Multiple Differential Cryptanalysis: A Rigorous Analysis

Subhabrata Samajder and Palash Sarkar
Applied Statistics Unit
Indian Statistical Institute
203, B.T.Road, Kolkata, India - 700108.
{subhabrata_r,palash}@isical.ac.in

April 22, 2016

Abstract

Statistical analysis of multiple differential attacks are considered in this paper. Following the work of Blondeau and Gérard, the most general situation of multiple differential attack where there are no restrictions on the set of differentials is studied. We obtain closed form expressions for the data complexity in terms of the success probability and the advantage of an attack. This is done under two scenarios – one, where an independence assumption used by Blondeau and Gérard is assumed to hold and second, where no such assumption is made. The first case employs the Chernoff bounds while the second case uses the Azuma-Hoeffding bounds from the theory of martingales. In both cases, we do not make use of any approximations in our analysis. As a consequence, the results are more generally applicable compared to previous works. The analysis without the independence assumption is the first of its kind in the literature. We believe that the current work places the statistical analysis of multiple differential attack on a more rigourous foundation than what was previously known.

Keywords: multiple differential cryptanalysis, Chernoff bounds, martingales, Azuma-Hoeffding bounds.

1 Introduction

One of the basic techniques for attacking a block cipher is differential cryptanalysis [3]. In its basic form, it considers the difference after several rounds of the encryption function to two plaintexts which themselves which differ by a fixed string. A pair of such input and output differences is called a differential. Initially, differential cryptanalysis was considered with respect to a single differential. Later work considered several differentials where either all the input differences are the same or all the output differences are the same. In its most general form, there are multiple differentials with no restrictions on either the input or the output differences.

Cryptanalysis of block ciphers has two conceptual phases. The first phase consists of a detailed study of the structure of the block cipher to discover some property which can be distinguished from randomness. Obtaining one or more differentials is one such property that a cryptanalyst looks for. The second phase involves using statistical methods to exploit such a property for deriving a portion of the secret key. Over the years, the second phase has received increasing focus [15, 12, 13, 14, 1, 5, 23, 11, 8, 6, 7].

To mount an attack, a cryptanalyst requires a number of plaintexts and the corresponding ciphertexts encrypted with the same secret key. In a key recovery attack, the goal is to obtain the correct value for a subset of the key bits. This subset is called the target sub-key. Suppose m bits of the key are to be recovered. Processing the obtained plaintext-ciphertext pairs will provide the cryptanalyst with a list of candidate values for the target sub-key. The attack is successful if the correct value is in the list of candidate values. The other parameter of interest is the size of the candidate list. If for some a, this list is of size 2^{m-a}, then the attack is said to have advantage at least a. A statistical analysis tries to determine the number N of plaintext-ciphertext pairs such that the attack is successful with probability at least P_S and has advantage a. The parameter N is called the data complexity of the attack. The goal of a statistical analysis is to obtain an expression for N in terms of P_S and a.

Statistical analysis of the most general form of multiple differential cryptanalysis was considered by Blondeau and Gérard [6]. Building on prior work [8], the paper provided two separate expressions for the data complexity and the success probability. The expression for the data complexity holds for success probability close to 0.5 and so for example, it cannot be applied to attack with success probability close to 1 (say 0.95). The expression for success probability, on the other hand, is complicated and does not directly involve the data complexity. Further, the analysis used several approximations involving the Poisson distributions and also required the approximations of tails of binomial distributions used in an earlier work [8]. The errors in such approximations, however, have not been rigorously analysed. Another feature of the analysis in [6] is that it is based on an independence assumption. Whether the assumption holds in general is not clear and the authors remark [6]: “This hypothesis is not so far to being true.”

Our Contributions

As in [6], the setting of our work is multiple differential cryptanalysis without any restrictions on the input or the output differences. We perform a new statistical analysis of this kind of attack. This analysis involves using the Chernoff bounds and the theory of martingales. These techniques have been earlier used in the context of block cipher cryptanalysis [22].

We build the statistical analysis in several steps. The first step considers the simpler scenario where there is only one single input difference while there can be several output differences. In the second step, we extend this to the general scenario of multiple differentials without any restriction on the input or the output differences. This step, however, utilises the independence assumption used in [6]. In the third step, we do away with the independence assumption. So, the analysis of the third step is the most general. The analysis of the first two steps uses the Chernoff bounds while the third step is based on the theory of martingales. There are two common features to all the three steps of our analysis.

1. Nowhere do we make any kind of approximation. So, the analysis holds for all settings.
2. In each case, we obtain explicit closed form expressions for the data complexity in terms of the success probability and the advantage. These expressions can be evaluated to obtain the data complexity for any values of the success probability and the advantage.

A set of multiple differentials were provided for a 64-bit toy cipher SMALLPRESENT in [6]. For these differentials, we compare the concrete values of the data complexities provided by our analysis with earlier works [6, 7]. It turns out that in all cases, the data complexities obtained by our methods is higher than what has been previously reported. For the analysis with the independence assumption, the data complexities are close, while for the analysis without the independence assumption, the new data complexity turns out to be significantly higher. There are two take-aways from these experimental results. First, if a cryptanalyst does not wish to rely on approximations which have not been rigorously analysed, then one will have to satisfied with the conservative estimates of the data complexity. Second, if one does not want to rely on a somewhat ad-hoc independence assumption, then one will have to be prepared for handling a much higher data complexity.

Previous and Related Works: Differential cryptanalysis was first proposed by Biham and Shamir in [3] for cryptanalysis of DES. Later in [4], the same authors improved upon the earlier version by considering multiple differentials with the same output difference. Knudsen [16] introduced truncated differential cryptanalysis. Other variants of differential cryptanalysis have been proposed. These include higher order differentials [17], cube attack [9], boomerang attack [25], impossible differential attack [2] and the improbable differential attack [24].

A statistical analysis of multiple differential attack with a single input difference was given in [20]. Selçuk [23] derived an expression for the data complexity of single differential cryptanalysis using the ranking methodology. The technique used by Selçuk was subsequently used by Blondeau et al. in [7] to derive data complexity of differential cryptanalysis using the log-likelihood (LLR) and chi-squared test statistic. This work considered differentials with a single input difference.

As mentioned earlier, the most general framework for differential cryptanalysis was considered in [6], where differentials were considered without any restrictions. The work proposed a new test statistic and showed that the distribution of the test statistic can be approximated by a Poisson distribution. It was subsequently pointed out that the Poisson approximation is not good for the tail probabilities and hence the technique of [8] was used to approximate the tail probabilities.

The task of deriving data complexity expressions without using approximations was carried out in [22] for several types of block cipher attacks. Chernoff bounds and the theory of martingales were used for this purpose. The present work employs these techniques to analyse the tail probabilities of the test statistic proposed in [6]. This leads to the aforementioned results on data complexities obtained here. For problematic issues regarding the use of approximations in cryptanalysis we refer the reader to [21].

2 Background

Let $E : \{0,1\}^k \times \{0,1\}^n \mapsto \{0,1\}^n$ be a block cipher so that for each $K \in \{0,1\}^k$, the function $E_K(\cdot) \overset{\Delta}{=} E(K,\cdot)$ is a bijection. Here K is called the secret key, the n-bit input to E_K is called the plaintext and the n-bit output of E_K is called the ciphertext.

We consider iterated block ciphers which are constructed by composing round functions. Let $R^{(0)}_{k^{(0)}}, R^{(1)}_{k^{(1)}}, \ldots$ be the round functions, where $k^{(0)}, k^{(1)}, \ldots$ denote the round keys. These round keys are produced by applying an expansion function on the secret key K, called the key scheduling algorithm. For a fixed key, the round functions are also bijections.

Denote by $K^{(i)}$ the concatenation of the first i round keys, i.e., $K^{(i)} = k^{(0)} \mid\mid \cdots \mid\mid k^{(i-1)}$ and by $E^{(i)}_{K^{(i)}}$ the composition of the first i round functions, i.e.,

$$E^{(1)}_{K^{(1)}} = R^{(0)}_{k^{(0)}}; \quad E^{(i)}_{K^{(i)}} = R^{(i-1)}_{k^{(i-1)}} \circ \cdots \circ R^{(0)}_{k^{(0)}} = R^{(i-1)}_{k^{(i-1)}} \circ E^{(i-1)}_{K^{(i-1)}}; \, i \geq 1.$$

Consider an attack on the first $(r+1)$ round of the block cipher E_K. For a plaintext P, denote by B the output after r rounds, i.e., $B = E^{(r)}_{K^{(r)}}(P)$ and by C the output after $r+1$ rounds, i.e., $C = E^{(r+1)}_{K^{(r+1)}} = R^{(r)}_{k^{(r)}}(B)$.

2.1 Differential Cryptanalysis

Let δ_0 and δ_r be n-bit strings where δ_0 is not the all-zero string. For a plaintext P, denote $P' = P \oplus \delta_0$. Since $\delta_0 \neq 0^n$, $P' \neq P$. Let B and B' denote the output after r rounds corresponding to P and P' respectively. From the bijectivity of the round functions, it follows that $B \neq B'$. For a fixed K, the quantities P', B and B' are completely determined by P.

Let P be chosen uniformly at random and let p be such that $p = \Pr[B \oplus B' = \delta_r]$. On the other hand, if B and B' are chosen uniformly and without replacement from $\{0,1\}^n$, then $\Pr[B \oplus B' = \delta_r] = 1/(2^n - 1)$. Let $p_w = 1/(2^n - 1)$. The first and a non-trivial step of a differential cryptanalysis is to make a detailed study of the block cipher to unearth δ_0 and δ_r such that p is 'significantly' different from p_w.

Target sub-key: Suppose there are m bits of the key such that knowledge of these m bits is sufficient to invert the last round and obtain B from C. These m bits could be a subset of bits of the last round key. If it turns out that $m < n$, then a differential cryptanalysis can be attempted. We will call this set of m bits as the target sub-key. There are 2^m possible choices of the target sub-key out of which only one is correct. We will denote the correct choice of the target sub-key as κ^*. The goal of the attack is to find κ^*.

The attack will proceed by testing each possible value of the target sub-key. If the choice of the target sub-key is correct, then $\Pr[B \oplus B' = \delta_r]$ will be equal to p. On the other hand, if the choice of the target sub-key is incorrect, then it is conventional to assume that the block cipher behaves like a random permutation and so in this case, $\Pr[B \oplus B' = \delta_r]$ will be equal to p_w.

Multiple differentials: We follow [6] in considering multiple differentials. Our notation, though, is somewhat different.

In the above, we have considered a single pair (δ_0, δ_r). More generally, one can consider a set Δ of such pairs. An attack which aims to utilise such a set of differentials is called a multiple differential attack. Let $\nu = |\Delta|$. Define $\Delta_0 = \{\delta_0 :$ there is a δ_r such that $(\delta_0, \delta_r) \in \Delta\}$. In other words, Δ_0 consists of the set of all n-bit strings which occur as the first component of some pair in Δ. These are all the distinct input differences. Let $\nu_0 = |\Delta_0|$ and enumerate the input differences in Δ_0 as $\Delta_0 = \{\delta_0^{(1)}, \ldots, \delta_0^{(\nu_0)}\}$. For $\delta_0^{(i)}$, define $\Delta_r^{(i)}$ as $\Delta_r^{(i)} = \{\delta_r : (\delta_0^{(i)}, \delta_r) \in \Delta\}$. The set $\Delta_r^{(i)}$ consists of the set of all possible output difference corresponding to the input difference $\delta_0^{(i)}$. Let $\nu_i = |\Delta_r^{(i)}|$ and enumerate the set $\Delta_r^{(i)}$ as $\Delta_r^{(i)} = \{\delta_r^{(i,1)}, \ldots, \delta_r^{(i,\nu_i)}\}$. Then the set of differentials can be written as

$$\Delta = \{(\delta_0^{(i)}, \delta_r^{(i,j)}) \mid i = 1, \ldots, \nu_0 \text{ and } j = 1, \ldots, \nu_i\}.$$

Suppose P is chosen uniformly at random. For a particular choice κ of the target sub-key and $\delta_0^{(i)}$, let $S_{\kappa,i}$ be the following random variable.

$$S_{\kappa,i} = \left(R_\kappa^{(r)}\right)^{-1} (E_{K^{(r)}}(P)) \oplus \left(R_\kappa^{(r)}\right)^{-1} \left(E_{K^{(r)}}(P \oplus \delta_0^{(i)})\right). \tag{1}$$

Here $\left(R_\kappa^{(r)}\right)^{-1}$ refers to the inversion of the r-th round using κ as the value of the target sub-key.

Extending the probability notions from the case of a single differential, for $i = 1, \ldots, \nu_0$ and $j = 1, \ldots, \nu_i$, we define

$$\begin{array}{rcl} \Pr\left[S_{\kappa,i} = \delta_r^{(i,j)}\right] & = & \left\{ \begin{array}{ll} p_{i,j} & \text{if } \kappa = \kappa^*; \\ q_{i,j} & \text{otherwise;} \end{array} \right. \\ q_{i,j} & = & 1/(2^n - 1); \\ p_i & = & \sum_{j=1}^{\nu_i} p_{i,j}; \\ \hat{p} & = & \left(\sum_{i=1}^{\nu_0} p_i\right)/\nu_0; \\ q_i & = & \sum_{j=1}^{\nu_i} q_{i,j}; \\ \hat{q} & = & \left(\sum_{i=1}^{\nu_0} q_i\right)/\nu_0. \end{array} \tag{2}$$

We note that the analysis of the block cipher will provide the set Δ and also the probabilities $p_{i,j}$ for $i = 1, \ldots, \nu_0$ and $j = 1, \ldots, \nu_i$. These form the starting point of a statistical analysis.

Setting of the attack: Suppose that $P_1, \ldots, P_N$ are chosen independently and uniformly at random from $\{0,1\}^n$. For each $\eta = 1, \ldots, N$, the adversary obtains the ciphertext C_η corresponding to the encryption of P_η under some key K. Further, for each input difference $\delta_0^{(i)} \in \Delta_0$ the adversary obtains the encryption of $P \oplus \delta_0^{(i)}$ under the same key K. So, in total the adversary has $(\nu_0 + 1)N$ pairs of plaintext and ciphertexts. Therefore,

N denotes the number of independently and uniformly chosen plaintexts whereas the number of encryptions required is $(\nu_0 + 1)N$. By data complexity we will denote N.

Recall that κ^* is the correct choice of the target sub-key corresponding to K. The goal of the adversary is to obtain κ^*.

We extend the definition of $S_{\kappa,i}$ in the following manner. For a choice κ of the target sub-key, $i = 1, \ldots, \nu_0$ and $\eta = 1, \ldots, N$, define

$$S_{\kappa,i,\eta} = \left(R_\kappa^{(r)}\right)^{-1} (E_{K^{(r)}}(P_\eta)) \oplus \left(R_\kappa^{(r)}\right)^{-1} \left(E_{K^{(r)}}(P_\eta \oplus \delta_0^{(i)})\right). \tag{3}$$

Since $P_1, \ldots, P_N$ are independent, the random variables $S_{\kappa,i,1}, \ldots, S_{\kappa,i,N}$ are also independent and each of these is distributed as $S_{\kappa,i}$.

Define a binary valued random variable $T_{\kappa,i,\eta}$ as follows: $T_{\kappa,i,\eta} = 1$ if $S_{\kappa,i,\eta}$ is in $\Delta_r^{(i)}$ and it is 0 otherwise. It follows that $T_{\kappa,i,1}, \ldots, T_{\kappa,i,N}$ are also independent. For a fixed κ, i and η, the random variable $S_{\kappa,i,\eta}$ can take at most one value in $\Delta_r^{(i)}$. As a result,

$$\Pr[T_{\kappa,i,\eta} = 1] = \begin{cases} \sum_{\eta=1}^{\nu_i} p_{i,j} = p_i & \text{if } \kappa = \kappa^*; \\ \nu_i/(2^n - 1) = q_i & \text{if } \kappa \neq \kappa^*. \end{cases} \tag{4}$$

From the $(\nu_0 + 1)N$ plaintext-ciphertext pairs; for $i = 1, \ldots, \nu_0$; and for each choice κ of the target sub-key; the adversary can compute the values of $T_{\kappa,i,1}, \ldots, T_{\kappa,i,N}$ in time $N\nu_0 2^m$.

The test statistic: For a choice κ of the target sub-key and $\eta = 1, \ldots, N$, define

$$T_{\kappa,\eta} = \sum_{i=1}^{\nu_0} T_{\kappa,i,\eta}. \tag{5}$$

For a choice κ of the target sub-key, the test statistic is defined to be

$$T_\kappa = \sum_{\eta=1}^{N} T_{\kappa,\eta} = \sum_{i=1}^{\nu_0} \sum_{\eta=1}^{N} T_{\kappa,i,\eta}. \tag{6}$$

Success probability and advantage of an attack: An attack will ultimately provide a list of candidate values of the target sub-key. The attack is said to be successful, if κ^* is in the list and the probability of this event is said to be the success probability. This probability is denoted as P_S. The size of the list is another factor which determines the efficacy of an attack. The attack is said to have advantage a, if the size of the list is 2^{m-a}.

The goal of a statistical analysis is to obtain an expression for the data complexity in terms of the success probability and the advantage of an attack.

2.2 Summary of the Blondeau-Gèrard Analysis

In [6], Blondeau and Gèrard provided a statistical analysis of multiple differential attack as outlined above. Their analysis used the following assumption.

Assumption 1 ([6, Hypothesis 2]). *For any sub-key κ (including κ^*) and for any $\eta = 1, \ldots, N$, the random variables $T_{\kappa,1,\eta}, \ldots, T_{\kappa,\nu_0,\eta}$ are independent.*

Based on this assumption and using an asymptotic result it was shown that each of the random variables $T_{\kappa,\eta}$ follow a Poisson distribution with parameter $\lambda = \sum_{i=0}^{\nu_0} p_i$ for $\kappa = \kappa^*$ and follows a Poisson distribution with parameter $\lambda = \sum_{i=0}^{\nu_0} q_i$ for $\kappa \neq \kappa^*$ [6, Theorem 1]. Since $T_\kappa = \sum_{\eta=1}^{N} T_{\kappa,\eta}$ and $T_{\kappa,1}, \ldots, T_{\kappa,N}$ are independent,

T_κ also follows a Poisson distribution with parameter $N\lambda$. It was further mentioned that this approximation does not give a good estimate of the tails of the cumulative distribution function of the test statistic T_κ. Hence, they used another approximation to get the tail of the distribution. Eventually, the full distributions for both the correct and the incorrect choice of the target sub-key were given. It was found that these distributions are similar to the distributions of [8]. Therefore, the framework from [8] was used to estimate the data complexity and the success probability of the multiple differential cryptanalysis. We restate the results here.

Corollary 1 of [6]. The data complexity of multiple differential cryptanalysis with success probability close to 0.5 is given by

$$N = -2 \cdot \frac{\ln(2\iota\sqrt{\pi}2^{-m})}{\nu_0 D(\hat{p} \mid\mid \hat{q})}; \tag{7}$$

where ι is the size reduced list of the candidate keys and $D(\hat{p} \mid\mid \hat{q})$ is the Kullback-Leibler divergence between the distributions $(\hat{p}, 1-\hat{p})$ and $(\hat{q}, 1-\hat{q})$. Typically, $\iota = 2^{m-a}$ for an attack with a-bit advantage.

Corollary 2 of [6]. Let $G^*(x)$ (resp. $G(x)$) be the estimate of the cumulative distribution function of T_{κ^*} (resp. , T_κ) defined by [6, Proposition 1]. The success probability, P_S, of a multiple differential cryptanalysis is given by

$$P_S \approx 1 - G^*\left(G^{-1}\left(1 - \frac{\iota - 1}{2^m - 2}\right) - 1\right); \tag{8}$$

where the pseudo-inverse of G is defined by $G^{-1}(y) = \min\{x \mid G(x) \geq y\}$.

Remark 1: Putting $\iota = 2^{m-a}$ in (7), we have

$$N \quad = \quad -2 \cdot \frac{\ln(2\sqrt{\pi}/2^a)}{\nu_0 D(\hat{p} \mid\mid \hat{q})}. \tag{9}$$

Therefore, $N > 0$ only for $a > \lg(2\pi) = 1 + \lg \pi$. So, the data complexity expression given by equation (7) is not meaningful for small values of 'a'.

3 Hypothesis Testing Framework

We briefly outline the hypothesis testing framework that we use to perform the statistical analysis of multiple differential attack.

The test statistic is T_κ. Let $\mu_0 = E[T_{\kappa^*}]$ and $\mu_1 = E[T_\kappa]$ for $\kappa \neq \kappa^*$. Assume that $\mu_0 > \mu_1$ and consider the following test of hypothesis.

Hypothesis Test-1:
H_0: "κ is correct" versus H_1: "κ is incorrect."
Decision rule: Reject H_0 if $T_\kappa \leq t$, where t is a value in (μ_1, μ_0).

Note: If $\mu_0 < \mu_1$, then the decision rule is to reject H_0 if $T_\kappa \geq t$ for some t in (μ_0, μ_1). The analysis of this case is similar to the case of $\mu_0 > \mu_1$ and provides the same expression for the data complexity. So, we do not consider the details of this case.

The Type-1 error probability is defined to be Pr[Type-1 Error] = $\Pr[T_\kappa \leq t | H_0$ holds] and the Type-2 error probability is defined to be Pr[Type-2 Error] = $\Pr[T_\kappa > t | H_1$ holds]. We obtain (upper bounds) on the Type-1 and Type-2 error probabilities which are denoted as α and β respectively. These expressions involve both t and N and it turns out to be possible to obtain expressions for t and N in terms of α and β.

The test is applied with all the 2^m values of the target sub-key and a list of values of κ for which H_0 is not rejected is returned. If a Type-1 error occurs, then the list does not contain the correct value of the target sub-key and so the success probability of the attack is $1 - \Pr[\text{Type-1 Error}]$. In particular, we set $P_S = 1 - \alpha$ so that if α is an upper bound on the Type-1 error probability, then P_S is a lower bound on the success probability.

Each Type-2 error results in classifying an incorrect value of κ as a candidate key. Since the tests with the $2^m - 1$ incorrect choices of the target sub-keys are independent, the expected number of wrong keys returned is $(2^m - 1) \times \Pr[\text{Type-2 error}]$. If β is an upper bound on $\Pr[\text{Type-2 Error}]$, then the expected number of misclassifications is at most $(2^m - 1)\beta < 2^m\beta$. For an attack with advantage a, the size of the returned list is 2^{m-a}. Setting $2^m\beta = 2^{m-a}$ gives $\beta = 2^{-a}$. So, if the Type-2 error probability is at most 2^{-a}, then the attack has expected advantage at least a.

Upper Bounds: The hypothesis test is applied to a particular test statistic. The corresponding data complexity obtained is a lower bound on the number of plaintexts to achieve specified (upper bounds on the) Type-1 and Type-2 error probabilities, *if* the particular test statistic is used. This leaves open the possibility that there may be other test statistics for which the (lower bound on the) data complexity required to achieve the same Type-1 and Type-2 error probabilities is lower. So, the expressions that we obtain are upper bounds on the minimum possible data complexities to achieve specified error probabilities.

4 Single Input Difference

In this section, we analyze the particular case of $\nu_0 = 1$, i.e., the case where all the differentials have the same input difference. The number of differentials in Δ can be 1 or more, i.e., $\nu \geq 1$, but, for all of these the input difference will be the same. This case is the same as the one studied in [7] and later in [22].

Since $\nu_0 = 1$, it follows that $\nu_1 = \nu$. So, for every $\kappa \in \{0,1\}^m$ and $\eta = 1, \ldots, N$, there is a single random variable $T_{\kappa,1,j}$. This variable takes the value 1 if $S_{\kappa,1,\eta} \in \Delta_r^{(1)}$ and 0 otherwise. As mentioned earlier, $\Pr[T_{\kappa,1,\eta} = 1] = p_1$ if $\kappa = \kappa^*$ and $\Pr[T_{\kappa,1,\eta} = 1] = q_1 = \nu/(2^n - 1)$ if $\kappa \neq \kappa^*$. Further, for any fixed κ, the random variables $T_{\kappa,1,1}, \ldots, T_{\kappa,1,N}$ are independent. Being independently distributed binary valued random variables, these can be considered the outcomes of Poisson trials.

The test statistic T_κ in this case becomes

$$T_\kappa = T_{\kappa,1,1} + \cdots + T_{\kappa,1,N}. \tag{10}$$

We define

$$\begin{aligned} \mu_0 &= E[T_\kappa] = Np_1 \text{ if } \kappa = \kappa^*; \\ \mu_1 &= E[T_\kappa] = Nq_1 \text{ if } \kappa \neq \kappa^*. \end{aligned}$$

Suppose $\mu_0 > \mu_1$. Hypothesis Test-1 is applied with the test statistic T_κ as described in Section 3. For the analysis of Hypothesis Test-1, we need to determine the Type-1 and Type-2 error probabilities. Since T_κ is the sum of independent Bernoulli distributed random variables, the Chernoff bounds can be used to obtain bounds on the tail probabilities of T_κ for both correct and incorrect choices of κ. We refer to Appendix A for the precise statement of the Chernoff bounds. We have the following result.

Proposition 1. *Suppose that $\nu_0 = 1$, i.e., there is only a single input difference. Let $0 < \alpha, \beta < 1$ and N be such that*

$$N \geq \frac{3\left(\sqrt{p_1 \ln(1/\alpha)} + \sqrt{q_1 \ln(1/\beta)}\right)^2}{(p_1 - q_1)^2}. \tag{11}$$

Then the probabilities of the Type-1 and Type-2 errors in Hypothesis Test-1 are upper bounded by α and β respectively. Putting $\alpha = 1 - P_S$ and $\beta = 2^{-a}$, it follows that for

$$N \geq \frac{3\left(\sqrt{p_1 \ln(1/(1-P_S))} + \sqrt{aq_1 \ln 2}\right)^2}{(p_1 - q_1)^2}. \tag{12}$$

the success probability will be at least P_S and the advantage will be at least a.

Proof. Recall that $t \in (\mu_1, \mu_0)$ and H_0 is rejected if $T_\kappa \leq t$. Let $\gamma_0 = 1 - t/\mu_0$ and so $\gamma_0 \in (0, 1)$.

$$\begin{aligned} \Pr[\text{Type-1 Error}] &= \Pr[T_\kappa \leq t \mid H_0 \text{ holds}] = \Pr[T_\kappa \leq (1-\gamma_0)\mu_0] \\ &\leq \exp\left(-\mu_0\gamma_0^2/2\right) \leq \exp\left(-\mu_0\gamma_0^2/3\right) = \exp\left(-\frac{(\mu_0 - t)^2}{3\mu_0}\right) = \alpha; \text{ (say)} \\ \Rightarrow t &= \mu_0 - \sqrt{3\mu_0 \ln(1/\alpha)}. \end{aligned} \tag{13}$$

Similarly, define $\gamma_1 = t/\mu_1 - 1$ so that $\gamma_1 \in (0, 1)$.

$$\begin{aligned} \Pr[\text{Type-2 Error}] &= \Pr[T_\kappa > t \mid H_1 \text{ holds}] = \Pr[T_\kappa > (1+\gamma_1)\mu_1] \\ &\leq \exp\left(-\mu_1\gamma_1^2/3\right) = \exp\left(-\frac{(t - \mu_1)^2}{3\mu_1}\right) = \beta; \text{ (say)} \\ \Rightarrow t &= \mu_1 + \sqrt{3\mu_1 \ln(1/\beta)}. \end{aligned} \tag{14}$$

Eliminating t from equations (13) and (14) and using $\mu_0 = Np_1$, $\mu_1 = Nq_1$, we obtain an expression for N which is given by the right hand side of (11). For any N greater than this value, the probabilities of Type-1 and Type-2 errors are at most α and β respectively. □

5 Multiple Input Differences Under Assumption 1

Under Assumption 1, for each κ and η, the random variables $T_{\kappa,1,j}, \ldots, T_{\kappa,\nu_0,j}$ are independent. If we further consider the values of η from 1 to N, we get that the following sequence of binary valued random variables are independent:

$$\begin{array}{cccc} T_{\kappa,1,1} & T_{\kappa,2,1} & \cdots & T_{\kappa,\nu_0,1} \\ T_{\kappa,1,2} & T_{\kappa,2,2} & \cdots & T_{\kappa,\nu_0,2} \\ \cdot & \cdot & \cdots & \cdot \\ T_{\kappa,1,N} & T_{\kappa,2,N} & \cdots & T_{\kappa,\nu_0,N}. \end{array} \tag{15}$$

So, these can be considered as the outcomes of $N\nu_0$ Poisson trials. The test statistic T_κ is the sum of the above random variables and hence the Chernoff bounds can be applied to T_κ. Let as before $\mu_0 = E[T_\kappa]$ when H_0 holds and $\mu_1 = E[T_\kappa]$ when H_1 holds. Then, we have

$$\begin{aligned} \mu_0 &= E[T_\kappa \mid H_0 \text{ holds}] \\ &= E\left[\sum_{\eta=1}^{N} \sum_{i=1}^{\nu_0} T_{\kappa,i,\eta} \mid H_0 \text{ holds}\right] \\ &= \sum_{\eta=1}^{N} \sum_{i=1}^{\nu_0} E[T_{\kappa,i,\eta} \mid H_0 \text{ holds}] \\ &= \sum_{\eta=1}^{N} \sum_{i=1}^{\nu_0} p_i = \sum_{\eta=1}^{N} \nu_0 \hat{p} = N\nu_0\hat{p}; \\ \text{Similarly, } \mu_1 &= E[T_\kappa \mid H_1 \text{ holds}] = N\nu_0\hat{q}. \end{aligned} \tag{16}$$

With these values of μ_0 and μ_1 and following the proof of Proposition 1 almost verbatim we obtain the following result.

Proposition 2. *Suppose that $\nu_0 \geq 1$ and Assumption 1 holds. Let $0 < \alpha, \beta < 1$ and N be such that*

$$N \quad \geq \quad \frac{3\left(\sqrt{\hat{p}\ln(1/\alpha)} + \sqrt{\hat{q}\ln(1/\beta)}\right)^2}{\nu_0(\hat{p}-\hat{q})^2}. \tag{17}$$

Then the probabilities of the Type-1 and Type-2 errors in Hypothesis Test-1 are upper bounded by α and β respectively. Putting $\alpha = 1 - P_S$ and $\beta = 2^{-a}$, it follows that for

$$N \quad \geq \quad \frac{3\left(\sqrt{\hat{p}\ln(1/(1-P_S))} + \sqrt{a\hat{q}\ln 2}\right)^2}{\nu_0(\hat{p}-\hat{q})^2}. \tag{18}$$

the success probability will be at least P_S and the advantage will be at least a.

For $\nu_0 = 1$, Proposition 2 reduces to Proposition 1. Note that for $\nu_0 = 1$, Assumption 1 is vacuous.

6 Multiple Input Differences Without Independence Assumption

It is not clear that Assumption 1 holds in general. In this section, we consider the problem of deriving an expression for data complexity without using Assumption 1. Without this assumption, we can no longer assume that the rows of (15) are independent and so T_κ cannot be written as the sum of outcomes of Possion trials. This, in particular, means that we cannot apply the Chernoff bounds to bound the tail probabilities of T_κ.

To tackle this situation, we use the theory of martingales as utilised in [22]. Appendix B provides a brief review of the relevant theory of martingales. The test statistic is still T_κ and Hypothesis Test-1 is applied. The analysis of Type-1 and Type-2 error probabilities change.

Proposition 3. *Suppose that $\nu_0 \geq 1$. Let $0 < \alpha, \beta < 1$ and N be such that*

$$N \quad \geq \quad \frac{2\left(\sqrt{\ln(1/\alpha)} + \sqrt{\ln(1/\beta)}\right)^2}{(\hat{p}-\hat{q})^2}. \tag{19}$$

Then the probabilities of the Type-1 and Type-2 errors in Hypothesis Test-1 are upper bounded by α and β respectively. Putting $\alpha = 1 - P_S$ and $\beta = 2^{-a}$, it follows that for

$$N \quad \geq \quad \frac{2\left(\sqrt{\ln(1/(1-P_S))} + \sqrt{a\ln 2}\right)^2}{(\hat{p}-\hat{q})^2}. \tag{20}$$

the success probability will be at least P_S and the advantage will be at least a.

Proof. Since the computation of the expectation of a sum of random variables does not depend on whether these random variables are independent, the values of $\mu_0 = E[T_\kappa | H_0 \text{ holds}]$ and $\mu_1 = E[T_\kappa | H_1 \text{ holds}]$ are still given by (16).

Recall that for $\kappa \in \{0,1\}^m$ and $1 \leq \eta \leq N$, $T_{\kappa,\eta} = \sum_{i=1}^{\nu_0} T_{\kappa,i,\eta}$. Note that $T_{\kappa,\eta}$ takes values from the set $\{0, \ldots, \nu_0\}$. Define a sequence of random variables as follows.

$$\begin{aligned} Z_0 &= E[T_{\kappa,1} + \cdots + T_{\kappa,N}] = E[T_\kappa], \\ Z_\eta &= E[T_\kappa \mid T_{\kappa,1}, \ldots, T_{\kappa,\eta}]; \quad 1 \leq \eta \leq N. \end{aligned}$$

It can be shown that the sequence of random variables $\{Z_\eta\}_{\eta=0}^N$ forms a Doob Martingale with respect to the sequence $\{T_{\kappa,\eta}\}_{\eta=1}^N$. We refer to Appendix B for more details.

Let $f(x_1,\ldots,x_N) = x_1 + \cdots + x_N$; where $x_\eta \in \{0,1,\ldots,\nu_0\}$ and $\eta = 1,\ldots,N$. Let $x_1,\ldots,x_N,x'_\eta$ be any $N+1$ elements from the set $\{0,1,\ldots,\nu_0\}$. Then,

$$|f(x_1,\ldots,x_{\eta-1},x_\eta,x_{\eta+1},\ldots,x_N) - f(x_1,\ldots,x_{\eta-1},x'_\eta,x_{\eta+1},\ldots,x_N)| = |x_\eta - x'_\eta| \leq \nu_0.$$

This shows that f is ν_0-Lipschitz.

Note that $T_\kappa = f(T_{\kappa,1},\ldots,T_{\kappa,N})$. Since f is ν_0-Lipschitz and $T_{\kappa,\eta}$'s are independent, it follows that $|Z_\eta - Z_{\eta-1}| \leq \nu_0$ for all $\eta = 1,\ldots,N$. Further, $Z_N = E[T_{\kappa,1} + \cdots + T_{\kappa,N} \mid T_{\kappa,1},\ldots,T_{\kappa,N}] = T_\kappa$; $Z_0 = \mu_0$ if H_0 holds and $Z_0 = \mu_1$ if H_1 holds.

$$\begin{aligned}
\Pr[\text{Type-1 Error}] &= \Pr[T_\kappa \leq t \mid H_0 \text{ holds}] \\
&= \Pr[Z_N - Z_0 \leq -(Z_0 - t) \mid H_0 \text{ holds}] \\
&= \Pr[Z_N - \mu_0 \leq -(\mu_0 - t)] \\
&\leq \exp\left(-\frac{(\mu_0 - t)^2}{2N\nu_0^2}\right) = \alpha \text{ (say);} \quad \text{[By Azuma-Hoeffding inequality]} \\
\Rightarrow t &= \mu_0 - \nu_0\sqrt{2N\ln(1/\alpha)}; \qquad (21) \\
\Pr[\text{Type-2 Error}] &= \Pr[T_\kappa > t \mid H_1 \text{ holds}] \\
&= \Pr[Z_N - Z_0 > t - Z_0 \mid H_1 \text{ holds}] \\
&= \Pr[Z_N - \mu_1 > (t - \mu_1)] \\
&\leq \exp\left(-\frac{(t - \mu_1)^2}{2N\nu_0^2}\right) = \beta \text{ (say);} \quad \text{[By Azuma-Hoeffding inequality]} \\
\Rightarrow t &= \mu_1 + \nu_0\sqrt{2N\ln(1/\beta)}. \qquad (22)
\end{aligned}$$

Eliminating t from equation (21) and (22) and using $\mu_0 = N\nu_0\hat{p}$, $\mu_1 = N\nu_0\hat{q}$ gives the expression in the right hand side of (19). For any N which is at least this quantity, the Type-1 and Type-2 error probabilities are at most α and β. □

7 Comparison and Experimental Results

From a theoretical angle, our analysis provides the following advantages over the previous work of Blondeau and Gèrard [6].

No use of approximations: The analysis of [6] is based on several approximations. One place where this issue shows up is in the non-applicability of the data complexity expression in (9) for small values of a as has been discussed in Remark 1. Further, the accuracy of the estimate for other larger values of a has not been theoretically studied. Some experimental evidence of its accuracy has been provided in [6] for a particular toy cipher. Whether this extends to other real-life ciphers is not known.

Our analysis, on the other hand, does not make any approximations for the statistical analysis. As such, the data complexity expression applies to all ciphers and for all values of a and P_S.

Generality of the expression for data complexity: In [6], separate expressions for the data complexity and success probability is obtained. The data complexity is stated to hold for success probability close to 0.5 and the expression for the success probability is quite complicated. In contrast, our analysis results in a single expression for the data complexity which is explicitly expressed in terms of the success probability and the advantage. No such expression for the data complexity is provided in [7]. For example, if one were to require the success probability to be 0.95, then the corresponding data complexity cannot be derived from the analysis in [7].

Analysis without independence assumption: Assumption 1 is used for the analysis in [7]. We also use Assumption 1 for one of our analysis. Whether Assumption 1 holds in general is not known. We show how to obtain an expression for the data complexity without using Assumption 1. There is no previous work in the literature which does this.

7.1 Experimental Comparison

We provide an experimental comparison of the data complexities obtained using our method and the methods available in previous works.

The experiments have been conducted using the set of differentials of the 64-bit toy block cipher SMALLPRESENT [8], that was given in [6, Table 6]. The table gives 3 estimates of the probabilities for the same input and output differences, namely, theoretical, 40-bit key schedule and 80-bit key schedule. In our experiments, we have made comparisons for each of these 3 probability estimates. The target sub-key size was 32 bits. So, $n = 64$ and $m = 32$.

Comparison when $\nu_0 > 1$: The only prior work which analysed this case is [6] and the data complexity expression obtained in [6] is given by (7). This expression holds for success probability close to 0.5. Denote by N_{BG} the data complexity given by (9).

For $\nu_0 > 1$, we have obtained two expressions for the data complexity. One using Assumption 1, based on the Chernoff bound and is given by the right hand side of (18). The other does not require Assumption 1, is based on the theory of martingales and is given by the right hand side of (20). Denote the data complexity given by (18) as N_{Cher} and the data complexity given by (20) as N_{Mar}. The expressions for N_{Cher} and N_{Mar} do not require P_S to be necessarily 0.5. However, since N_{BG} requires this, we set $P_S = 0.5$ for these expressions.

Table 1 gives a comparison between the data complexities N_{BG}, N_{Cher} and N_{Mar} for $a = 20$ and $P_S = 1 - \alpha = 0.5$. From the table one can see that N_{Cher} is only slightly greater than N_{BG}. So, if Assumption 1 can be assumed to hold, then it is better to use the data complexity given by N_{Cher} since it is more generally applicable. On the other hand, if Assumption 1 cannot be assumed to hold, then one has to use N_{Mar} and the corresponding data complexity is much higher.

Probability Estimates	N_{BG}	N_{Cher}	N_{Mar}
expressions	$\frac{2a\ln 2-2\ln 2\sqrt{\pi}}{\nu_0 D(\hat{p}\|\|\hat{q})}$	$\frac{3\ln 2\left(\sqrt{\hat{p}}+\sqrt{a\hat{q}}\right)^2}{\nu_0(\hat{p}-\hat{q})^2}$	$\frac{2\ln 2\left(1+\sqrt{a}\right)^2}{(\hat{p}-\hat{q})^2}$
theoretical	2.51×10^{7}	4.65×10^{7}	7.47×10^{17}
40-bit	1.77×10^{7}	3.32×10^{7}	3.81×10^{17}
80-bit	6.66×10^{6}	1.30×10^{7}	5.86×10^{16}

Table 1: Table showing the comparison between the data complexities N_{BG}, N_{Cher} and N_{Mar}. For compatibility with N_{BG}, P_S has been taken to be 0.5 in N_{Cher} and N_{Mar}.

Comparison when $\nu_0 = 1$: In this case, there is only a single input difference. The data complexity expression from [6] given in (9) can be specialised to the case $\nu_0 = 1$. As before we denote by N_{BG} the data complexity arising from (9) by setting $\nu_0 = 1$. Also, let N_{Cher} denote the data complexity given by (12).

Table 6 in [6] provides six groups of differentials with each group having the same input difference. We separately consider each of these six groups for the three probability estimates. This leads to 18 cases. For each of the 18 cases, we compare N_{BG} and N_{Cher}. Since N_{BG} requires $P_S = 0.5$ we have computed the data complexities for all the cases with this value of P_S. The advantage a was varied from 1 to 32. In each case, the comparative nature of the three data complexities are similar and so we report only the data complexities for $a = 20$. These are shown in Table 2. From the table, one can see that the data complexities are close and so, there is not a significant penalty for working with a data complexity expression which applies more generally.

Probability Estimates	Input Difference	N_{BG}	N_{Cher}
Theoretical	0x3	2.01×10^8	3.68×10^8
	0x7	1.09×10^8	2.04×10^8
	0xD	1.54×10^8	2.87×10^8
	0x5	1.86×10^8	3.42×10^8
	0xB	1.99×10^8	3.65×10^8
	0xF	1.13×10^8	2.11×10^8
40-bit	0x3	1.32×10^8	2.45×10^8
	0x7	8.22×10^7	1.56×10^8
	0xD	1.08×10^8	2.03×10^8
	0x5	1.29×10^8	2.40×10^8
	0xB	1.27×10^8	2.37×10^8
	0xF	8.40×10^7	1.59×10^8
80-bit	0x3	1.42×10^8	2.63×10^8
	0x7	8.90×10^6	1.84×10^7
	0xD	1.20×10^8	2.25×10^8
	0x5	1.28×10^8	2.39×10^8
	0xB	1.39×10^8	2.58×10^8
	0xF	8.56×10^7	1.62×10^8

Table 2: Table showing the comparison between the data complexities N_{BG} and N_{Cher} for SMALLPRESENT with $\nu_0 = 1$, $P_S = 0.5$ and $a = 20$.

The case $\nu_0 = 1$ was earlier studied in [7]. Two approaches were used. One was based on the log-likelihood ratio (LLR) and the other on the chi-squared statistic and corresponding expressions for data complexities were obtained. We tried to compare these data complexities with those in Table 2. However, this turned out to be problematic. For both the LLR and the chi-squared approaches from [7], the values obtained for the data complexities turned out to be meaningless. We further investigated the reasons for this and came to the following conclusions.

For the chi-squared based approach in [7], certain conditions are required to hold for the approximations to be valid. For the distributions of SMALLPRESENT these conditions are violated and so the applicability of the data complexity expression becomes invalid. The problem with the LLR-based approach from [7] is different. It turns out that putting $P_S = 0.5$ in the LLR-based data complexity expression given in [7] leads to meaningless values, while the expression provides meaningful values of data complexity for higher values of P_S. This indicates that the data complexity expression obtained from the LLR-based approach is invalid for $P_S = 0.5$. An earlier LLR-based data complexity expression for linear cryptanalysis [11, Theorem 2] had explicitly required $P_S > 0.5$. Though this is not mentioned in the analysis in [7], it seems that this condition still applies.

It is interesting to note that between the different approximate data complexity expressions, the one in [6] requires P_S to be close to 0.5; the LLR-based data complexity expression in [7] requires $P_S > 0.5$ and the chi-square based data complexity expression does not apply to SMALLPRESENT. This indicates various troublesome issues in using approximations. As mentioned earlier, the analysis of data complexity carried out in this paper, does not require any approximations and applies for all values of P_S.

8 Conclusion

This work considered multiple differential cryptanalysis without any restrictions on the input and the output differences. Expressions for data complexities were derived. These are obtained as closed-form formulas in terms of the success probability and the advantage of an attack. A main point of the work was to avoid making any approximation. As a result the obtained expressions are generally applicable.

References

[1] Thomas Baignères, Pascal Junod, and Serge Vaudenay. How Far Can We Go Beyond Linear Cryptanalysis? In *Advances in Cryptology–ASIACRYPT 2004*, pages 432–450. Springer, 2004.

[2] Eli Biham, Alex Biryukov, and Adi Shamir. Cryptanalysis of Skipjack Reduced to 31 Rounds Using Impossible Differentials. In *Advances in Cryptology–Eurocrypt99*, pages 12–23. Springer, 1999.

[3] Eli Biham and Adi Shamir. Differential Cryptanalysis of DES-like Cryptosystems. In *Advances in Cryptology–CRYPTO'90*, pages 2–21. Springer, 1990.

[4] Eli Biham and Adi Shamir. Differential Cryptanalysis of DES-like Cryptosystems. *Journal of CRYPTOLOGY*, 4(1):3–72, 1991.

[5] Alex Biryukov, Christophe De Cannière, and Michaël Quisquater. On Multiple Linear Approximations. In *Advances in Cryptology–CRYPTO 2004*, pages 1–22. Springer, 2004.

[6] Céline Blondeau and Benoît Gérard. Multiple Differential Cryptanalysis: Theory and Practice. In *Fast Software Encryption*, pages 35–54. Springer, 2011.

[7] Céline Blondeau, Benoît Gérard, and Kaisa Nyberg. Multiple Differential Cryptanalysis using LLR and χ^2 Statistics. In *Security and Cryptography for Networks*, pages 343–360. Springer, 2012.

[8] Céline Blondeau, Benoît Gérard, and Jean-Pierre Tillich. Accurate Estimates of the Data Complexity and Success Probability for Various Cryptanalyses. *Designs, Codes and Cryptography*, 59(1-3):3–34, 2011.

[9] Itai Dinur and Adi Shamir. Cube Attacks on Tweakable Black Box Polynomials. *Advances in Cryptology–EUROCRYPT 2009*, pages 278–299, 2009.

[10] Geoffrey Grimmett and David Stirzaker. *Probability and Random Processes*. Oxford university press, 2001.

[11] Miia Hermelin, Joo Yeon Cho, and Kaisa Nyberg. Multidimensional Extension of Matsuis Algorithm 2. In *Fast Software Encryption*, pages 209–227. Springer, 2009.

[12] Pascal Junod. On the Complexity of Matsuis Attack. In *Selected Areas in Cryptography*, pages 199–211. Springer, 2001.

[13] Pascal Junod. On the Optimality of Linear, Differential, and Sequential Distinguishers. In *Advances in Cryptology–EUROCRYPT 2003*, pages 17–32. Springer, 2003.

[14] Pascal Junod and Serge Vaudenay. Optimal Key Ranking Procedures in a Statistical Cryptanalysis. In *Fast Software Encryption*, pages 235–246. Springer, 2003.

[15] Burton S Kaliski Jr and Matthew JB Robshaw. Linear Cryptanalysis Using Multiple Approximations. In *Advances in Cryptology–Crypto94*, pages 26–39. Springer, 1994.

[16] Lars R Knudsen. Truncated and Higher Order Differentials. In *Fast Software Encryption*, pages 196–211. Springer, 1995.

[17] Xuejia Lai. Higher order derivatives and differential cryptanalysis. In *Communications and Cryptography*, pages 227–233. Springer, 1994.

[18] Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005.

[19] Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms*. Chapman & Hall/CRC, 2010.

[20] Sean Murphy. The Analysis of Simultaneous Differences in Differential Cryptanalysis. Technical Report RHUL-MA-2012-13, Royal Holloway, University of London, 2011. Availabel at `http://www.isg.rhul.ac.uk/~sean/SimDiffA.pdf`.

[21] Subhabrata Samajder and Palash Sarkar. Another look at normal approximations in cryptanalysis. Cryptology ePrint Archive, Report 2015/679, 2015. `http://eprint.iacr.org/`.

[22] Subhabrata Samajder and Palash Sarkar. Rigorous Upper Bounds on Data Complexities of Block Cipher Cryptanalysis. *IACR Cryptology ePrint Archive*, 2015:916, 2015. `http://eprint.iacr.org/2015/916`.

[23] Ali Aydın Selçuk. On Probability of Success in Linear and Differential Cryptanalysis. *Journal of Cryptology*, 21(1):131–147, 2008.

[24] Cihangir Tezcan. The Improbable Differential Attack: Cryptanalysis of Reduced Round CLEFIA. In *Progress in Cryptology-INDOCRYPT 2010*, pages 197–209. Springer, 2010.

[25] David Wagner. The Boomerang Attack. In *Fast Software Encryption*, pages 156–170. Springer, 1999.

A Chernoff Bounds

We briefly recall some results on tail probabilities of sums of Poisson trials that will be used. These results can be found in standard texts such as [19, 18] and are usually referred to as the Chernoff bounds.

Theorem 4. *Let $X_1, X_2, \ldots, X_\lambda$ be a sequence of independent Poisson trials such that for $1 \leq i \leq \lambda$, $\Pr[X_i = 1] = p_i$. Then for $X = \sum_{i=1}^{\lambda} X_i$ and $\mu = E[X] = \sum_{i=1}^{\lambda} p_i$ the following bounds hold:*

$$\text{For any } \gamma > 0,\ \Pr[X \geq (1+\gamma)\mu] < \left(\frac{e^{-\gamma}}{(1+\gamma)^{(1+\gamma)}}\right)^{\mu}. \tag{23}$$

$$\text{For any } 0 < \gamma < 1,\ \Pr[X \leq (1-\gamma)\mu] \leq \left(\frac{e^{-\gamma}}{(1-\gamma)^{(1-\gamma)}}\right)^{\mu}. \tag{24}$$

These bounds can be simplified to the following form.

$$\text{For any } 0 < \gamma \leq 1,\ \Pr[X \geq (1+\gamma)\mu] \leq e^{-\mu\gamma^2/3}. \tag{25}$$

$$\text{For any } 0 < \gamma < 1,\ \Pr[X \leq (1-\gamma)\mu] \leq e^{-\mu\gamma^2/2}. \tag{26}$$

B Martingales

This section gives a brief description of martingales for discrete random variables. Further details can be found in standard texts such as [10, 18]. We start with the definition of conditional expectation.

Definition 1 (Conditional Expectation)**.** *Let X and Y be two random variables such that $E\left[X\right] < \infty$. Define*

$$\psi\left(y\right) \triangleq E\left[X|\, Y = y\right] = \sum_x x \Pr\left[X = x|\, Y = y\right].$$

Thus, $E\left[X|\, Y = y\right]$ is a function of y. The conditional expectation of X given Y is defined to be $\psi\left(Y\right)$ and is written as $\psi\left(Y\right) \triangleq E\left[X|\, Y\right]$. So, the conditional expectation of X given Y is a random variable $\psi\left(Y\right)$ which is a function of the random variable Y.

The following are several standard properties of conditional expectation.

Proposition 5. *1. $E\left[E\left[Y \mid X\right]\right] = E\left[X\right].$*

2. If X has a finite expectation and if g is a function such that $Xg(Y)$ has a finite expectation, then $E\left[Xg(Y) \mid Y\right] = E\left[X \mid Y\right] g(Y)$.

3. $E\left[(X - g(Y))^2\right] \geq E\left[(X - E\left[X \mid Y\right])^2\right]$ for any pair of random variables X and Y such that X^2 and $g(Y)^2$ have finite expectations.

4. For any function g, such that $g(X)$ has finite expectation, $E\left[g(X) \mid Y = y\right] = \sum_x g(x) \Pr\left[X = x \mid Y = y\right].$

5. $\mid E\left[X \mid Y\right] \mid \leq E\left[\mid X \mid\mid Y\right].$

6. $E\left[E\left[X \mid Y, Z\right] \mid Y\right] = E\left[X \mid Y\right].$

7. $E\left[E\left[g(X,Y) \mid Z, W\right] \mid Z\right] = E\left[g(X,Y) \mid Z\right].$

Definition 2 (Martingale)**.** *A sequence of random variables $Z_1, Z_2, Z_3, \ldots$ is a martingale with respect to another sequence of random variables $Y_1, Y_2, Y_3, \ldots$ if for all $n \geq 1$ the following two conditions hold.*

1. $E\left[|Z_n|\right] < \infty$.

2. $E\left[Z_{n+1}|\, Y_1, Y_2, \ldots, Y_n\right] = Z_n$.

If $Z_n = Y_n$ for all $n \geq 1$ then the sequence is a martingale with respect to itself.

The basic Azuma-Hoeffding inequality for martingales is the following.

Theorem 6. *Let, $Z_0, Z_1, Z_2, \ldots$ be a martingale with respect $Y_0, Y_1, Y_2, \ldots$ and suppose that there exists a sequence $\upsilon_1, \upsilon_2, \ldots$ of real numbers such that for all $i \geq 1$, $\mid Z_i - Z_{i-1} \mid \leq \upsilon_i$. Then for any integer $\lambda > 0$ and real $\delta > 0$*

$$\Pr\left[Z_\lambda - Z_0 \geq \delta\right] \leq e^{-\delta^2/\left(2\sum_{i=1}^{\lambda}\upsilon_i^2\right)}; \tag{27}$$

$$\Pr\left[Z_\lambda - Z_0 \leq -\delta\right] \leq e^{-\delta^2/\left(2\sum_{i=1}^{\lambda}\upsilon_i^2\right)}. \tag{28}$$

Given a random variable Y with $E\left[\mid Y \mid\right] < \infty$ and a sequence of random variables $Y_0, Y_1, \ldots Y_\lambda$, a simple way to construct a martingale is the following. Define $Z_i = E\left[Y \mid Y_0, Y_1, \ldots, Y_i\right]$ for $i = 0, 1, \ldots, n$. It can be shown through a routine calculation using properties of conditional expectation given in Proposition 5 that the following condition holds.

$$E\left[Z_{i+1} \mid Y_0, Y_1, \ldots, Y_i\right] \quad = \quad Z_i.$$

Thus, the sequence of random variables $\{Z_\lambda\}$ forms a martingale with respect to sequence $\{Y_\lambda\}$. A martingale of this type is called a **Doob Martingale**.

Recall that the Azuma-Hoeffding inequality can be applied if the differences $|Z_i - Z_{i-1}|$ are bounded. A general technique for obtaining a Doob martingale with bounded differences is as follows. We call a function $f(y_1, y_2, \ldots, y_\lambda)$ to be υ-**Lipschitz condition**, if for any i and for any set of values $y_1, y_2, \ldots, y_\lambda$ and y_i',

$$\mid f(y_1, y_2, \ldots, y_{i-1}, y_i, y_{i+1}, \ldots, y_\lambda) - f(y_1, y_2, \ldots, y_{i-1}, y_i', y_{i+1}, \ldots, y_\lambda) \mid \leq \upsilon.$$

In other words, changing the value of any single coordinate changes the value of the function by at most υ. Let $Y_1, \ldots, Y_\lambda$ be a finite sequence of random variables. Define,

$$\begin{aligned} Z_0 &= E\left[f(Y_1, Y_2, \ldots, Y_\lambda)\right] \\ Z_i &= E\left[f(Y_1, Y_2, \ldots, Y_\lambda) \mid Y_1, Y_2, \ldots, Y_i\right]. \end{aligned}$$

Then $Z_0, Z_1, \ldots, Z_\lambda$ form a Doob martingale with respect to $Y_1, \ldots, Y_\lambda$. Further, assume that the random variables Y_i's are *independent*. Then it can be shown that $|Z_i - Z_{i-1}| \leq \upsilon$. The martingale $Z_0, \ldots, Z_\lambda$ satisfies the conditions of Theorem 6 and so the inequality stated in the theorem applies to this martingale.