CONVERGENCE OF CYCLIC RANDOM WALKS WITH AN APPLICATION TO CRYPTANALYSIS

CLIFFORD BERGMAN[1] AND SUNDER SETHURAMAN[2]

Imagine that you and some friends are playing a version of roulette. The wheel is divided into 36 sectors, alternately colored red and black. Before spinning the wheel, the contestant chooses a color and then wins or loses depending on whether or not his color comes up.

You, the master player, have honed an ability to spin the wheel exactly 3620° with high probability. Thus, if the wheel is initially on a red sector, then after your spin, it will again be on a red sector, and similarly for black. Of course, nobody's perfect, so let us say that 90% of your spins return the wheel to the same color on which they begin.

After you've cleaned out your friends a couple of times, they begin to wise up. One of them proposes a small change in the rules. Instead of a single spin, the contestant must spin the wheel 10 consecutive times. It is only if his initial guess matches the outcome after the tenth spin that he wins the game.

Is this fellow on to something? Will the new rule blunt your advantage? Let us assume that you continue to bet on the wheel's starting color, and think of each spin as a coin toss in which the probability of 'heads' is 0.9 (i.e., the wheel returns to its starting color after one spin). Then you will win the game if the number of tails after 10 tosses is an even number. The probability of this is easily computed to be $\sum_{k=0}^{5} \binom{10}{2k} (.1)^{2k} (.9)^{10-2k} \approx 0.55$. It seems clear from Figure 1 that as the required number of spins increases, your advantage diminishes. When used with a large number of spins, the game resembles a fair coin-toss, no matter how biased is a single spin.

The behavior of the "bias" of an iterated Bernoulli variable when computed modulo 2, and generalizations to iterations modulo m for $m > 2$, is the subject of this article. This equalizing phenomenon has been understood at least since the 1950's in the context of cyclic random walks, Feller [7, section 16.2(d)]; random number generation, Horton and Smith [14] and Dvoretsky and Wolfowitz [5]; and card-shuffling, Aldous and Diaconis [1], among other

2000 *Mathematics Subject Classification.* primary 60B10; secondary 60B15, 94B60.

Key words and phrases. random walk, circulant matrix, DES cipher, cyclic group.

Address: 400 Carver Hall, Department of Mathematics, Iowa State University, Ames, IA 50011

[1]E-mail: cbergman@iastate.edu.

[2]E-mail: sethuram@iastate.edu. Research supported in part by NSF grant DMS-0071504.

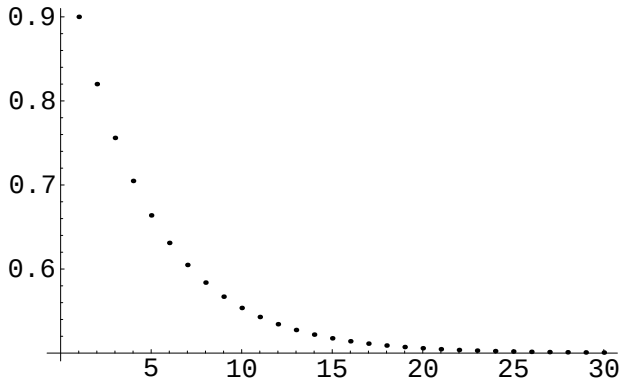

FIGURE 1. probability of winning the roulette game as a function of the number of spins

areas. In recent times, this behavior has also been exploited in cryptanalysis, Matsui [16], which we shall explain in Section 1. We had fun revisiting this old topic and the purpose of this note is to survey some of the interesting results and to give, in some cases, simpler and more probabilistic proofs than are found in the literature.

1. The Basic Problem

Let us formalize the above discussion as follows. Assume we have a coin for which 0 (Heads) occurs with probability p and 1 (Tails) occurs with chance $q = 1 - p$. Let $X_1, X_2, \ldots$ be a sequence of independent coin-tosses. Let the operation $\oplus$ represent the "exclusive-or" operation. In other words, $X_1 \oplus X_2$ is nothing more than $X_1 + X_2 \bmod_2$,

$$X_1 \oplus X_2 = \begin{cases} 0 & \text{when } X_1 = X_2 = 0 \text{ or } X_1 = X_2 = 1 \\ 1 & \text{when } X_1 \neq X_2. \end{cases}$$

Our fundamental observation is that no matter how biased the individual tosses, their exclusive-or resembles a fair coin asymptotically.

Proposition 1. *Let $X_1, X_2, \ldots$ be a sequence of independent, identically distributed coin tosses with probability of 0 equal to $0 < p < 1$. Then $\lim_{n\to\infty} P(X_1 \oplus X_2 \oplus \cdots \oplus X_n = 0) = \frac{1}{2}$.*

If we use μ to denote the probability distribution of a fair coin, then the proposition says that as n tends to infinity, the sequence $X_1 \oplus X_2 \oplus \cdots \oplus X_n$, *converges in measure* to μ, in symbols, $X_1 \oplus X_2 \oplus \cdots \oplus X_n \Rightarrow \mu$.

We shall present several proofs of this proposition: a combinatorial proof, one using Markov chains, and later an "eigenvalue" argument which provides some additional insight into the phenomenon.

Proof. For the combinatorial proof, observe that the variable $X_1 \oplus X_2 \oplus \cdots \oplus X_n$ is equal to 0 precisely when an even number of the individual

tosses return Heads. From the binomial theorem,

$$\begin{aligned}(q+p)^n + (q-p)^n &= \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} + \sum_{k=0}^{n} (-1)^k \binom{n}{k} p^k q^{n-k} \\ &= 2 \sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n}{2k} p^{2k} q^{n-2k} \\ &= 2P(X_1 \oplus \cdots \oplus X_n = 0).\end{aligned}$$

Of course, $q+p = 1$ and $0 < |q-p| < 1$. Thus, for large n, the left-hand side of the last equation is close to 1. Hence, $P(X_1 \oplus X_2 \oplus \cdots \oplus X_n = 0) \approx 1/2$.

For our second proof, we begin with the sequence $Z_n = X_1 \oplus \cdots \oplus X_n$, for $n = 1, 2, 3, \ldots$ This is a Markov sequence with transition matrix

$$P = \begin{pmatrix} p & q \\ q & p \end{pmatrix}.$$

In other words, $P(Z_{n+1} = 0 \mid Z_n = 0) = p$, $P(Z_{n+1} = 0 \mid Z_n = 1) = q$, $P(Z_{n+1} = 1 \mid Z_n = 0) = q$, and $P(Z_{n+1} = 1 \mid Z_n = 1) = p$. Thus

$$\begin{aligned}P(Z_n = 0) &= pP(Z_{n-1} = 0) + qP(Z_{n-1} = 1) \\ &= q + (p-q)P(Z_{n-1} = 0).\end{aligned}$$

The last equality follows from the fact that $P(Z_{n-1} = 0) + P(Z_{n-1} = 1) = 1$. Iterating this expression yields

$$\begin{aligned}P(Z_n = 0) &= q \sum_{r=0}^{n-2} (p-q)^{n-1} + (p-q)^{n-1} p \\ &= \frac{q}{1-(p-q)} + \left[p - \frac{q}{1-(p-q)} \right] (p-q)^{n-1},\end{aligned}$$

for $n \geq 2$. As $q/[1-(p-q)] = 1/2$, the above quantity converges to $1/2$ completing the Markov chain proof. □

In Proposition 1 we assumed that the coin tosses were both independent and identically distributed. In fact, a far weaker condition than identical distribution is sufficient for the proposition to hold. To make this precise, we introduce a quantity which measures the bias of a Bernoulli variable.

Definition 2. Let X be a random variable taking on two values with probabilities p and $q = 1-p$. Then the *bias of* X, denoted $B(X)$, is the quantity $|p-q|$.

The bias $B(X)$ has several nice properties. First, it is symmetric in p and q. Second, we always have $0 \leq B(X) \leq 1$, attaining the lower bound precisely when X is a fair coin and the upper bound when X is a constant (i.e., $p = 1$ or $q = 1$). Third, $B(X)$ is the magnitude of the smallest eigenvalue of the matrix P. Indeed, 1 and $p-q$ are eigenvalues with respective eigenvectors $(1,1)$ and $(1,-1)$. The utility of the bias is seen from the following identity.

Proposition 3. *Let X and Y be independent coin-flips, and let $Z = X \oplus Y$. Then $B(Z) = B(X) \cdot B(Y)$.*

Before proving this result, we show how it immediately implies our earlier claim (Proposition 1). Applying the above identity, we obtain

$$\big|P(Z_n = 0) - P(Z_n = 1)\big| = |p - q|^n$$

which vanishes (exponentially) as n tends to infinity.

Proof of Proposition 3. The identity can certainly be checked by direct computation, but for further development we use the following argument. Since the particular values taken on by X play no role in $B(X)$, let us identify Heads with 1 and Tails with -1. With this identification, the bias $B(X)$ is nothing but $|E[X]|$, the magnitude of the expected value of X. Furthermore, $X \oplus Y$ becomes the product $X \cdot Y$. Since expected value respects the product of independent observations, $B(XY) = |E[XY]| = |E[X]| \cdot |E[Y]| = B(X)B(Y)$. □

By employing the bias, we can improve Proposition 1 dramatically.

Corollary 4. *Let $X_1, X_2, \ldots$ be a sequence of (non-identically distributed) independent coin-flips. Let $Z_n = X_1 \oplus \cdots \oplus X_n$. Then,*

$$Z_n \Rightarrow \mu \quad \text{iff} \quad \lim_{n\to\infty} \prod_{i=1}^{n} B(X_i) = 0.$$

It is interesting to note that the condition at right allows for the coin-flips to become increasingly biased at some rate.

2. An Application to Cryptanalysis

Our interest in the topic of this paper was first piqued while reading Matsui's exposition of his linear cryptanalysis of the DES cipher, [16]. In Matsui's paper, a crucial role is played by the "piling-up lemma", which is essentially our Proposition 3. In this section we explain this phenomonon.

Cryptography is the art and science of designing ciphers—methods for communicating secretly. Cryptanalysis by contrast, is the opposing art of cracking ciphers—reading the secret communications without authorization. DES (the "Data Encryption Standard") has been, for the past 25 years, the cipher most commonly employed in commercial products. For this reason, an enormous amount of effort has been expended attempting to crack DES. Linear cryptanalysis is the most effective technique yet devised in this regard. In order to explain how it works, we must first describe the DES cipher in a few paragraphs—no easy task!

In mathematical terms, a cipher is nothing but an invertible function $E\colon \mathcal{P} \to \mathcal{C}$. The sets $\mathcal{P}$ and $\mathcal{C}$ are the spaces of *plaintext blocks* and *ciphertext blocks,* respectively. Since it is inadvisable for each individual user to construct his or her own encryption function, a modern cipher is conceived

x	0000	0001	0010	0011	0100	0101	0110	0111
$g_0(x)$	1101	0000	1011	0111	0100	1001	0001	1010
$g_1(x)$	0001	0100	1011	1101	1100	0011	0111	1110
x	1000	1001	1010	1011	1100	1101	1110	1111
$g_0(x)$	1110	0011	0101	1100	0010	1111	1000	0110
$g_1(x)$	1010	1111	0110	1000	0000	0101	1001	0010

FIGURE 2. Permutations for baby lucifer

as a cryptosystem: an indexed family $\mathbb{E} = \{ E_k : k \in \mathcal{K} \}$ of invertible functions. The set $\mathcal{K}$ is called the *key space.* In practice, this is implemented as a single function $\mathbb{E}\colon \mathcal{K} \times \mathcal{P} \to \mathcal{C}$, with $E_k(x) = \mathbb{E}(k, x)$. The algorithm describing $\mathbb{E}$ is known to the public. In order to communicate securely, two parties need only choose a key at random from $\mathcal{K}$ and keep it secret.

In most modern cryptosystems, $\mathcal{P} = \mathcal{C}$ and the function E_k is constructed as $F_{s_r} \circ F_{s_{r-1}} \circ \cdots \circ F_{s_2} \circ F_{s_1}$, where each s_i is a subkey derived from the original key k according to an algorithm called the *key schedule.* The integer r is the number of *rounds* in the cipher. The idea is that each F_{s_i} is a relatively simple permutation of the set $\mathcal{P}$ and once again, there is a single function $\mathbb{F}$ with $F_s(x) = \mathbb{F}(s, x)$. Even though the individual permutations are simple, when enough of them are composed, the resulting permutation (on a very big set) appears to be quite complex.

Before discussing DES, we will demonstrate these ideas on a very simple cryptosystem we call baby lucifer. The lucifer cipher, like DES, was developed at IBM, and is considered to be the ancestor of DES. Our baby lucifer is a (highly) abbreviated version of lucifer as it was described in [8]. In this cipher, $\mathcal{P} = \mathcal{C} = \mathcal{K} = \{0,1\}^8$, in other words each plaintext block, ciphertext block and key is a bitstring of length 8. The raw materials consist of two permutations g_0, g_1 of the set $\{0,1\}^4$. Those functions are given in Figure 2.

Baby lucifer is a 4-round cipher. Let $k = k_1 k_2 \ldots k_8$ denote an 8-bit key. We define the subkeys s_i to be $k_{2i-1}k_{2i}$ for $i = 1, 2, 3, 4$. For a subkey $s = uv$ and an input block $x = x_1 x_2 \ldots x_8$ the function $F_s(x)$ is defined to be the following pair of operations.

1. $y_1 y_2 y_3 y_4 = g_u(x_1 x_2 x_3 x_4) \qquad y_5 y_6 y_7 y_8 = g_v(x_5 x_6 x_7 x_8)$
2. $F_s(x) = y_7 y_8 y_5 y_2 y_1 y_3 y_6 y_4$.

For example, suppose the plaintext block $x = 01110010$ is to be encrypted with the key $k = 01110110$. The first two subkeys are 01 and 11. The computation involved in the first two rounds looks like this.

$$0111\,0010 \xrightarrow{(g_0,g_1)} 1010\,1011 \longrightarrow 1110\,1100$$

$$1110\,1100 \xrightarrow{(g_1,g_1)} 1001\,0000 \longrightarrow 00001001$$

The interested reader can verify that the states after the subsequent rounds are 11111011 and 00100110. This last string would be the output of the

entire cipher on the input string. Despite its small size, baby lucifer does a good job of scrambling the data.

We shall not attempt to describe DES in detail. The sets $\mathcal{P}$ and $\mathcal{C}$ are both taken to be $\{0,1\}^{64}$, while $\mathcal{K} = \{0,1\}^{56}$. There are 16 rounds to the cipher. DES is a far more complicated cipher than is baby lucifer. The precise details of the function $\mathbb{F}(s,x)$ and the key schedule are quite complex and unnecessary for our discussion. A full description can be found in any reference on cryptography, for example Menezes et al. [17].

Now we turn to the issue of cryptanalysis. Let us imagine that Alice and Bob have agreed on a secret key $k \in \mathcal{K}$ and are sending encrypted messages $y_i = E_k(x_i)$, $i = 1, 2, \ldots, t$ to each other. A third party named Eve obtains all of the pairs (x_i, y_i) and wishes to use this information to determine k. This is called a *known plaintext attack* since the evesdropper knows not just some ciphertext, but the corresponding plaintext as well.

Of course, one approach Eve might take is to exhaustively check each possible key, usually called a *brute-force* attack. It can be described using the following pseudocode.

```
for each l ∈ K do
    if E_l(x_i) = y_i for i = 1, 2, ..., t then
      return(Key=l).
```

If Alice and Bob are using baby lucifer, this little routine will test at most $|\mathcal{K}| = 2^8 = 256$ keys, which is feasible even by hand. This is one reason that a modern cipher requires a very large key space. On the other hand, if Alice and Bob communicate using DES, then $|\mathcal{K}| = 2^{56}$. In this case the loop will take a very long time to execute. Linear cryptanalysis is an improved attack that reduces the number of individual keys that must be tested.

We shall sketch the attack for a general cryptosystem that has r rounds and a block size of n bits. Let $K \in \mathcal{K}$ and $X \in \mathcal{P}$ denote uniform random variables and set $Y = E_K(X)$ and $Z = (F_{S_r})^{-1}(Y)$. Note that Z depends entirely on X and K. We can think of $X = (X_1, X_2, \ldots, X_n)$, $Z = (Z_1, \ldots, Z_n)$ and $K = (K_1, \ldots, K_m)$ as random sequences in the field $\mathbf{GF}(2)$ of integers modulo 2. Consider linear functions $U(X)$, $V(Z)$ and $W(K)$ over the field $\mathbf{GF}(2)$, that is

$$\begin{aligned} U(X) &= c_1X_1 + c_2X_2 + \cdots + c_nX_n \\ V(Z) &= d_1Z_1 + d_2Z_2 + \cdots + d_nZ_n \\ W(K) &= e_1K_1 + e_2K_2 + \cdots + e_mK_m \end{aligned}$$

with coefficients coming from $\mathbf{GF}(2)$ and define $L(X,Z,K) = U(X) + V(Z) + W(K)$. Then the value of $L(X,Z,K)$ behaves as a random variable on $\{0,1\}$ and has a bias between 0 and 1. The attack depends on the existence of a linear function L whose bias $B(L)$ is "large" and which depends on relatively few of the bits comprising K.

In a good cryptosystem, we expect the behavior of different keys to be indistinguishable. Thus, it is reasonable to assume that for any *fixed* key k,

the bias of $L(X, Z, k)$ should be approximately equal to $B(L)$. Notice also that the assertion that $L(X, Z, k)$ is biased means that the equation

$$U(X) + V(Z) = W(k) \tag{1}$$

is biased in the sense that it is either true or false with high probability. If k is an unknown key, then $W(k)$ is simply an unkown constant (either 0 or 1), so equation (1) simply asserts that $U(X)+V(Z)$ is also a biased random variable.

Suppose for the moment that the biased expression L has been found. Eve proceeds as follows. To save a subscript, let (x, y) denote any of her obtained pairs (x_i, y_i). She takes a guess, $\widetilde{S}_r$ at the subkey s_r. Using her guess she computes $\widetilde{Z} = F^{-1}_{\widetilde{S}_r}(y)$. If she has guessed right, she expects that the values of $U(x)+V(\widetilde{Z})$ will exhibit the bias $B(L)$ as (x, y) ranges over all of the pairs (x_i, y_i). On the other hand, if she has guessed wrong, then the bits in x and $\tilde{Z}$ will be independent, and Eve expects to see no bias in the values of $U(x)+V(\widetilde{Z})$. In this way, Eve can determine whether her guess at s_r is correct. This enables a classic divide-and-conquor strategy: try every possible value for S_r until the correct one is found. This determines some portion of the corrrect key k. Then try every possible combination for the remaining bits. If the length of the key is m bits and the length of a subkey is j bits then this strategy requires a search of $2^j + 2^{m-j} \ll 2^m$ keys in general.

The success of this strategy hinges on finding an appropriate function $L(X, Z, K)$ and determining its bias. Let us write $Z^0 = X$ and recursively define $Z^i = F_{K_i}(Z^{i-1})$. Thus Z^i denotes the internal state of the cipher after the i^{th} round. Matsui's suggested that by studying the structure of the round function $\mathbb{F}$, one can chose a series of linear functions $L_i(Z^{i-1}, Z^i, S_i) = U_i(Z^{i-1}) - V_i(Z^i) + W_i(S_i)$ for $i = 1, \ldots, r-1$ that are *linked* in the sense that $V_i = U_{i+1}$. In this way one obtains a telescoping sum

$$L(X, Z, K) = \sum_{i=1}^{r-1} L_i(Z^{i-1}, Z^i, S_i). \tag{2}$$

The bias of L can now be determined by an application of Proposition 3. If we assume that the variables $L_i(Z^{i-1}, Z^i, S_i)$ behave independently, then $B(L) = B(L_1) \cdot B(L_2) \cdots B(L_{r-1})$. Each function L_i has (hopefully) only a few nonzero coefficients, so its bias can be computed directly.

We shall illustrate this procedure using baby lucifer. Let A_1, A_2, A_3, A_4 denote random input bits and K a random key bit. Define $B_1B_2B_3B_4 = g_K(A_1A_2A_3A_4)$ (see Figure 2). Consider the linear combinations

$$\begin{aligned} M_1 &= A_4 + B_1 + B_2 + B_3 + K \\ M_2 &= A_1 + A_2 + B_3 + K \\ M_3 &= A_1 + A_2 + A_4 + B_2. \end{aligned} \tag{3}$$

By checking all 32 combinations of A_1, A_2, A_3, A_4, K, we see that $M_1 = 1$ with probability $24/32$, hence has bias equal to $(24-8)/32 = 1/2$. Similarly, $M_2 = 1$ with probability $22/32$ so has bias $3/8$ and $M_3 = 0$ with probability $24/32$ and has bias $1/2$.

We are now ready to construct the functions L_1, L_2, L_3 corresponding to the first three rounds of baby lucifer. By applying M_3 to the first four bits in round 1, we obtain the expression

$$L_1 = Z_1^0 + Z_2^0 + Z_4^0 + Z_4^1 + K_1, \qquad B(L_1) = 1/2.$$

(Don't forget that the second output bit of the function g_s becomes the *fourth* output bit of the round.) Similarly, by applying M_1 to the first four bits of round 2 yields

$$L_2 = Z_4^1 + Z_5^2 + Z_4^2 + Z_6^2 + K_3, \qquad B(L_2) = 1/2.$$

By applying M_1 to the first four and M_2 to the second four bits of round three yields two expressions

$$\begin{aligned} L_3' &= Z_4^2 + Z_5^3 + Z_4^3 + Z_6^3 + K_5, & B(L_3') &= 1/2 \\ L_3'' &= Z_5^2 + Z_6^2 + Z_1^3 + K_6, & B(L_3'') &= 3/8. \end{aligned}$$

Adding these two expressions together yields L_3 with a bias of $3/16$. But more to the point, adding all four expressions together yields

$$\begin{aligned} L &= X_1 + X_2 + X_4 + Z_1 + Z_4 + Z_5 + Z_6 + K_1 + K_3 + K_5 + K_6 \\ B(L) &= \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{3}{8} = \frac{3}{64} \end{aligned} \tag{4}$$

Eve can use the expression in (4) to determine the unknown key. For each of the four choices of $k_7 k_8$, she estimates the bias of $X_1 + X_2 + X_4 + Z_1 + Z_4 + Z_5 + Z_6$ using several pairs (x_i, y_i). One of these four choices should stand out as the one inducing a bias on this expression. Once $k_7 k_8$ has been determined, Eve can exhaustively try all candidates for the first six key bits. In this way, Eve will test $2^2 + 2^6 = 68$ keys, which is a savings of about 50% on a brute-force search.

Actually, Eve can do a bit better. By reviewing the compuations in equations (3), we see that the expression in (4) is biased toward the value 1. Therefore, under the correct key, the value of $X_1+X_2+X_4+Z_1+Z_4+Z_5+Z_6$ should be biased toward $1+k_1+k_3+k_5+k_6$. This allows Eve to obtain one bit for free since she has already determined which way the former expression is biased. Thus Eve's total work is only $2^2 + 2^5 = 36$ keys.

Recall that baby lucifer has an 8-bit block size, an 8-bit key and 4 rounds. DES, by contrast has a 64-bit block, 56-bit key and 16 rounds. The basic approach to finding a biased linear expression is the same for DES as it is for baby lucifer, but the details are somewhat more complex. The expression Matsui found has 30 variables and a bias of approximately $1.49 \cdot 2^{-23}$. This is not very large, but it does give Eve something to work with.

Matsui was also able to choose V_{15} so that it had very few nonzero coefficients. Consequently, only 6 bits of $\widetilde{S}_{16}$ have any influence on the value of $U(X) + V(\widetilde{Z})$. (This is partly a result of the particular design of DES.) Thus by trying all 64 possible 6-bit strings against all of the pairs (x_i, y_i) Eve can determine 6 bits of the actual key. She can also use this procedure in reverse—that is think of decrypting each y_i to get x_i—to determine 6 more bits of the key. And just as in the baby lucifer example, she can find an additional two bits for free along the way. This still leaves 42 bits, which she must find by exhaustive search. This certainly requires a great deal of computation, but the total number of combinations considered, $2^6 + 2^6 + 2^{42}$, is far less than the 2^{56} that is necessary for a simple-minded brute-force attack on DES.

To put this in human terms: if some computer requires 1 year to execute the brute-force attack, then that same computer could use linear cryptanalysis to find the key in about a half-hour. (For those interested in the state-of-the-art, in 1998, the Electronic Freedom Foundation [6] constructed a special-purpose DES-cracking machine that can find a key in about 12 hours. It uses nothing but a parallelized version of brute-force search.)

There is one other consideration when comparing linear cryptanalysis to the brute-force attack: how large must t (the number of pairs (x_i, y_i)) be in order for the attack to succeed with high probability? It is not hard to see that $t = 2$ is sufficient to ensure that the brute-force attack will almost surely find only the correct key (and no "false alarms"). On the other hand, for linear cryptanalysis, Eve needs enough pairs to be sure that the bias in Equation (2) is detectable. Using a standard normal approximation, one can see that to detect the bias with probability .95 requires that t be approximately 2^{46}. Thus, in order to mount this attack, Eve would need to intercept (and Alice and Bob would have to send!) a staggering amount of data. For this reason, linear cryptanalysis has never been considered a practical threat to DES.

One interesting aspect of this application is how it differs from most other applications of the ideas of this paper that have appeared in the literature. In those discussions, the emphasis is on the convergence of an iterated variable to the uniform distribution, whereas here we are trying to find iterated variables that are as far from uniform as possible.

In this discussion, we have omitted several details of DES as well as of the attack itself. In addition to Matsui's paper [16], the interested reader may wish to consult Harpes et al. [11] for a more complete treatment as well as a generalization.

3. Iterations Modulo m

As the exclusive-or operation can be identified with mod_2 addition, it is natural to wonder about mod_m-generalizations of the results in Section 1.

This situation occurs in applications such as card shuffling and random number generation. We provide a few references in Section 4.

Let $X_1, X_2, \ldots$ be a sequence of independent rolls of m-sided dice where the distribution of X_i is given by

$$X_i = l \qquad \text{with chance } p_i(l), \qquad \text{for } l = 0, 1, \ldots, m-1.$$

As before, we let μ denote the uniform distribution on $\mathbb{Z}_m = \{0, 1, 2, \ldots, m-1\}$, i.e., $\mu(l) = 1/m$ for $0 \leq l \leq m-1$. Does $Z_n = X_1 + \cdots + X_n \bmod_m$ converge to μ? In fact, this will be the case when "degeneracies" are avoided.

Although the combinatorics when $m > 2$ becomes difficult, we can still interpret the asymptotics of Z_n through Markov chain analysis when the observations $\{X_i : i \geq 1\}$ are identically distributed with common distribution $\mathbf{p} = \langle p(l) : 0 \leq l \leq m-1 \rangle$. As before, $\{Z_n : n \geq 1\}$ forms a Markov sequence. A moment's thought reveals that the transition matrix P now takes the form

$$P = \begin{pmatrix} p(0) & p(1) & \cdots & p(m-1) \\ p(m-1) & p(0) & \cdots & p(m-2) \\ & & \vdots & \\ p(1) & p(2) & \cdots & p(0) \end{pmatrix}. \tag{5}$$

It will be useful later to note that P is a $m \times m$ circulant matrix corresponding to the vector $\mathbf{p} = \langle p(0), p(1), \ldots, p(m-1) \rangle$ The analysis that the distribution of Z_n converges to μ, however, is not as direct as for the $m = 2$ case, but relies on the ergodic theorem for finite-state space Markov chains. We omit discussion of this theorem as the methods below are simpler and more general.

Let us now consider the general case when the X_i's are not necessarily identically distributed. For this analysis, it is convenient to work with a multiplicative cyclic group, rather than the additive group of integers modulo m. Let $\omega_m = e^{2\pi i/m}$ denote a primitive mth root of unity. The group $\mathbb{Z}_m = \{0, 1, \ldots, m-1\}$ under addition mod m is isomorphic to $\{1, \omega_m, \omega_m^2, \ldots, \omega_m^{m-1}\}$ under ordinary multiplication of complex numbers. This isomorphism maps an element $j \in \mathbb{Z}_m$ to ω_m^j. As in Proposition 3, $Z_n = X_1 + X_2 + \cdots + X_n \bmod_m$ corresponds to $X_1 \cdot X_2 \cdots X_n$. The sequence $\{Z_n : n \geq 1\}$ can be thought of now as a random rotation on the unit circle, or a random walk on the cyclic group of order m.

For a random variable X on the mth roots of unity with distribution $\mathbf{p}$, we now express the eigenvalues of the associated $m \times m$ circulant matrix P in terms of the moments of X (in analogy to the case $m = 2$).

Proposition 5. *Let X be a random variable on the mth roots of unity with distribution $\mathbf{p}$. Let P be the $m \times m$ circulant matrix corresponding to this distribution. Then, for $k = 0, 1, \ldots, m-1$, the eigenvalues λ_k and*

eigenvectors v_k of P are given by

$$\lambda_k = E[X^k]\ [= p(0) + p(1)\omega_m^k + p(2)\omega_m^{2k} + \cdots + p(m-1)\omega_m^{(m-1)k}],$$
$$v_k = \left(1, \omega_m^k, (\omega_m^k)^2, \ldots, (\omega_m^k)^{m-1}\right).$$

Proof. It is easy to check by direct calculation that if λ_k and v_k are defined as in the proposition then $Pv_k = \lambda_k v_k$, for $k = 0, 1, \ldots, m-1$. Now it is possible for the sequence $\lambda_0, \lambda_1, \ldots, \lambda_{m-1}$ to contain repeats. However, the matrix whose rows are the vectors v_k is

$$V = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega_m & \omega_m^2 & \cdots & \omega_m^{m-1} \\ 1 & \omega_m^2 & (\omega_m^2)^2 & \cdots & (\omega_m^2)^{m-1} \\ & & \vdots & & \\ 1 & \omega_m^{m-1} & (\omega_m^{m-1})^2 & \cdots & (\omega_m^{m-1})^{m-1} \end{pmatrix}$$

which we recognize as a Vandermonde matrix. The determinant of V is $\prod_{j<k}(\omega_m^k - \omega_m^j) \neq 0$. Consequently, $\{v_0, v_1, \ldots, v_{m-1}\}$ is a linearly independent set, and provides us with a basis of eigenvectors for P. □

Notice that the circulant matrix P depends only on the probabilities of the outcomes for the variable X and not on the values taken by X (although the equality of λ_k and $E[X^k]$ does depend on the values of X). In analogy with Definition 2, we write $\lambda_k(X)$, $k = 1, 2, \ldots, m-1$, for the eigenvalues of P.

Theorem 6. *Let $X_1, X_2, \ldots$ be independent variables on the m^{th} roots of unity, and let $Z_n = X_1 \cdot X_2 \cdots X_n$, for $n \geq 1$. Then,*

$$Z_n \Rightarrow \mu \iff \lim_{n\to\infty} \prod_{i=1}^{n} |\lambda_k(X_i)| = 0, \text{ for each } 1 \leq k \leq m-1.$$

Proof. Since $E[Z_n^k] = E[X_1^k] \cdot E[X_2^k] \cdots E[x_n^k] = \prod_{i=1}^n \lambda_k(X_i)$ by Proposition 5, it is enough to show that $Z_n \Rightarrow \mu$ if and only if the first $m-1$ moments of Z_n tend to zero.

One direction is trivial. The condition $Z_n \Rightarrow \mu$ is equivalent to the assertion $\lim_n P(Z_n = \omega_m^l) = 1/m$ for all $l = 0, 1, \ldots, m-1$. This gives, for $1 \leq k \leq m-1$, that

$$\begin{aligned} \lim_{n\to\infty} E[Z_n^k] &= \frac{1}{m}(1 + \omega_m^k + \cdots + \omega_m^{(m-1)k}) \\ &= \frac{1}{m} \cdot \frac{1 - \omega_m^{mk}}{1 - \omega_m^k} \end{aligned}$$

which vanishes as $\omega_m^m = 1$.

For the other direction, suppose that for every $1 \leq k < m$, we have that $\lim_{n\to\infty} E[Z_n^k] = 0$. Expanding, we have

$$\lim_{n\to\infty} \sum_{l=0}^{m-1} (\omega_m^l)^k P(Z_n = \omega_m^l) = 0, \qquad \text{for } k = 1, \ldots, m-1. \tag{6}$$

Consider the sequence

$$\mathbf{q}_n = \langle P(Z_n = 1), P(Z_n = \omega_m), \ldots, P(Z_n = \omega_m^{m-1}) \rangle$$

of points in $\mathbb{R}^m$ for $n \geq 1$. The assertion that $Z_n \Rightarrow \mu$ is equivalent to the claim that $\mathbf{q}_n \to \langle 1/m, 1/m, \ldots, 1/m \rangle$ as $n \to \infty$. Suppose this is not the case. That is, $\mathbf{q}_n$ has a subsequence $\{\mathbf{q}_{n_j} : j \geq 1\}$ that is bounded away from $\langle 1/m, \ldots, 1/m \rangle$. Since the sequence $\{\mathbf{q}_{n_j} : j \geq 1\}$ is uniformly bounded, it must itself have a convergent subsequence converging, let us say, to $\langle a_0, a_1, \ldots, a_m \rangle \neq \langle 1/m, \ldots, 1/m \rangle$.

The equations in (6) imply that

$$a_0 + \omega_m^k a_1 + \omega_m^{2k} a_2 \; + \; \cdots + \omega_m^{(m-1)k} a_{m-1} = 0$$

for $k = 1, \ldots, m-1$. In addition we have the equation

$$a_0 + \quad a_1 + \quad a_2 \; + \; \cdots + \qquad\qquad a_{m-1} = 1$$

since, for all n, $\sum_l P(Z_n = \omega_m^l) = 1$. Thus we have a system of m linear equations whose matrix (if the last equation is moved to the top) is none other than V, which we have already determined to be nonsingular. Consequently, the system has a unique solution. Since

$$\langle a_0, a_1, \ldots, a_{m-1} \rangle = \langle 1/m, 1/m, \ldots, 1/m \rangle$$

is certainly a solution to the system, it is the only one, contradicting our assumption. □

The convergence in measure holds in particular for "non-degenerate" identically distributed independent variables. Motivated by the above result, let us say that the distribution $\mathbf{p}$ of a random variable X taking values on $\mathbb{Z}_m$ is *non-degenerate* when $|\lambda_k(X)| < 1$ for all $1 \leq k \leq m-1$. [Note that by the triangle inequality, $|\lambda_k(X)| \leq p(0) + \cdots + p(m-1) = 1$.] In other words, the distribution is non-degenerate exactly when there is a "positive spectral gap," that is, a positive gap between the eigenvalue $\lambda_0(X) = 1$ and $\max_{1 \leq k \leq m-1} |\lambda_k(X)|$. We develop now concrete conditions for non-degeneracy of a random variable X.

Let us again think of $\mathbb{Z}_m$ as the additive group of integers modulo m. For $i, j \in \mathbb{Z}_m$, let $\langle j \rangle + i = \{jx + i : x = 0, 1, \ldots, m-1\}$ be the coset of the cyclic subgroup generated by j displaced by i.

Proposition 7. *Let X be a random variable on $\mathbb{Z}_m$ with distribution $\mathbf{p}$. Define $H = \{i \in \mathbb{Z}_m : p(i) > 0\}$. Then, for $1 \leq k \leq m-1$, we have that*

$$|\lambda_k(X)| = 1 \iff H \subset \langle j \rangle + i$$

where $i = \min(H)$ and $j = m/\gcd(m, k)$.

Proof. The "⇒" proof follows from two observations.

(i) Let z and w be nonzero complex numbers such that $|z+w| = |z| + |w|$. Then $z = rw$ for some real number $r > 0$. This is the law of cosines.

(ii) Let $p(1), \ldots, p(k)$ be positive real numbers, and $\alpha_1, \ldots, \alpha_k$ be distinct complex numbers of magnitude 1. If $|p(1)\alpha_1 + \cdots + p(k)\alpha_k| = p(1) + \cdots + p(k)$, then $k = 1$. We will show this momentarily.

Given (i) and (ii), we prove the proposition. Assume $|\lambda_k| = 1$ for some $1 \leq k \leq m-1$. Then $|\sum_{i \in H} p(i)\omega_m^{ik}| = 1$. By (ii), we have $i, l \in H$ implies $\omega_m^{ik} = \omega_m^{lk}$. Let now $i = \min(H)$, $l \in H$, $d = \gcd(m, k)$, and $j = m/d$. Then, $\omega_m^{ik} = \omega_m^{lk}$ implies m divides $(l-i)k$, which further implies m/d divides $(l-i)k/d$, and so m/d divides $(l-i)$, which finally gives $l \in \langle j \rangle + i$.

To finish the proof, we argue (ii). Let $k > 1$ be the least counter-example of the statement. Then applying the triangle inequality twice,

$$\begin{aligned}\big|p(1)\alpha_1 + p(2)\alpha_2 + \cdots + p(k)\alpha_k\big| &= p(1) + \big(p(2) + \cdots + p(k)\big) \geq \\ |p(1)\alpha_1| + \big|p(2)\alpha_2 + \cdots + p(k)\alpha_k\big| &\geq \big|p(1)\alpha_1 + \cdots + p(k)\alpha_k\big|.\end{aligned}$$

This implies that $p(2) + \cdots + p(k) = |p(2)\alpha_2 + \cdots + p(k)\alpha_k|$. Therefore, by assumption, $k - 1 = 1$, that is $k = 2$. By (i), we must have $p(2)\alpha_2 = rp(1)\alpha_1$ for some $r > 0$. This gives $\alpha_2 = r(p(1)/p(2))\alpha_1$. But, $|\alpha_1| = |\alpha_2| = 1$, and so $rp(1)/p(2) = 1$ and $\alpha_2 = \alpha_1$. This is a contradiction.

For the "$\Leftarrow$" argument, suppose $H \subset \langle j \rangle + i$, where $j = m/\gcd(m, k)$ and $i = \min(H)$. Then, m divides kj, and so $\omega_m^{k(jx+i)} = \omega_m^{ki}$ for all $x \in \mathbb{Z}_m$. Hence, $|\lambda_k| = |\omega_m^{ki} \sum_{l \in H} p(l)| = 1$. □

Corollary 8. *Let $X_1, X_2, \ldots$ be a sequence of identically distributed, independent random variables taking values in $\mathbb{Z}_m$ with distribution $\mathbf{p}$. Let $Z_n = X_1 + X_2 + \cdots + X_n \ mod_m$ for $n \geq 1$. Then*

$$Z_n \Rightarrow \mu \iff \mathbf{p} \text{ is non-degenerate.}$$

Proof. For $0 \leq k \leq m-1$, denote by λ_k the common value of $\lambda_k(X_i)$ for $i \geq 1$. By Theorem 6, $Z_n \Rightarrow \mu$ if and only if $|\lambda_k|^n$ converges to 0 for all $1 \leq k \leq m-1$. This can happen if and only if $|\lambda_k| < 1$ for all $1 \leq k \leq m-1$ which by Proposition 7 is equivalent to nondegeneracy of $\mathbf{p}$. □

It may be worth noting that if m is prime then the only proper subgroup of $\mathbb{Z}_m$ is $\{0\}$. Hence Proposition 7 tells us that the only degenerate random variables are the constant ones.

4. Concluding Remarks

Corollary 8 is a case of the well-known Ito-Kawade theorem which gives equivalent conditions on a distribution ν in a compact group for the n-fold convolutions ν^{*n} to converge to the Haar measure of the group; see Grenander [10], Heyer [12], and Högnäs and Mukherjea [13]; a different flavor of results can be found in Diaconis [3].

It seems, however, that the convergence of convolutions of non-identical distributions in general groups are not as well understood (cf. [13, Section

2.4]). But see there, in particular, Theorem 2.49 of [13] which gives a necessary and sufficient condition for convergence of convolutions of non-identical measures on discrete groups. In this context, Theorem 6 is a particular case of this result with respect to the finite group $\mathbb{Z}_m$ first proved by Dvoretzky and Wolfowitz [5] with Fourier methods which was motivated by random number generation questions. See also Aldous and Diaconis [1], and Goel and Gulati [9] for appliations to cardshuffling and statistics among other things.

Circulant matrices appear naturally in many applications. A proof of Corollary 8 purely in terms of circulant properties is found in Krafft and Schaefer [15]. The book Davis [2] is a comprehensive reference. See also Diaconis [4] for a discussion of interesting generalizations of these matrices and their use.

References

[1] Aldous, D., and Diaconis, P. (1986) Shuffling Cards and Stopping Times. *Amer. Math. Monthly* **93** 333–348.

[2] Davis, P. (1979) *Circulant Matrices* Wiley, New York.

[3] Diaconis, P. (1988) *Group Representations in Probability and Statistics.* Institute of Mathematical Statistics Lecture Notes-Monograph Series **11**, Hayward, CA.

[4] Diaconis, P. (1990) Patterned Matrices. *Proceedings of Symposia in Applied Mathematics, Ed. C. Johnson.* **40** 37–58.

[5] Dvoretzky, A., and Wolfowitz, J. (1951) Sums of Random Integers Reduced Modulo *m*. *Duke Math. Journal* **18** 501–507.

[6] EFF DES Cracker Project, `http://www.eff.org/descracker`

[7] Feller, W. (1968) *An Introduction to Probability Theory and its Applications I.* Wiley, New York.

[8] Feistel, H. (1973) Cryptography and computer privacy. *Scientific American* **228** (May 1973) 15–23.

[9] Goel, P., and Gulati, C. (2001) Monotone Decreasing Distance between Distributions of Sums of Unfair Coins and a Fair Coin. *Math. Sci.* **26** 34–40.

[10] Grenander, U. (1963) *Probabilities on Algebraic Structures.* Wiley, New York.

[11] Harpes, C., Kramer, G. and Massey, J. (1995) A Generalization of Linear Cryptanalysis and the Applicability of Matsui's Piling-up Lemma. *Advances in Cryptology–Eurocrypt '95,* Lecture Notes in Computer Science, vol. 921, pp. 24–38, Springer-Verlag, New York.

[12] Heyer, H. (1975) *Probability Measures on Locally Compact Groups.* Springer, Berlin-Heidelberg-New York.

[13] Högnäs, G., and Mukherjea, A. (1995) *Probability Measures on Semigroups.* Plenum Press, New York.

[14] Horton, H.B., Smith, R.T. (1949) A Direct Method for Producing Random Digits in any Number System. *Ann. Math. Stat.* **20** 82–90.

[15] Krafft, O., and Schaefer, M. (1990) Convergence of the Powers of a Circulant Stochastic Matrix. *Linear Alg. and Appl.* **127** 59–69.

[16] Matsui, M. (1993) Linear cryptanalysis method for DES cipher. *Advances in Cryptology—Eurocrypt '93,* Lecture Notes in Computer Science vol. 765, pp. 386–397, Springer-Verlag, New York.

[17] Menezes, A., van Oorschot, and Vanstone, S. (1997) *Handbook of Applied Cryptography.* CRC Press, Boca Raton.