Can Large Deviation Theory be Used for Estimating Data Complexity?

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

Abstract

Statistical analysis of attacks on block ciphers have mostly used normal approximations. A few recent works have proposed doing away with normal approximations and instead use Chernoff and Hoeffding bounds to obtain rigorous bounds on data complexities of several attacks. This opens up the question of whether even better general bounds can be obtained using the statistical theory of large deviations. In this note we examine this question. Our conclusion is that while in theory this is indeed possible, in general obtaining meaningful expressions for data complexity presents several difficulties. This leaves open the question of whether this can be done for specific attacks.

1 Introduction

A key recovery attack on a block cipher aims to recover the correct value of a sub-key, i.e., a portion of the secret key. It has two phases. In the first phase, a list of possible candidate values of the sub-key is returned and the second phase performs a (brute force) checking of all the candidate keys.

Statistical analysis of attacks on block ciphers allows estimating the number of plaintext-ciphertext pairs required for an attack. The number of such plaintext-ciphertext pairs is called the data complexity of an attack. The data complexity depends on two parameters, namely the success probability and the advantage (which measures the size of the candidate list returned by the first phase). The goal of a statistical analysis is to be able to obtain an explicit expression for the data complexity in terms of the success probability and the advantage.

While some form of statistical analysis has always accompanied attacks on block ciphers, a systematic approach to such analysis was given in [9]. This approach is based on the earlier idea [4] of ranking of keys by the value of a test statistic associated with each key. A result on the normal approximation of order statistics was used to formalise the idea of ranking. It was used to analyse linear and differential cryptanalysis and was later used in several works [3, 1] to analyse more advanced attacks. A recent paper [5] has studied in details the order statistics based approach including the error in normal approximation. This shows several shortcomings of the approach and calls into question its widespread use in analysing attacks on block ciphers.

An alternative statistical approach is to use the standard theory of hypothesis testing. In the past, this approach has been used for distinguishing attacks, but, in the context of key recovery attack, hypothesis testing has not received much attention except for a passing mention in [2]. It is only recently [5] that hypothesis testing has been systematically used to re-derive expressions for data complexities obtained earlier in [9, 3, 1].

To perform a statistical analysis, it is required to identify a test statistic irrespective of whether the order statistics or the hypothesis testing based approach is used. This test statistic depends on a value of the sub-key and has two different distributions according as whether the sub-key is correct or incorrect[1] Most of the works in the literature have used the normal approximation to estimate the two distributions. The errors in such approximations have not been analysed. The work [5] takes a careful look at the errors in the normal approximations used in several works in the literature and concludes that such approximations restrict the applicability of the analyses.

Using the hypothesis testing based approach requires bounding the Type-I and Type-II error probabilites. This turns out to be essentially the task of bounding the tail probabilities of certain distributions. The Type-I error probability is the tail probability of the test statistic under the condition that the null hypothesis holds (i.e., the choice of the sub-key is correct) while the Type-II error probability is the tail probability of the test statistic under the condition that the alternate hypothesis holds (i.e., the choice of the sub-key is incorrect).

In almost all cases, the test statistic turns out to be the sum of some independent and identically distributed random variables taking values from a finite set. So, the requirement is to bound the tail probabilities of a sum of independent random variables (under both the null and the alternate hypotheses). There are known general bounds for doing this. When the individual random variables are Bernoulli distributed, the Chernoff bound can be used. If the set from which the random variables take values has more than two elements, then applying the Chernoff bound becomes difficult. In this case, the more general Hoeffding bound can be applied[2].

Both the Chernoff and Hoeffding bounds do not require any approximations and can be considered to provide rigorous bounds on the data complexity. Recently, several works [6, 8, 7] have followed this approach to obtain expressions for bounds on data complexities of several attacks on block ciphers. This opens up the question of whether these bounds can be improved in general.

Theory of large deviations. The branch of probability/statistics dealing with probability of rare events is known as the theory of large deviations. The tail probability can also be tackled using standard tools from this theory. So, the question arises as to whether it is possible to obtain better bounds for data complexities using tools from the theory of large deviations.

In this note, we take a look at some of the basic results from the theory of large deviations to determine whether these can be used for analysing data complexity. In theory, this can certainly be done. Our conclusion, however, is that in general there are several difficulties in obtaining bounds which can be actually be computed in practice to estimate the data complexity of an attack. We discuss these difficulties in some details.

[1] More recently, there has been work on considering different distributions for different incorrect keys. Here we will not consider this issue.

[2] A roundabout way of doing this is to go through the theory of martingales and applying the Azuma-Hoeffding bound. The obtained bound is (almost) the same

The question of whether these difficulties can be overcome for specific attacks is left open.

2 Basic Results from Theory of Large Deviations

Let $X_1, \ldots, X_N$ be independent identically distributed random variables with mean μ and $S_N = X_1 + \cdots + X_N$. The following short computation establishes a general bound called the Chernoff bound. For any $x > \mu$ and $\theta > 0$,

$$\begin{aligned}
\Pr[S_N > xN] &\leq \Pr\left[e^{\theta S_N} > e^{\theta N x}\right] \\
&= \frac{E[\exp(\theta S_N)]}{\exp(\theta x N)} && \text{(Markov Inequality)} \\
&= \frac{E\left[\exp\left(\theta \sum_{i=1}^{N} X_i\right)\right]}{\exp(\theta x N)} \\
&= \frac{E\left[\prod_{i=1}^{N} \exp(\theta X_i)\right]}{\exp(\theta x N)} \\
&= \frac{\prod_{i=1}^{N} E[\exp(\theta X_i)]}{\exp(\theta x N)} && \text{(independence)} \\
&= \left(\frac{E[\exp(\theta X_1)]}{\exp(\theta x)}\right)^N && \text{(identically distributed)} \\
&= \exp\left(-N\left(x\theta - \ln M_{X_1}(\theta)\right)\right).
\end{aligned}$$

Here $M_{X_1}(\theta) = E[\exp(\theta X_1)]$ is the moment generating function of the random variable X_1.

To obtain a good bound, the goal is to minimise the right hand side over all possible θ. To this end, a function $I(x)$ called the rate function is defined as follows:

$$I(x) = \sup_{\theta} \left(x\theta - \ln M_{X_1}(\theta)\right). \tag{1}$$

The function $I(x)$ is the Legendre transform of $M_{X_1}(\theta)$ (or the random variable X_1). Using $I(x)$, we obtain

$$\Pr[S_N > xN] \leq \exp(-NI(x)). \tag{2}$$

Similarly, one can show that if $x < \mu$, then

$$\Pr[S_N < xN] \leq \exp(-NI(x)). \tag{3}$$

Consider now the setting of hypothesis testing. There are two hypothesis H_0 and H_1 and assume that under H_i, the means of the X_i's are μ_i, $i = 0, 1$. The distribution of the random variables X_i are different under H_0 and H_1 and so the rate function will also change. Denote by $I_i(x)$, $i = 0, 1$, the rate function corresponding to the hypothesis H_i.

For the sake of convenience, assume that $\mu_1 < \mu_0$, the other case being similar. Let the test statistics be S_n and suppose that the test take the following form:

"Reject H_0 if $S_N \leq Nt$ for some $t \in (\mu_1, \mu_0)$."

In the context of applying hypothesis testing to block cipher cryptanalysis, the test takes the above form.

The bounds on the Type-I and Type-II error probabilities are as follows:

$$\begin{aligned}
\Pr[\text{Type-I error}] &= \Pr[S_N \leq tN |\ H_0 \text{ holds}] \leq e^{-NI_0(t)}; \\
\Pr[\text{Type-II error}] &= \Pr[S_N > tN |\ H_1 \text{ holds}] \leq e^{-NI_1(t)}.
\end{aligned}$$

Denote the upper bound on Type-I error probability by α and the upper bound on the Type-II error probability by β. Then we get

$$\begin{aligned} NI_0(t) &= \ln(1/\alpha); \\ NI_1(t) &= \ln(1/\beta). \end{aligned} \tag{4}$$

To obtain an expression for N, it is required to eliminate t from these two equations.

3 Difficulties

We identify three difficulties in applying the above scenario to the context of block cipher cryptanalysis.

3.1 Difficulty 1: Limited Information on the Distribution of X_1

Note that computing $M_{X_1}(\theta)$ requires knowing the distribution of the random variable X. In the context of cryptanalysis, this is mostly not known. Analysis of the block cipher only provides an estimate of the expectation of X. So, computing $M_{X_1}(\theta)$ is in general not possible.

3.2 Difficulty 2: Computing the Rate Function

Suppose that it were possible to somehow obtain $M_{X_1}(\theta)$. The rate function requires taking a supremum over all possible values of θ. In general this is difficult to do. One way would be to use the standard approach of differentiating, setting to 0 and then solving for θ.

In the context of cryptanalysis, one encounters random variables which take values from a finite set. For a random variable X taking ρ values $v_1, \ldots, v_\rho$ with corresponding probabilities $p_1, \ldots, p_\rho$,

$$M_{X_1}(\theta) = E\left[e^{X\theta}\right] = \sum_{i=1}^{\rho} p_i e^{\theta v_i}.$$

For a general value of ρ, differentiating $x\theta - M_{X_1}(\theta)$ with respect to θ and solving for θ does not seem to be possible. When $\rho = 2$ and each X_i follows the Ber(p) distribution it can be shown that

$$I(x) = x\ln(x/p) + (1-x)\ln((1-x)/(1-p)). \tag{5}$$

3.3 Difficulty 3: Inverting the Rate Function

Suppose it is possible to obtain an expression for the rate function. Even then it would be required to invert it so as to be able to eliminate t from (4). Even in the simplest case of Bernoulli trials, from the form of $I(x)$ given by (5), there does not appear to be any simple way of eliminating t from (4).

4 Other Concentration Bounds

As mentioned earlier, the Chernoff and the Hoeffding bounds have been successfully used to obtain rigorous bounds on data complexities. Apart from these two bounds, there are several other concentration inequalities. A summary can be found at

https://en.wikipedia.org/wiki/Concentration_inequality#Bounds_on_sums_of_independent_variables.

Of specific interest are the Bennett's and Bernstein's inequalities. Applying these bounds require knowledge of the variances of the individual random variables. This is often difficult to obtain. Further, applying Bennett's inequality presents a difficulty similar to that of inverting the rate function discussed above. At this point, we are unaware of any concentration inequality which can be applied in general to the context of statistical analysis of block ciphers and which allows obtaining meaningful and improved expressions for data complexities than those given by the Chernoff and the Hoeffding bounds.

5 Conclusion

We have considered the possibility of applying the theory of large deviations for estimating data complexity of attacks on block ciphers. While in theory this can be done, obtaining meaningful expressions is in general difficult in practice. These difficulties are summarised below.

1. Sufficient information about a random variable may not be available so as to be able to compute the moment generating function.
2. Even if the moment generating function is known, it may not be possible to compute the expression for the rate function.
3. Even if the expression for the rate function is known, using it to obtain an expression for the data complexity may not be possible.

We conclude by noting that while there seems to be general difficulty in obtaining expressions for data complexity from the theory of large deviations, there remains the possibility of obtaining estimates of data complexity in specific cases.

References

[1] 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.

[2] 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.

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

[4] Mitsuru Matsui. The First Experimental Cryptanalysis of the Data Encryption Standard. In Y. G. Desmedt, editor, *Advances in Cryptology–Crypto94*, pages 1–11. Springer, 1994.

[5] Subhabrata Samajder and Palash Sarkar. Another Look at Normal Approximations in Cryptanalysis. *Journal of Mathematical Cryptology*. to appear.

[6] Subhabrata Samajder and Palash Sarkar. Rigorous upper bounds on data complexities of block cipher cryptanalysis. Cryptology ePrint Archive, Report 2015/916, 2015. http://eprint.iacr.org/.

[7] Subhabrata Samajder and Palash Sarkar. Multiple differential cryptanalysis: A rigorous analysis. Cryptology ePrint Archive, Report 2016/405, 2016. `http://eprint.iacr.org/`.

[8] Subhabrata Samajder and Palash Sarkar. A new test statistic for key recovery attacks using multiple linear approximations. Cryptology ePrint Archive, Report 2016/404, 2016. `http://eprint.iacr.org/`.

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