a_i, b_i), $i = 1, \ldots, t$, where either the set a_i, $i = 1, \ldots, t$, or b_i, $i = 1, \ldots, t$, is linearly independent. Assume a_i, $i = 1, \ldots, t$, are not linearly independent and let a_{i_j}, $j = 1, \ldots, k$, be a subset such that

$$\sum_{j=1}^{k} a_{i_j} = 0.$$

Since (a_i, b_i), $i = 1, \ldots, t$, are linearly independent, it must be the case that

$$\sum_{j=1}^{k} a_{i_j} \neq 0.$$

Then L contains a mask pair of the form $(a, 0)$, which contradicts the assumption. □

Let us note that, a Davies-Mayer approximation does not contain any trivial approximation. Hence in the presentation (13) of L as $L = U \oplus V \oplus W$, the multidimensional approximation W is a Davies-Mayer approximation. Also we see that L is a Davies-Mayer approximation if and only if $U = V = \{(0, 0)\}$. If P is a permutation, then $x + (A^t \circ P)(x)$ has the known fixed point $x_0 = (A^t \circ P)^{-1}(0)$, but this property is not known to be detected from the statistical behavior of the linear approximations of a random permutation P.

To illustrate the distribution of such a linear approximation we depict the distribution of capacity over 2^{14} random keys in Figure 8. The capacity is computed for the 6-dimensional linear approximation over 20 rounds of SMALLPRESENT-[4] spanned by mask pairs (e_9, e_9), (e_{10}, e_{10}), (e_{11}, e_{11}), (e_{13}, e_{13}), (e_{14}, e_{14}), and (e_{15}, e_{15}), where we denoted by e_j the bit vector with a single 1-bit in position j.

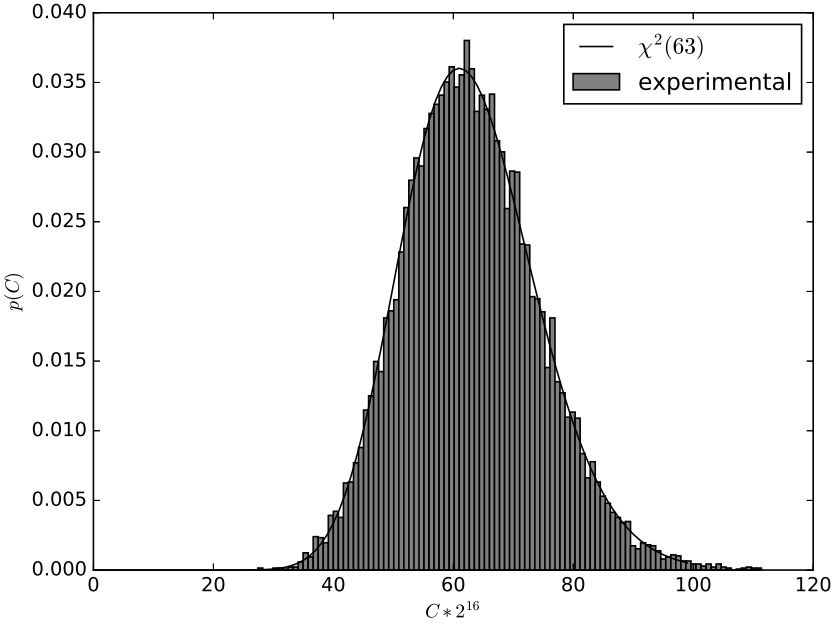


Figure 8: Probability distribution of $2^{16}C$ for capacity C of a 6-dimensional linear approximation with no mask pairs of the form $(a, 0)$ or $(0, b)$, $a \neq 0$, $b \neq 0$

5.5 Multidimensional linear approximation of Serpent

Some of the first applications of multidimensional linear cryptanalysis used block cipher Serpent [HCN09]. The multidimensional approximation L for Serpent was built by taking the linear space spanned by a linearly independent set of m strong base approximations of the form $(a_1, b), \ldots, (a_m, b)$ all with the same output mask b. Then L is of the form $U \oplus V$, where $u = m$ and $v = 1$. Moreover, the Davies-Mayer part W was non-existent. It means that all the linear combinations of the base approximations involving an even number of base approximations had output mask equal to zero, and hence, correlation zero. In the cryptanalysis all $2^{m+1} - 1$ non-zero approximations were involved including those $2^m - 1$ of the form $(a, 0)$ with correlation zero. It was mentioned that such approximations can be ignored in the computation of the empirical correlation. Nevertheless they cannot be ignored when the degree of freedom of the sampled χ^2 statistic is determined as will be explained in Subsection 7.2.

Recently it was proposed by Nyberg to remove the subspace of trivial linear approximations and consider only the remaining set that forms an affine subspce [Nyb19]. Let us apply this idea to the multidimensional approximation of Serpent discussed above. Take the $m-1$ dimensional subspace spanned by masks $(a_2 \oplus a_1, 0), \ldots, (a_m \oplus a_1, 0)$ and denote it by H. Then the affine subspace $(a_1, b) \oplus H$ is only a half of the size of the original linear space and still contains all m strong base approximations. Moreover, for each key,

the full codebook capacity of the affine set of approximations is exactly the same as the full codebook capacity of the original set, while the degrees of freedom of the χ^2 statistic is reduced by one half.

To conclude this section let us mention that the structure of multidimensional linear approximation must be taken in consideration also for non-bijective functions. Then only the mask pairs of the form $(a, 0)$ are trivial with mean and variance of the correlation equal to zero. For example, if in the above example the block cipher Serpent is replaced by some non-bijective function but the same set of linear approximations are used, then removing the trivial approximations leads to an affine set of approximations.

Next we study the distribution of the full codebook capacity for an affine set of linear approximations of a random permutation. In Subsection 7.3 we will recall the sampled χ^2 statistic from [Nyb19] with the following essential improvements: sampling without replacement and key-dependent capacity. The full distribution is then given by integration of the probability distribution of the key-dependent capacity to the sampled χ^2 statistic.

6 Capacity of an affine space of linear approximations

The approach for constructing an affine set of linear approximations which does not contain trivial approximations but has a statistical model without artificial independence assumptions, was proposed by Nyberg [Nyb19]. Such a set can be constructed, for example, by taking an affine subspace of input masks and an affine subspace of output masks to get a set of the form

$$A = (a_0 \oplus U') \times (b_0 \oplus V') = (a_0, b_0) \oplus (U' \times V'),$$

where the dimensions of U' and V' are positive, $a_0 \notin U'$ and $b_0 \notin V'$. We denote

$$U = \{(a, 0) \,|\, a \in U'\} \text{ and } V = \{(0, b) \,|\, b \in V'\}. \tag{16}$$

Then the smallest linear space that contains A is

$$U \oplus V \oplus \{(0, 0), (a_0, b_0)\} = (U \oplus V) \cup A,$$

that is, the space W in the expression (13) has dimension one. But using the multidimensional linear approximation defined by this set of masks instead of using only the set A would add all trivial linear approximations from U and V to this set and reduce the strength of the attack. To avoid wasting attack resources, such as memory and time, we want to exclude the linear approximations with masks in $U \oplus V$.

More generally, let us consider such a statistic $T(A)$ for any affine set of the form $A = (a_0, b_0) \oplus H$ where H is a linear subspace of masks and $(a_0, b_0) \notin H$. Moreover, we assume that A does not contain trivial masks. Let Λ be the multidimensional linear approximation defined by the linear space of masks $L = \{(0, 0), (a_0, b_0)\} \oplus H$. Let Λ' the multidimensional linear approximation defined by H and $\Lambda' = U \oplus VW$ be its presentation in the form (13). We define the affine test statistic as follows

$$T(A) = 2^n \sum_{(a,b)\in A} \mathrm{cor}(a, b)^2 = T(\Lambda) - T(\Lambda'). \tag{17}$$

We denote the dimension of Λ by t. Hence we can express Λ as $\Lambda = (f_0, \Lambda')$, where f_0 is the Boolean function $f_0(x) = a_0 \cdot x + b_0 \cdot P(x)$. Then the values of Λ are given as (δ, η), where δ is a bit and η is a $(t - 1)$-bit vector.

Since the correlations of the linear approximations are not independent, we cannot examine the distribution of $T(A)$ directly from its expression as a sum of squared correlations. We can, however, do this if instead we express $T(A)$ in terms of value distribution $p_{(\delta,\eta)}$ of Λ as given by the following lemma.

Lemma 4. *In the setting defined above, we have*

$$T(A) \quad = \quad 2^n 2^{t-1} \sum_{\eta} (p_{(1,\eta)} - p_{(0,\eta)})^2. \tag{18}$$

Proof. By applying (14) to $T(\Lambda')$ we obtain

$$\begin{aligned}
&2^n 2^{t-1} \sum_{\eta} (p_{(1,\eta)} - p_{(0,\eta)})^2 + T(\Lambda') \\
&= 2^n 2^{t-1} \sum_{\eta} \left(p_{(1,\eta)} - 2^{-t}) - (p_{(0,\eta)} - 2^{-t})\right)^2 \\
&+2^n 2^{t-1} \sum_{\eta} \left(p_{(1,\eta)} + (p_{(0,\eta)} - 2^{-(t-1)})\right)^2 \\
&= 2^n 2^{t-1} \sum_{\eta} \left(p_{(1,\eta)} - 2^{-t}) - (p_{(0,\eta)} - 2^{-t})\right)^2 \\
&+2^n 2^{t-1} \sum_{\eta} \left(p_{(1,\eta)} - 2^{-t}) + (p_{(0,\eta)} - 2^{-t})\right)^2,
\end{aligned}$$

which simplifies to the expression of $T(\Lambda)$ given by Equation (14). □

To compute $T(A)$ according to Equation (18) for a random permutation, all n-bit inputs x are distributed to 2^{t-1} categories according to the value η of $\Lambda'(x)$. Further, within each category the inputs x are divided into two subsets according to their value $f_0(x)$. The resulting value in category η is the difference of the sizes of its two subsets.

Since the expected distribution of the values (δ, η) of Λ is uniform, the expected value of the differences $p_{(1,\eta)} - p_{(0,\eta)}$ is zero. Hence we propose to use Pearson's χ^2 test for the values obtained in this way in 2^{t-1} categories. The related χ^2 test statistic is $T(A)$.

To determine the degree of freedom of $T(A)$, we observe that, taken together, the 2^{t-1} variables $p_{(1,\eta)} + p_{(0,\eta)}$ and the 2^{t-1} variables $p_{(1,\eta)} - p_{(0,\eta)}$, where η is a $t-1$-bit vector, uniquely determine the value distribution of Λ with probabilities $p_{\delta,\eta}$, where δ is a bit and η is a $t-1$-bit vector, which by Conjecture 1 has $2^t - 2^u - 2^v + 1$ free variables. Since the masks in $U \oplus V$ (if any) belong also to the multidimensional linear approximation Λ', the value distribution of Λ' has $2^{t-1} - 2^u - 2^v + 1$ free variables, also by Conjecture 1. Since $T(A) + T(\Lambda') = T(\Lambda)$, it follows that $T(A)$ must have at least 2^{t-1} degrees of freedom. On the other hand, by its expression (18) $T(A)$ has at most 2^{t-1} degrees of freedom, and hence exactly 2^{t-1} degrees of freedom.

We conclude that under Theorem 4 and Conjecture 1 for random permutations, $T(A)$ is χ^2 distributed with 2^{t-1} degrees of freedom and summarize the result as follows.

Theorem 6. *Let $A = (a_0, b_0) \oplus H$ be an affine subspace of linear approximations of a random permutation such that it does not contain any trivial linear approximations and assume that the multidimensional linear approximations defined by the linear spaces H and $L = \{(0,0), (a_0, b_0)\} \oplus H$ satisfy Conjecture 1. Then the statistic*

$$T(A) = 2^n \mathrm{Cap}(A) = 2^n \sum_{(a,b) \in A} \mathrm{cor}(a,b)^2$$

follows χ^2 distribution with $|A|$ degrees of freedom.

7 Data Sampling

A linear attack can be seen as composed of two parts: (i) finding an approximation with good correlation and (ii) detecting this correlation in a collection of input-output pairs.

When viewed this way, linear cryptanalysis is mainly a parameter estimation problem and the influence of data sampling is only on the second part. The distribution of the correlation over the keys is determined by the structure of the block cipher.

Using a sample of the codebook introduces an error to this parameter estimation problem. The empirical correlation is therefore a random variable in the key and the choice of the sample of plaintexts

In Sections 5 and 6 we presented the probability distributions of correlations and capacities computed over the full codebook for random functions and permutations. The goal of this section is to integrate a random variate data sample of fixed size to these probability distributions.

7.1 Sampling single linear approximation

In classical studies on linear cryptanalysis the plaintext is assumed to be drawn randomly and independently, which implies sampling with replacement. This convention was also adopted in [Nyb19] where the first model on affine multidimensional linear cryptanalysis was given. In this paper, we give a statistical model of affine cryptanalysis for random sampling of plaintext without replacement. There are two main reasons for doing this. First, duplicated plaintexts do not give new information. Secondly, we want our model to be valid also for large sample sizes and ultimately also for the full codebook without duplications.

Given a linear approximation (a, b) and a data sample S of size N drawn for a random function $F : \mathbb{F}_2^n \to \mathbb{F}_2^s$, let us denote by $\widehat{w}(a, b)$ the number of inputs $x \in S$ for which the linear approximation $a \cdot x + b \cdot F(x)$ takes the value zero. Let $w(a, b)$ be the number of zeros of $a \cdot x + b \cdot F(x)$ over all inputs $x \in \mathbb{F}_2^n$. Then

$$\widehat{w}(a, b) \sim \mathcal{HG}(2^n, w(a, b), N).$$

For a random F, we know by Theorem 2 that for $b \neq 0$

$$w = w(a, b) \sim \mathcal{B}(2^n, 1/2).$$

Then the distribution of $\widehat{w}(a, b)$ taken over random permutations and random data samples S of size N has the following probability distribution

$$\begin{aligned}\Pr(\widehat{w}(a, b) = k) &= \sum_{w=0}^{2^n} \left(\frac{1}{2}\right)^{2^n} \binom{2^n}{w} \frac{\binom{w}{k}\binom{2^n - w}{N - k}}{\binom{2^n}{N}} \\ &= (\frac{1}{2})^N \binom{N}{k}.\end{aligned}$$

Let us denote by $\widehat{\mathrm{cor}}(a, b)$ the sampled correlation, that is,

$$\widehat{\mathrm{cor}}(a, b) = \frac{1}{N}(2\widehat{w}(a, b) - N).$$

We have shown that $\widehat{w}(a, b)$ follows the binomial distribution. By normal approximation, we obtain the following result.

Theorem 7. *Let $\widehat{\mathrm{cor}}(a, b)$ be the correlation of the sampled linear approximation of a random function considered over all data samples of size N (of distinct plaintexts) and random functions from $\mathbb{F}_2^n$ to $\mathbb{F}_2^s$, where $2^n \geq N$. The probability distribution of $\widehat{\mathrm{cor}}(a, b)$ can be approximated by the normal distribution $\mathcal{N}(0, 1/N)$.*

To prove the corresponding result for random permutations we use the normal approximation from the beginning.

Theorem 8. *Let $\widehat{\mathrm{cor}}(a,b)$ be the correlation of the sampled linear approximation of a random permutation considered over all data samples of size N (of distinct plaintexts) and random permutations from $\mathbb{F}_2^n$ to $\mathbb{F}_2^n$, where $2^n \geq N$. The probability distribution of $\widehat{\mathrm{cor}}(a,b)$ can be approximated by the normal distribution $\mathcal{N}(0,1/N)$.*

Proof. Given a linear approximation (a,b) and a data sample S of size N drawn for a permutation $P : \mathbb{F}_2^n \to \mathbb{F}_2^n$, the sampled correlation is $\widehat{\mathrm{cor}}(a,b) = \frac{1}{N}(2\widehat{w}(a,b) - N)$ where $\widehat{w}(a,b) \sim \mathcal{HG}(2^n, w(a,b), N)$. Then by normal approximation

$$\widehat{\mathrm{cor}}(a,b) \sim \mathcal{N}\left(\mathrm{cor}(a,b), \frac{B}{N}(1-\mathrm{cor}(a,b)^2\right),$$

where $\mathrm{cor}(a,b) = 2^{-n}(2w(a,b) - 2^n)$. By Corollary 3 we have

$$\mathrm{cor}(a,b) \sim \mathcal{N}(0, 2^{-n}).$$

Then the distribution of $\widehat{\mathrm{cor}}(a,b)$ taken over all permutations and samples of size N is approximately normal with mean $\mathrm{Exp}\,\mathrm{cor}(a,b) = 0$ and variance equal to

$$\mathrm{Var}\left(\mathrm{cor}(a,b)\right) + \mathrm{Exp}\left(\mathrm{Var}(\widehat{\mathrm{cor}}(a,b)\right) = 2^{-n} + \frac{B}{N} - \frac{B}{N}2^{-n} \approx \frac{1}{N}.$$

□

7.2 Sampling multidimensional linear approximation

Let us now recall the sampled test statistic of a multidimensional linear approximation Λ. It is obtained by taking (14) and replacing 2^n by N and correlations $\mathrm{cor}\,(a,b)$ by sampled correlations $\widehat{\mathrm{cor}}\,(a,b)$ as follows

$$T_N(\Lambda) = N \sum_{(a,b)\in L, (a,b)\neq 0} \widehat{\mathrm{cor}}\,(a,b)^2\,. \tag{19}$$

Let us first derive the probability distribution of $T_N(\Lambda)$ for an arbitrary fixed key and randomly chosen sample of distinct plaintexts. The corresponding probability distribution for $T_N(\Lambda)$ is given by the following result originally stated in [BN17b]. The proof given in [BN17b] assumed independent hypergeometric distributions. In [BTV18] the validity of this result was questioned due to the artificial assumption of independence. Therefore another proof will be given here by applying the standard statistical argument of *finite population correction* to the χ^2 distributed variable as given by Lemma 1. In our context, the size of the population is 2^n and the size of the random sample of distinct elements from the population is N.

Theorem 9. *Let Λ be a multidimensional linear approximation of dimension t applied to a permutation of the space of n-bit vectors. Let $T_N(\Lambda)$ be the statistic defined by Equation (19) computed for a sample of size N of distinct plaintexts. Then $B^{-1}T_N(\Lambda)$ follows non-central χ^2 distribution with $2^t - 1$ degrees of freedom and non-centrality parameter $B^{-1}N\mathrm{Cap}(\Lambda)$, where B is as defined by (6).*

Proof. We denote by $\widehat{p}_\zeta$ the probabilities of the distribution of the t-bit values $\zeta \in \Lambda$ computed for a sample of size N of inputs x. We apply Equation (12) to this distribution to write $T_N(\Lambda)$ as follows

$$T_N(\Lambda) = N2^t \sum_\zeta (\widehat{p}_\zeta - 2^{-t})^2 = \sum_\zeta \frac{(N\widehat{p}_\zeta - N2^{-t})^2}{N2^{-t}}. \tag{20}$$

Then $T_N(\Lambda)$ is Pearson's χ^2-test statistic with $2^t - 1$ degrees of freedom. Since the sample is without replacement we apply Lemma 1 and get that $B^{-1}T_N(\Lambda)$ is non-centrally χ^2 distributed and has expected value equal to $2^t - 1 + \delta$, where δ is the non-centrality parameter. Then the expected value of $T_N(\Lambda)$ is equal to $B(2^t - 1) + B\delta$. To determine δ we compute the expected value of $T_N(\Lambda)$ directly. Expanding the expression (20) we get

$$T_N(\Lambda) = \sum_{\zeta} \frac{(N\widehat{p}_\zeta - Np_\zeta)^2}{N2^{-t}} \tag{21}$$

$$+ \quad N2^t \sum_{\zeta} (p_\zeta - 2^{-t})^2 \tag{22}$$

$$+ \quad N2^{t+1} \sum_{\zeta} p_\zeta(\widehat{p}_\zeta - p_\zeta), \tag{23}$$

where p_ζ is the full codebook probability of the t-bit value ζ in the image space of Λ. Note that in the expansion (21-23) the term $N2^{t+1}\sum_\zeta(\widehat{p}_\zeta - p_\zeta)$ was omitted because it is equal to zero. Now expression (21) is Pearson's χ^2-test statistic with $2^t - 1$ degrees of freedom by using the standard approximation $Np_\zeta \approx N2^{-t}$ in the denominator. Moreover it is central, since for each ζ the expected value of $\widehat{p}_\zeta$ is equal to p_ζ. Since the sampling is without replacement, we get that the expected value of (21) is equal to $B(2^t - 1)$. The expression (22) is constant and equal to $N\mathrm{Cap}(\Lambda)$, and the expected value of (23) is equal to zero. Solving δ from the equation

$$B(2^t - 1) + B\delta = B(2^t - 1) + N\mathrm{Cap}(\Lambda)$$

gives the non-centrality parameter as claimed. □

As the sample size N grows, the sampled statistic $T_N(\Lambda)$ approaches the full-codebook statistic $T(\Lambda)$. In general, the χ^2-variables computed for the full-codebook may not have the same degree of freedom as we saw in Subsection 5.2, which complicates the analysis of the joint distribution of the statistic $T_N(\Lambda)$ considered over random permutations and random samples of size N. In the case when Λ does not contain any trivial approximations we get the following result. The proof is similar to the proof of Corollary 4 and is omitted here.

Theorem 10. *Let Λ be a Davies-Mayer approximation applied to a random permutation of the space of n-bit vectors. Let $T_N(\Lambda)$ be the statistic defined by Equation (19) computed for a sample of size N of distinct plaintexts and considered as a random variable over the random permutations and random samples of size N. Then the mean of $T_N(\Lambda)$ is $|\Lambda| - 1$ and the variance is $2\,(|\Lambda| - 1)$.*

We have seen that constructions of multidimensional and affine linear approximations that do not contain any trivial approximations have a simple and clear theory for random permutations. Also for approximations that contain trivial approximations it is quite straightforward to derive the mean and the variance of the sampled statistic. For permutations originating from ciphers the theory is not that clear. The least one can say is that linear approximations of permutation ciphers have the same trivial linear approximations as for random permutations. The problem of trivial approximations was observed also in [HCN09] where it was recommended to exclude them in the computation of the empirical correlation. While this helps in speeding up the cryptanalysis, the problem of accuracy still remains. By constructions that do not contain trivial approximations, the degrees of freedom of the χ^2 distribution originate only from linear approximations with potentially significant correlations. In the case of [HCN09] the trivial linear approximations could have been easily excluded by considering the related affine set as discussed in Subsection 5.5.

7.3 Sampling affine approximation

Given an affine subspace A of linear approximations defined by two multidimensional linear approximations Λ and Λ' of dimensions t and $t-1$ respectively, we define the sampled test statistic $T_N(A)$ analogically to (17) as follows

$$T_N(A) = N \sum_{(a,b)\in A} \widehat{\mathrm{cor}}\,(a,b)^2 = T_N(\Lambda) - T_N(\Lambda'). \tag{24}$$

By repeating the derivations of Section 6, but now for the sampled statistic $T_N(A)$ and using Theorem 9 we get the following result.

Theorem 11. *Let A be an affine set of linear approximations applied to a permutation of the space of n-bit vectors and assume it does not contain trivial approximations. Let $T_N(A)$ be the statistic defined by Equation (24) computed for a sample of size N of distinct plaintexts. Then $B^{-1}T_N(A)$ follows the non-central χ^2 distribution with $|A|$ degrees of freedom and non-centrality parameter $B^{-1}N\mathrm{Cap}(A)$, where B is as defined by (6).*

The noncentrality parameter $B^{-1}N\mathrm{Cap}(A)$ of the distribution of $T_N(A)$ depends on the permutation that is used to compute the outputs for the linear approximations in A. If the permutation is chosen randomly from the set of all permutations of n-bit vectors, the distribution of $T(A) = 2^n\mathrm{Cap}(A)$ is given by Theorem 6. We get the following corollary.

Corollary 4. *Let A be an affine set of linear approximations applied to a random permutation of the space of n-bit vectors and let us assume that A does not contain trivial approximations. Let $T_N(A)$ be the statistic defined by Equation (24) computed for a sample of size N of distinct plaintexts and considered as a random variable over the random permutations and random samples of size N. Then the mean of $T_N(A)$ is $|A|$ and the variance is $2|A|$.*

Proof. Using the non-central χ^2 distribution of $B^{-1}T_N(A)$ for a fixed permutation with capacity $\mathrm{Cap}(A)$ given by Theorem 11, we get that the mean of $T_N(A)$ is equal to

$$B|A| + N\mathrm{Cap}(A) = B|A| + N2^{-n}T(A). \tag{25}$$

By taking the mean over random permutations, we get the mean $|A|$ as claimed.

Similarly, by Theorem 11, we get that the variance of $T_N(A)$ is equal to

$$B^2\left(2|A| + 4B^{-1}N\mathrm{Cap}(A)\right). \tag{26}$$

Then the total variance over random permutation is computed as the sum of the mean of (26) and the variance of (25) to get

$$B^2\left(2|A| + 4B^{-1}N2^{-n}|A|\right) + \left(N2^{-n}\right)^2 \cdot 2|A| = 2B^2|A| + 4B(1-B)|A| + 2(1-B)^2|A| = 2|A|.$$

□

Based on these considerations one can argue that, when considered as a random variable over random permutations and random samples of N distinct plaintexts, the test statistic $T_N(A)$ follows the $\chi^2_{|A|}$ distribution.

8 Conclusion

In this paper we presented a model which captures the statistical behavior of the capacity of multidimensional linear approximations computed for a permutation and a sample of plaintext, when the permutation and the sample of distinct plaintext of fixed size are

selected uniformly at random. The additivity of the variances of squared correlations is achieved without any assumptions of statistical independence based only on standard statistical tools such as Pearson's χ^2 test and the finite population correction coefficient.

We showed for the first time that the degree of freedom of the related χ^2 distribution over the distribution depends on the structure of the multidimensional linear approximation and that it can be significantly smaller than assumed in previous works due to the existence of trivial approximations. We identify two types of sets of multiple linear approximations, the Davies-Mayer approximation and the affine approximation which do not have trivial approximations. Such types of approximations offer the most efficient χ^2-based linear attacks due to optimal degrees of freedom. When selecting sets of strong multiple linear approximations for actual ciphers such structures are recommended for consideration if possible. As an example, we mention the first multidimensional linear cryptanalysis on Serpent where restricting to an affine set of approximations could potentially improve the attack.

References

[BN17a] Céline Blondeau and Kaisa Nyberg. Improved parameter estimates for correlation and capacity deviates in linear cryptanalysis. *IACR Transactions on Symmetric Cryptology*, 2016(2):162–191, 2017.

[BN17b] Céline Blondeau and Kaisa Nyberg. Joint data and key distribution of simple, multiple, and multidimensional linear cryptanalysis test statistic and its impact to data complexity. *Des. Codes Cryptography*, 82(1-2):319–349, 2017.

[BR14] Andrey Bogdanov and Vincent Rijmen. Linear hulls with correlation zero and linear cryptanalysis of block ciphers. *Designs, Codes and Cryptography*, 70(3):369–383, 2014.

[BT13] Andrey Bogdanov and Elmar Tischhauser. On the wrong key randomisation and key equivalence hypotheses in Matsui's Algorithm 2. In Moriai [Mor14], pages 19–38.

[BTV16] Andrey Bogdanov, Elmar Tischhauser, and Philip S. Vejre. Multivariate Linear Cryptanalysis: The Past and Future of PRESENT. *IACR Cryptology ePrint Archive*, 2016:667, 5 Jul 2016. http://eprint.iacr.org/2016/667.

[BTV18] Andrey Bogdanov, Elmar Tischhauser, and Philip S. Vejre. Multivariate profiling of hulls for linear cryptanalysis. *IACR Transactions on Symmetric Cryptology*, 2018(1), 2018.

[DR07] Joan Daemen and Vincent Rijmen. Probability distributions of correlation and differentials in block ciphers. *J. Mathematical Cryptology*, 1(3):221–242, 2007.

[Dun09] Orr Dunkelman, editor. *Fast Software Encryption, 16th International Workshop, FSE 2009, Leuven, Belgium, February 22-25, 2009, Revised Selected Papers*, volume 5665 of *Lecture Notes in Computer Science*. Springer, 2009.

[HCN09] Miia Hermelin, Joo Yeon Cho, and Kaisa Nyberg. Multidimensional extension of Matsui's Algorithm 2. In Dunkelman [Dun09], pages 209–227.

[HCN19] Miia Hermelin, Joo Yeon Cho, and Kaisa Nyberg. Multidimensional linear cryptanalysis. *J. Cryptology*, 32(1):1–34, 2019.

[Lea10] Gregor Leander. Small scale variants of the block cipher PRESENT. *IACR Cryptology ePrint Archive*, 2010:143, 2010.

[Mat94] Mitsuru Matsui. The First Experimental Cryptanalysis of the Data Encryption Standard. In Yvo Desmedt, editor, *Advances in Cryptology - CRYPTO '94, 14th Annual International Cryptology Conference, Santa Barbara, California, USA, August 21-25, 1994*, volume 839 of *Lecture Notes in Computer Science*, pages 1–11. Springer, 1994.

[Mor14] Shiho Moriai, editor. *Fast Software Encryption - 20th International Workshop, FSE 2013, Singapore, March 11-13, 2013. Revised Selected Papers*, volume 8424 of *Lecture Notes in Computer Science*. Springer, 2014.

[Nyb19] Kaisa Nyberg. Affine linear cryptanalysis. *Cryptography and Communications*, 11(3):367–377, 2019.

[O'C95] Luke O'Connor. Properties of linear approximation tables. In Bart Preneel, editor, *FSE 1994*, volume 1008 of *LNCS*, pages 131–136. Springer, Heidelberg, 1995.

[RT89] J. N. K. Rao and D. R. Thomas. Chi-squared tests for contingency tables. In C. J. Skinner, D. Holt, and T. M. F. Smith, editors, *Analysis of Complex Surveys*, pages 89–104. John Wiley & Sons, Chichester, UK, 1989.

A Proof of Theorem 1

Lemma 5. *Multinomial variables x_i and x_j and some of their first powers and products have the following expected values taken over the multinomial distribution.*

$$\begin{aligned}
\mathrm{Exp}(x_i) &= mp_i \\
\mathrm{Exp}(x_i x_j) &= m(m-1)p_i p_j \\
\mathrm{Exp}(x_i^2) &= mp_i + m(m-1)p_i^2 \\
\mathrm{Exp}(x_i^2 x_j) &= m(m-1)p_i p_j + m(m-1)(m-2)p_i^2 p_j \\
\mathrm{Exp}(x_i^3) &= mp_i + 3m(m-1)p_i^2 + m(m-1)(m-2)p_i^3 \\
\mathrm{Exp}(x_i x_j x_\ell x_t) &= m(m-1)(m-2)(m-3)p_i p_j p_\ell p_t \\
\mathrm{Exp}(x_i^2 x_j x_\ell) &= m(m-1)(m-2)p_i p_j p_\ell + m(m-1)(m-2)(m-3)p_i^2 p_j p_\ell \\
\mathrm{Exp}(x_i^2 x_j^2) &= m(m-1)p_i p_j + m(m-1)(m-2)(p_i^2 p_j + p_i p_j^2) \\
&+ m(m-1)(m-2)(m-3)p_i^2 p_j^2 \\
\mathrm{Exp}(x_i^3 x_j) &= m(m-1)p_i p_j + 3m(m-1)(m-2)p_i^2 p_j \\
&+ m(m-1)(m-2)(m-3)p_i^3 p_j \\
\mathrm{Exp}(x_i^4) &= mp_i + 7m(m-1)p_i^2 + 6m(m-1)(m-2)p_i^3 \\
&+ m(m-1)(m-2)(m-3)p_i^4
\end{aligned}$$

Next we give the proof of Theorem 1.

Proof. Let us start by writing the capacity C in the form

$$C = \frac{k}{m^2}\sum_{i=1}^{k}(x_i - \frac{m}{k})^2 = \frac{k}{m^2}\sum_{i=1}^{k} x_i^2 - 1.$$

To compute the variance of the capacity it suffices to do it for the sum $\sum_{i=1}^{k} x_i^2$. We write

$$\operatorname{Var} \sum_i x_i^2 = \operatorname{Exp} \left(\sum_i x_i^2 \right)^2 - \left(\operatorname{Exp} \sum_i x_i^2 \right)^2 \tag{27}$$

$$= \operatorname{Exp} \left(\sum_i x_i^4 \right) \tag{28}$$

$$+ \operatorname{Exp} \left(\sum_i \sum_{j \neq i} x_i^2 x_j^2 \right) \tag{29}$$

$$- \left(\operatorname{Exp} \sum_i x_i^2 \right)^2 . \tag{30}$$

By Lemma 5 (28) can be expressed as

$$\frac{1}{m^3} + \frac{7(m-1)}{m^3} P_2 + \frac{6(m-1)(m-2)}{m^3} P_3 + \frac{(m-1)(m-2)(m-3)}{m^3} P_4,$$

where we have denoted

$$P_4 = \sum_{i=1}^{k} p_i^4.$$

Similarly, (29) can be expressed as

$$\frac{m-1}{m^3} + \frac{(2m-5)(m-1)}{m^3} P_2 + \frac{(m-1)(m-2)(m-3)}{m^3} P_2^2 - \frac{2(m-1)(m-2)}{m^3} P_3 - \frac{(m-1)(m-2)(m-3)}{m^3} P_4,$$

and (30) as

$$\frac{1}{m^2} + \frac{2(m-1)}{m^2} P_2 + \frac{(m-1)^2}{m^2} P_2^2.$$

By combining these expressions, we get the claimed result. The derivation of the mean is similar, but simpler. □

B Number of zeroes of linear approximation of random Boolean function

.

Proof. The number of zeros of $a \cdot x + f(x)$ can be written as

$$\#\{x \in \mathbb{F}_2^n \mid f(x) = 0,\, a \cdot x = 0\} + \#\{x \in \mathbb{F}_2^n \mid f(x) = 1,\, a \cdot x = 1\},$$

where by Lemma 3, the number $z = \#\{x \in \mathbb{F}_2^n \mid f(x) = 0,\, a \cdot x = 0\}$ follows $\mathcal{HG}(2^n, 2^{n-1}, N_0)$. As f varies over all Boolean function, the number of zeros N_0 follows the binomial distribution $\mathcal{B}(2^n, \frac{1}{2})$. That is

$$\Pr(N_0 = w) = (\frac{1}{2})^{2^n} \begin{pmatrix} 2^n \\ w \end{pmatrix}.$$

Then

$$\Pr(z = k) = \sum_w \Pr(N_0 = w) \Pr(z = k \mid N_0 = w),$$

where the bounds for w and k are as follows

$$k \leq w \leq 2^{n-1} + k \text{ and } 0 \leq k \leq 2^{n-1}.$$

We get

$$\begin{aligned} \Pr(z = k) &= \left(\frac{1}{2}\right)^{2^n} \left(\begin{array}{c} 2^{n-1} \\ k \end{array} \right) \sum_{w=k}^{2^{n-1}+k} \left(\begin{array}{c} 2^{n-1} \\ w-k \end{array} \right) \\ &= \left(\frac{1}{2}\right)^{2^{n-1}} \left(\begin{array}{c} 2^{n-1} \\ k \end{array} \right), \end{aligned}$$

that is, $z \sim \mathcal{B}(2^{n-1}, \frac{1}{2})$. Similarly, it can be shown that

$$\#\{x \in \mathbb{F}_2^n \mid f(x) = 1,\, a \cdot x = 1\} \sim \mathcal{B}(2^{n-1}, \frac{1}{2}).$$

As the sum of two $\mathcal{B}(2^{n-1}, \frac{1}{2})$-variables, the number of zeros of the linear approximation $a \cdot x + f(x)$ follows $\mathcal{B}(2^n, \frac{1}{2})$. □