A Theoretical Approach to Bayesian Cryptanalysis

Bhaskar Biswas* Sourabh Bhattacharya#
Kwangjo Kim*
* School of Computing, KAIST, S Korea
# Indian Statistical Institute, Kolkata, India

Abstract

Till date, the cryptanalysis of iterated ciphers are mainly based on dedicated metrics like linear *bias* or *differential* or their variants (mostly). The attacks are dependent on design of cipher[Hey02]. To the best of our knowledge, no successful attempt been made to generalize the choice or selection of metrics; known or unknown. In this paper we try to use the state of the art statistical tools to derive a method of generic attack against iterated ciphers. We present only the theoretical setup here. The simulation results are work under progress.

1 Introduction

Statistics plays a key role in cryptanalysis[1]. Although statistical analysis alone will rarely give solutions to cryptographic systems, it often plays the central role in a method of attack. The probabilistic variation of plaintext, or possibly of keys, forms the basis of many cryptanalytic techniques [Kul76]. The Ph.D. thesis of P. Junod[Jun05] gives an overview. Statistical inference is the process of making statements about the properties of a population based on a sample of possible observations and any other available information. Bayesian inference utilises Bayes' theorem, extended to include currently available information, to modify opinion by experience. [Lin65, Pre89] give good introductions to Bayesian inference. A region S_α is a $100(1-\alpha)$ credible region if

$$\int_{S_\alpha} \pi(\theta|x)d\theta = 1-\alpha$$

The problem, similar to that encountered in classical inference, is that there may be any number of regions containing a $(1-\alpha)$ proportion of the posterior distribution. However it seems reasonable to choose S_α so that S_α does not exclude any value of θ more probable than any value of θ already in S_α.

In this paper we use the Bayesian Inference techniques to analyze the block ciphers. We first derive a *basis* function which shall allow us to compute component-wise *likelihood* for any iterated cipher. Then Gibbs sampling method is used for posterior sampling. We follow the Markov Chain - Monte Carlo method to achieve conditional distributions and eventually the *credible regions*.

2 The methodology

Let $\boldsymbol{p} = (p_1, \ldots, p_n)$ be an input (plain-text), $\boldsymbol{\kappa}$ be the secret key, which gives birth to the r (number of rounds in the cipher) round keys $= (\kappa^1, \ldots, \kappa^r)$ each of n bits by the key-generation algorithm. The encrypted message $\boldsymbol{c} = (c_1, \ldots, c_n)$. We denote the bit positions by subscipts and different sets by superscripts.

Let our data set be $\boldsymbol{D} = (\boldsymbol{c}^1 = \boldsymbol{f}(\boldsymbol{p}^1, \boldsymbol{\kappa}), \ldots, \boldsymbol{c}^m = \boldsymbol{f}(\boldsymbol{p}^m, \boldsymbol{\kappa}))'$ be the set of outputs corresponding to inputs $\{(\boldsymbol{p}^1, \boldsymbol{\kappa}), \ldots, (\boldsymbol{p}^m, \boldsymbol{\kappa})\}$.

Now let, $\boldsymbol{c}^{m+1} = \boldsymbol{f}(\boldsymbol{p}^{m+1}, \boldsymbol{\kappa})$, where $\boldsymbol{f}(\cdot,\cdot)$ is the cryptographic function, $\boldsymbol{c}^{m+1}$ is known but $\boldsymbol{p}^{m+1}$ and

[1]**Acknowledgement :** This work was partially supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (No. NRF-2015R1A2A2A01006812).

$\boldsymbol{\kappa}$ are unknown, our objective is to obtain the posterior distribution of $(\boldsymbol{p}^{m+1}, \boldsymbol{\kappa})$, given $\boldsymbol{D}$ and $\boldsymbol{c}^{m+1}$; in particular, we are interested in the marginal posterior of $\boldsymbol{\kappa}$, given $\boldsymbol{D}$ and $\boldsymbol{c}^{m+1}$. Thus, we are in an inverse problem set-up: given the inputs $(\boldsymbol{p}, \boldsymbol{\kappa})$, the response $\boldsymbol{c} = \boldsymbol{f}(\boldsymbol{p}, \boldsymbol{\kappa})$ is known, but we are interested in obtaining the inverse of $\boldsymbol{c}$, given by $\boldsymbol{f}^{-1}(\boldsymbol{c})$. Note that, given $\boldsymbol{\kappa}$, $\boldsymbol{p} = \boldsymbol{f}^{-1}(\boldsymbol{c}, \boldsymbol{\kappa})$ is known completely.

2.1 The Reed-Muller (basis) representation

It is well-known that any function $g : \mathbb{F}^N \longmapsto \mathbb{F}$ can be represented using the following Reed-Muller expansion:

$$g(\boldsymbol{w}) = \oplus_{\boldsymbol{\alpha} \in \mathbb{F}^N} C_\alpha \mu_\alpha(\boldsymbol{w}), \tag{1}$$

where N is integer, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)'$, $\mu_{\boldsymbol{\alpha}}(\boldsymbol{w}) = \prod_{j=1}^N w_j^{\alpha_j}$ are the basis functions, with $\alpha_j \in \{0, 1\}$ for all $j = 1, 2, \ldots, N$; the coefficients $C_{\boldsymbol{\alpha}} \in \{0, 1\}$ for $\boldsymbol{\alpha} \in \mathbb{F}^N$. In the above, we define $w_j^0 = 1$ and $w_j^1 = w_j$.

We exploit the above Reed-Muller expansion principle to represent our unknown, multidimensional, Boolean function $\boldsymbol{f} : \mathbb{F}_2^n \longmapsto \mathbb{F}_2$.

Componentwise Reed-Muller representation

Note that a valid Reed-Muller based basis function representation of the i-th component $f_i(\boldsymbol{x}, \boldsymbol{\kappa})$ is as follows:

$$f_i(\boldsymbol{x}, \boldsymbol{\kappa}) = \oplus_{\boldsymbol{\alpha} \in \mathbb{F}^{2n}} C_{i,\boldsymbol{\alpha}} \mu_{\boldsymbol{\alpha}}(\boldsymbol{x}, \boldsymbol{\kappa}), \tag{2}$$

where the coefficients $C_{i,\boldsymbol{\alpha}}$ and $\mu_{\boldsymbol{\alpha}}(x)$ are of the same forms as in Eq.(1).

Prior on the Reed-Muller coefficients

We assume that, for each i and $\boldsymbol{\alpha}$,

$$C_{i,\boldsymbol{\alpha}} \overset{iid}{\sim} Bernoulli(p_{\boldsymbol{c}}). \tag{3}$$

Defining

$$r_{\boldsymbol{C}} = \sum_{i=1}^{n} \sum_{\boldsymbol{\alpha} \in \mathbb{F}^{2n}} C_{i,\alpha}, \tag{4}$$

the prior on $\boldsymbol{C} = \{C_{i,\boldsymbol{\alpha}} : i = 1, \ldots, n, \boldsymbol{\alpha} \in \mathbb{F}^{2n}\}$ is given by

$$[\boldsymbol{C}] \;=\; p_{\boldsymbol{c}}^{r_{\boldsymbol{C}}} (1 - p_{\boldsymbol{c}})^{nZ - r_{\boldsymbol{C}}}, \tag{5}$$

where $Z = 2^{2n}$.

The likelihood

We write $\boldsymbol{D}_{m+1} = (\boldsymbol{D}', \boldsymbol{y}_{m+1})'$, the data set $\boldsymbol{D}$, augmented with $\boldsymbol{y}_{m+1}$. The likelihood is given by

$$\begin{aligned} L(\boldsymbol{D}_{m+1}; \boldsymbol{c}, \boldsymbol{x}, \boldsymbol{\kappa}) \;&=\; \prod_{j=1}^{m} \delta_{\{\boldsymbol{y}_j\}} \left(f(\boldsymbol{x}_j, \boldsymbol{\kappa}_j)\right) \\ \times \delta_{\{\boldsymbol{y}_{m+1}\}} \left(f(\boldsymbol{x}, \kappa)\right) \qquad (6) \\ &=\; \prod_{j=1}^{m} \delta_{\{\boldsymbol{y}_j\}} \left(f_{\boldsymbol{c}}(\boldsymbol{x}_j, \boldsymbol{\kappa}_j)\right) \\ \times \delta_{\{\boldsymbol{y}_{m+1}\}} \left(f_{\boldsymbol{c}}(\boldsymbol{x}, \boldsymbol{\kappa})\right) \qquad (7) \end{aligned}$$

In practice we will approximate (7) by

$$\begin{aligned} L_\epsilon(\boldsymbol{D}_{m+1}; \boldsymbol{c}, \boldsymbol{x}, \boldsymbol{\kappa}) \;&=\; \prod_{j=1}^{m} \delta_{B(\boldsymbol{y}_j, \epsilon)} \left(\boldsymbol{f}_{\boldsymbol{c}}(\boldsymbol{x}_j, \boldsymbol{\kappa}_j)\right) \\ \times \delta_{B(\boldsymbol{y}_{m+1}, \epsilon)} \left(\boldsymbol{f}_{\boldsymbol{c}}(\boldsymbol{x}, \boldsymbol{\kappa})\right) \qquad (8) \end{aligned}$$

In Eq.(8), $B(\boldsymbol{y}, \epsilon) = \{\boldsymbol{z} : d(\boldsymbol{y}, \boldsymbol{z}) \leq \epsilon\}$ is an ϵ-neighbourhood of $\boldsymbol{y}$, where $\epsilon > 0$ and d is a suitably chosen metric that is capable of measuring distances between any two values in $\mathbb{F}^{2n}$. We consider the Hamming distance as a suitable metric between any two elements $\boldsymbol{u} = u_1 u_2 \cdots u_{2n}$ and $\boldsymbol{v} = v_1 v_2 \cdots v_{2n}$ where $u_i, v_i \in \{0, 1\}$ for each i:

$$\begin{aligned} d(\boldsymbol{u}, \boldsymbol{v}) \;&=\; \sum_{i=1}^{2n} (u_i \oplus v_i) \qquad (9) \\ &=\; \sum_{i=1}^{2n} |u_i - v_i| \qquad (10) \end{aligned}$$

The above distance, which can be easily seen to satisfy all the properties of a matric, can be interpreted as the number of position-wise mismatches.

Clearly, as $\epsilon \to 0$, $L_\epsilon(\boldsymbol{D}_{m+1}; \boldsymbol{c}, \boldsymbol{x}, \boldsymbol{\kappa}) \to L_\epsilon(\boldsymbol{D}_{m+1}; \boldsymbol{c}, \boldsymbol{x}, \boldsymbol{\kappa})$ pointwise. Given a suitable choice of ϵ, the number of basis functions used in the representation $\boldsymbol{f}_{\boldsymbol{c}}(\cdot, \cdot)$ of $\boldsymbol{f}(\cdot, \cdot)$ is the least number such that $f_{\boldsymbol{c}}(\boldsymbol{y}_j, \boldsymbol{\kappa}_j) \in B(\boldsymbol{y}_j, \epsilon)$ for each $j = 1, \ldots, m$.

Prior on $(\boldsymbol{x}, \boldsymbol{\kappa})$

We assume that *a priori*, for each j, $x_j \sim Bernoulli(p_{\boldsymbol{x}})$ independently, and $\kappa_j \sim Bernoulli(p_{\boldsymbol{\kappa}})$, independently.

The joint posterior

The joint posterior is given, up to a proportionality constant, by

$$[\boldsymbol{C}=\boldsymbol{c},\boldsymbol{X}=\boldsymbol{x},\boldsymbol{K}=\boldsymbol{\kappa}|\boldsymbol{D}_{m+1}] \propto [\boldsymbol{C}=\boldsymbol{c}]\times[\boldsymbol{X}=\boldsymbol{x}]\times[\boldsymbol{K}=\boldsymbol{\kappa}] \times L_\epsilon(\boldsymbol{D}_{m+1};\boldsymbol{c},\boldsymbol{x},\boldsymbol{\kappa}).$$

2.2 Computation of the posterior using Gibbs sampling

Given arbitrary initial values of $\boldsymbol{x},\boldsymbol{\kappa}$, drawn perhaps from their respective prior distributions[CG92]. The sample is drawn from the full conditional distributions $[\boldsymbol{C}|\boldsymbol{D}_{m+1},\boldsymbol{X},\boldsymbol{K}]$, $[\boldsymbol{K}|\boldsymbol{D}_{m+1},\boldsymbol{C},\boldsymbol{X}]$, and $[\boldsymbol{X}|\boldsymbol{D}_{m+1},\boldsymbol{C},\boldsymbol{K}]$. In fact, in what follows, we will find it easier to sample, for each $i=1,2,\ldots,n$, sequentially from $[C_{i,\boldsymbol{\alpha}}|\boldsymbol{D}_{m+1},\boldsymbol{c}_{-i,-\boldsymbol{\alpha}},\boldsymbol{x},\boldsymbol{\kappa}]$, $[\kappa_i|\boldsymbol{D}_{m+1},\boldsymbol{c},\boldsymbol{x},\boldsymbol{\kappa}_{-i}]$, $[x_i|\boldsymbol{D}_{m+1},\boldsymbol{c},\boldsymbol{x}_{-i},\boldsymbol{\kappa}]$, where $\boldsymbol{c}_{-i,-\boldsymbol{\alpha}}=\boldsymbol{c}\backslash\{c_{i,\boldsymbol{\alpha}}$, $\boldsymbol{\kappa}_{-i}=\boldsymbol{\kappa}\backslash\kappa_i$, and $\boldsymbol{x}_{-i}=\boldsymbol{x}\backslash x_i$.

Given an initial value $(\boldsymbol{C}^{(0)},\boldsymbol{K}^{(0)},\boldsymbol{X}^{(0)})$, for each $i=1,2,\ldots,n$, at each iteration $t=1,2,\ldots$, we generate samples sequentially

$C_{i,\boldsymbol{\alpha}}^{(t)}\sim[C_{i,\boldsymbol{\alpha}}|\boldsymbol{D}_{m+1},\boldsymbol{c}_{-i,-\boldsymbol{\alpha}}^{(t-1)},\boldsymbol{x}^{(t-1)},\boldsymbol{\kappa}^{(t-1)}]$;
$\kappa_i^{(t)}\sim[\kappa_i|\boldsymbol{D}_{m+1},\boldsymbol{c}_{i,\boldsymbol{\alpha}}^{(t)},\boldsymbol{c}_{-i,-\boldsymbol{\alpha}}^{(t-1)},\boldsymbol{x}^{(t-1)},\boldsymbol{\kappa}_{-i}^{(t-1)}]$; and
$x^{(t)}\sim[x_i|\boldsymbol{D}_{m+1},\boldsymbol{c}_{i,\boldsymbol{\alpha}}^{(t)},\boldsymbol{c}_{-i,-\boldsymbol{\alpha}}^{(t-1)},\boldsymbol{x}_{-i}^{(t-1)},\kappa_i^{(t)},\boldsymbol{\kappa}_{-i}^{(t-1)}]$.

We discard the samples corresponding to the first B iterations as burn-in (a thumb rule says that $B=20,000$ is often adequate), and then keep the samples corresponding to the next T ($T=50,000$ is often adequate) iterations for the inference.

Adequate choice of the initial values $(\boldsymbol{c}^{(0)},\boldsymbol{\kappa}^{(0)},\boldsymbol{x}^{(0)})$ is important for fast convergence of the Gibbs sampler. Below we discuss the choice of the initial values, while also shedding light on the choice of ϵ and the number of basis components in the Reed-Muller representation of the system.

2.3 Deterministic simulation of $\boldsymbol{p}$ given $\boldsymbol{c}$ and $\boldsymbol{\kappa}$

Given $\boldsymbol{c}$, the function $\boldsymbol{f}_{\boldsymbol{c}}(\cdot,\cdot)$ is known. Since knowledge of the key $\boldsymbol{\kappa}$ facilitates direct inversion of the function $\boldsymbol{f}_{\boldsymbol{c}}$ to obtain $\boldsymbol{p}$, it seems that given our current value $\boldsymbol{\kappa}$ we can directly use the deterministic function (provided it is known) to obtain $\boldsymbol{p}$. This deterministic simulation scheme is likely to improve the accuracy of the simulations of the other unknowns $\boldsymbol{c}$ and $\boldsymbol{\kappa}$.

3 Summarization of the posterior distributions

For objects in $\mathbb{F}^n$, such as $\boldsymbol{X}$ and $\boldsymbol{K}$, the summaries are not straightforward to derive as in the cases of real random variables. For this we introduce below the concept of "central estimate"; our approach will be akin to the "central clustering" approach of [SMD11] in the context of Bayesian clustering.

3.1 Definition of Central Estimate

Motivated by the definition of mode in the case of parametric distributions, we define that $\boldsymbol{z}^*\in\mathbb{F}^n$ as "central", which, for a given small $\epsilon>0$, satisfies the following equation:

$$P\left(\{\boldsymbol{z}\in\mathbb{F}^n : d(\boldsymbol{z}^*,\boldsymbol{z})<\epsilon\}\right) = \sup_{\boldsymbol{z}'} P\left(\{\boldsymbol{z}\in\mathbb{F}^n : d(\boldsymbol{z}',\boldsymbol{z})<\epsilon\}\right). \tag{11}$$

Note that $\boldsymbol{z}^*$ is the global mode of the distribution as $\epsilon\to 0$. Thus, for a sufficiently small $\epsilon>0$, the probability of an ϵ-neighborhood of an arbitrary value $\boldsymbol{z}'$, of the form $\{\boldsymbol{z}\in\mathbb{F}^n : d(\boldsymbol{z}',\boldsymbol{z})<\epsilon\}$, is the highest when $\boldsymbol{z}'=\boldsymbol{z}^*$, the central estimate.

The above definition will hold for all positive ϵ if the distribution of $\boldsymbol{z}$ is unimodal. However, for multimodal distributions of $\boldsymbol{z}$, the central estimate will not remain the same for all such ϵ. For instance, due to discreteness of the distribution of $\boldsymbol{z}$, for some ϵ, the neighborhood of the global mode may contain just a few values of $\boldsymbol{z}$ (other than the global mode), while for the same ϵ, the neighborhood of some local mode may contain many more values. This would yield the local mode as another central estimate. Thus, by allowing ϵ to vary uniformly over $(0,1)$, all the modes of the distribution of $\boldsymbol{z}$ can be detected, including the global mode, the latter obtained by letting $\epsilon\to 0$.

3.2 Construction of desired credible regions of clusterings

Given a central estimate $\boldsymbol{z}^*$, one can then obtain, say, an approximate 95% posterior density credible region as the set $\{\boldsymbol{z} \in \mathbb{F}^n : d(\boldsymbol{z}, \boldsymbol{z}^*) < \epsilon^*\}$, where ϵ^* is such that

$$P\left(\{\boldsymbol{z} \in \mathbb{F}^n : d(\boldsymbol{z}, \boldsymbol{z}^*) < \epsilon^*\}\right) \approx 0.95. \qquad (12)$$

In Eq.(12) ϵ^* can be chosen adaptively by starting with $\epsilon^* = 0$ and then slightly increasing ϵ^* by a quantity ζ until (12) is satisfied. In our case we may chose $\zeta = 10^{-10}$.

Approximate Highest Posterior Density (HPD) regions can be constructed by taking the union of the highest density regions. We next discuss an adaptive methodology for constructing HPD regions[MHCI00].

3.3 Construction of desired HPD regions of clusterings

Assume that there are k modes, $\{\boldsymbol{z}_1^*, \ldots, \boldsymbol{z}_k^*\}$, obtained by varying ϵ of the neighborhoods $\{\boldsymbol{z} \in \mathbb{F}^n : d(\boldsymbol{z}, \boldsymbol{w}) < \epsilon\}; \boldsymbol{w} \in \mathbb{F}^n$, uniformly over the interval $(0, 1)$, and following the principle described in Section 3.1. Also consider k ϵ^*'s, $\{\epsilon_1^*, \ldots, \epsilon_k^*\}$. Consider the regions $R_j = \{\boldsymbol{z} \in \mathbb{F}^n : d(\boldsymbol{z}_j^*, \boldsymbol{z}) < \epsilon_j^*\}; j = 1, \ldots, k$. Set, initially, $\epsilon_1^* = \epsilon_2^* = \cdots = \epsilon_k^* = 0$.

Step 1 For $i = 1, \ldots, N(= 2^n)$, if the $\boldsymbol{z}_i \in \mathbb{F}^n$ does not fall in R_j for some j, then increase ϵ_j^* by a small quantity, say, ζ. As before, we may chose $\zeta = 10^{-10}$.

Step 2 Calculate the probability of $\cup_{j=1}^k R_j$ as $P = P\left(\cup_{j=1}^k R_j\right)$.

Step 3 Repeat steps (i) and (ii) until $P \approx 0.95$ or any desired probability.

Step 1 implicitly assumes that, since $\boldsymbol{z}_i \notin R_j$, R_j must be a region with low probability, so its expansion is necessary to increase the probability. This expansion is achieved by increasing ϵ_j^* by ζ. This step also ensures that the sets R_j are selected adaptively, by adaptively increasing ϵ_j^*. The final union of the C_j's is the desired approximate HPD region.

4 Future work

We have presented the theoretical outline of our proposed scheme. Immediate future work is to finish the simulations and see how we stand with respect to existent attacks based on attack complexities.

We have plan to extend our scheme for cipher only attach even when the function (cipher) is unknown.

References

[CG92] George Casella and Edward I George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

[Hey02] Howard M. Heys. A tutorial on linear and differential cryptanalysis. *Cryptologia*, 26(3):189–221, July 2002.

[Jun05] Pascal Junod. *Statistical cryptanalysis of block ciphers*. PhD thesis, Citeseer, 2005.

[Kul76] S. Kullback. Statistical methods in cryptanalysis. *Aegean Park Press*, 1976.

[Lin65] D. V. Lindley. *Introduction to Probability and Statistics from a Bayesian Viewpoint*. Cambridge University Press, 1965.

[MHCI00] Qi-Man Shao Ming-Hui Chen and Joseph G. Ibrahim. *Monte Carlo Methods in Bayesian Computation*. Springer-Verlag New York, 2000.

[Pre89] S. J. Press. Bayesian statistics: Principles, models and applications. *Rinehart and Winston*, 1989.

[SMD11] S. Bhattacharya S. Mukhopadhyay and K. Dihidar. On bayesian "central clustering" : Application to landscape classification of western ghats. *Annals of Applied Statistics*, 2011.