Linear Complexity and Random Sequences

Rainer A. Rueppel
Swiss Federal Institute of Technology
8092 Zurich/Switzerland
currently: CMRR, University of California, San Diego
La Jolla, CA 92093/U.S.A.

Abstract:

The problem of characterizing the randomness of finite sequences arises in cryptographic applications. The idea of randomness clearly reflects the difficulty of predicting the next digit of a sequence from all the previous ones. The approach taken in this paper is to measure the (linear) unpredictability of a sequence (finite or periodic) by the length of the shortest linear feedback shift register (LFSR) that is able to generate the given sequence. This length is often referred to in the literature as the linear complexity of the sequence. It is shown that the expected linear complexity of a sequence of n independent and uniformly distributed binary random variables is very close to n/2 and, that the variance of the linear complexity is virtually independent of the sequence length, i.e. is virtually a constant! For the practically interesting case of periodically repeating a finite truly random sequence of length 2^m or 2^m-1, it is shown that the linear complexity is close to the period length.

F. Pichler (Ed.): Advances in Cryptology - EUROCRYPT '85, LNCS 219, pp. 167-188, 1986.

Linear Complexity and Random Sequences

Stream ciphers utilize deterministically generated "random" sequences to encipher the message stream. Since the running key generator is a finite state machine, the key stream necessarily is (ultimately) periodic. Thus the best one can hope for is to make the first period of a periodic key stream resemble the output of a binary symmetric source (BSS). A BSS is a device which puts out with equal probability a zero or a one independently of the previous output bits, or in other words, a BSS realizes a fair coin tossing experiment. (Note that we have tacitly assumed the sequences under investigation to be defined over GF(2)). The period of the key stream necessarily is a finite quantity. Thus we are confronted with the problem of characterizing the randomness of a finite sequence. But how can this be done in light of the fact that every finite output sequence of a BSS is equally likely? It seems difficult to define adequately the concept of randomness (in a mathematical sense) for finite sequences. Still, nearly everyone would agree that something like a "typical" output sequence of a BSS exists. A finite coin tossing sequence, for example, would "typically" exhibit a balanced distribution of single bits, pairs, triples, etc. of bits, and long runs of one symbol would be very rare. This in contrast to infinite coin tossing sequences, where local non-randomness is sure to occur. D.E. Knuth (Knut 81) discusses various concepts of randomness for infinite sequences and gives a short description of how randomness of a finite sequence could be defined. By the above typicality-argument, one is led naturally to the criterion of distribution properties. A finite sequence of length T may be called "random" if every binary k-tuple for all k smaller than some upperbound (e.g. logT) appears about equally often. The "randomness postulates" of S. Golomb (Golo 67) based on this definition have gained widespread popularity (especially in the cryptographic community). Golomb proposed the following three requirements to measure the randomness of a periodic binary sequence. First, the disparity between zeros and ones within one period of the sequence does not exceed 1. Second, in every period, $(1/2^i)$th of the total number of runs has length i, as long as there are at least 2 runs of length i. Third, the periodic autocorrelation function is two-valued. Every sequence which satisfied these three randomness requirements was called by Golomb a pseudo-noise (PN) sequence. But although

Golomb called his requirements "randomness postulates", they do not define a general measure of randomness for finite sequences. These "randomness postulates" rather describe almost exclusively the sequences which have a primitive minimal polynomial (since they have maximum possible period, they are also called maximum-length sequences or m-sequences). But this means that the so-called PN-sequences are highly predictable, if L denotes the degree of the primitive minimal polynomial of the PN-sequence under investigation, then only 2L bits of the sequence suffice to specify completely the remainder of the period of length 2^L-1. Clearly the idea of randomness also reflects the impossibility of predicting the next digit of a sequence from all the previous ones. An interesting approach to a definition of randomness of finite sequences based on this concept of unpredictability was taken by R. Solomonov (Solo 64) and A. Kolmogorov (Kolm 65). They chararacterized the "patternlessness" of a finite sequence by the length of the shortest Turing machine program that could generate the sequence. Patternlessness may be equated with unpredictability or randomness. This concept was further developed by P. Martin-Loef (Mart 66). A different approach to evaluating the complexity of finite sequences was given by A. Lempel and J. Ziv (Lemp 76). Instead of an abstract model of computation such as a Turing machine one could directly use a linear feedback shift register (LFSR) model and measure the (linear) unpredictability of a sequence (finite or periodic) by the length of the shortest LFSR which is able to generate the given sequence. This approach is particularly appealing since there exists an efficient synthesis procedure (the Berlekamp-Massey LFSR synthesis algorithm (Mass 69)) for finding the shortest LFSR which generates a given sequence. This length is also referred to as the linear complexity associated to the sequence. The following sequence obtained by the author in 31 trials with a fair swiss coin may serve as an illustration for the concept of linear complexity as measure of randomness (or linear unpredictability).

$$\tilde{s} = (10001111010000110111110100010100)^{\infty} \qquad (1)$$

In Fig. 1, we compare the dynamic behaviour of the linear complexity of the periodically repeated swiss coin sequence (1) to that of a PN-sequence of period 31. $\Lambda(s^n)$ denotes the linear complexity of the first n digit subsequence of $\tilde{s}$.

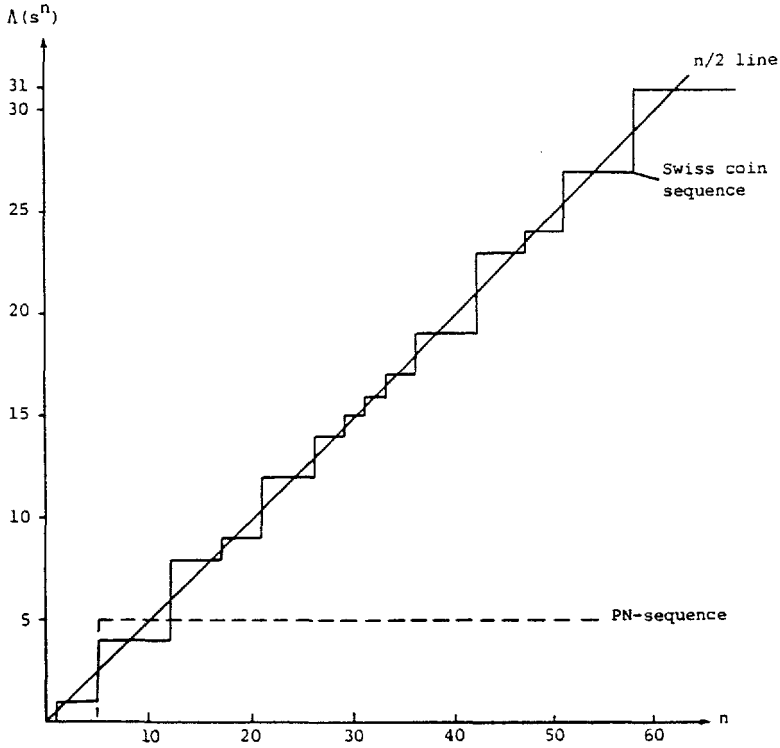

Fig. 1. Linear complexity profiles of the swiss coin sequence (1) and the PN-sequence generated by $<5,1+D^2+D^5>$ and initial state $[0,0,0,0,1]$

The linear complexity of the swiss coin sequence (1) grows approximately as n/2, where n denotes the number of processed bits, and stops at 31 which is the period of the sequence (1). Thus only the circulating shift register of length L = 31 is able to generate the swiss coin sequence. Conversely, the so-called PN-sequence of period 31 has a linear complexity of only 5 and is highly predictable. But note, a high linear complexity alone does not guarantee good randomness properties. As an example consider the sequence built by 30 consecutive 0's and an appended 1 which is periodically repeated. This sequence can also only be generated by the circulating shift register of length 31, but does not exhibit any randomness properties whatsoever. This could be seen in the associated linear complexity profile, in which the linear complexity remains at 0 until the 1 appears at the 31st position which causes the linear complexity to jump from 0 to 31 in one swoop. Consequently, we expect a "typical" random sequence to have associated a "typical" linear complexity profile closely following the n/2 line.

Let $s^n = s_0, s_1, \ldots, s_{n-1}$ denote a sequence of n independent and uniformly distributed binary random variables, and let $\Lambda(s^n)$ be the associated linear complexity. Our primary interest is in $N_n(L)$, the number of sequences of length n with linear complexity $\Lambda(s^n) = L$. Consider the basic recursion from $\Lambda(s^{n-1})$ to $\Lambda(s^n)$. The difference between the nth binary random variable s_{n-1} and the nth digit generated by the minimal-length LFSR which is able to generate s^{n-1} is called the next discrepancy δ_{n-1}. If the LFSR of length $\Lambda(s^{n-1})$ which generates s^{n-1} also generates s^n, then $\delta_{n-1} = 0$ and the linear complexity does not change. Conversely, if the LFSR of length $\Lambda(s^{n-1})$ which generates s^{n-1} fails to generate s^n, then $\delta_{n-1} = 1$ and the linear complexity increases when $\Lambda(s^{n-1})$ is smaller than $n/2$. The recursion describing the length change is basic to the LFSR-synthesis procedure (Mass 69):

$$\delta_{n-1} = 0 \qquad \Lambda(s^n) = \Lambda(s^{n-1}) \tag{2a}$$

$$\delta_{n-1} = 1 \qquad \begin{cases} \Lambda(s^n) = \Lambda(s^{n-1}) & \text{if } \Lambda(s^{n-1}) \geq \frac{n}{2} \\ \Lambda(s^n) = n - \Lambda(s^{n-1}) & \text{if } \Lambda(s^{n-1}) < \frac{n}{2} . \end{cases} \tag{2b}$$

Note that the linear complexity does not change (regardless of the value of the discrepancy) when $\Lambda(s^{n-1}) \geq \frac{n}{2}$. It is illuminating to represent graphically the linear complexity recursion (2) (see Fig. 2.)

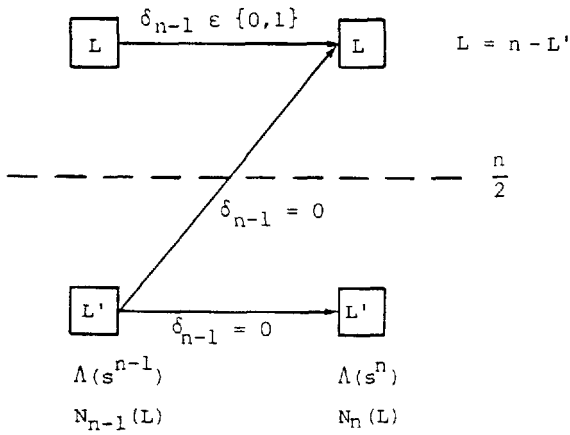

Fig. 2. Graphical illustration of the linear complexity growth process

From the diagram in Fig. 2, we may now directly read off the recursion for $N_n(L)$. If $\Lambda(s^{n-1}) = L' < \frac{n}{2}$, then $N_n(L') = N_{n-1}(L')$ since only one choice for s_{n-1} causes $\delta_{n-1} = 0$. The second choice for s_{n-1} causes $\delta_{n-1} = 1$ and thus transfers $N_{n-1}(L')$ sequences to the new complexity $L = n-L'$. If $\Lambda(s^{n-1}) = L > \frac{n}{2}$, then $\Lambda(s^n) = L$ (irrespective of δ_{n-1}) and $2N_{n-1}(L)$ sequences contribute to $N_n(L)$. The only exception to the sketched process in Fig. 2 occurs when n is even and $L = \frac{n}{2}$. In this case no path from $\Lambda(s_{n-1}) = L' < \frac{n}{2}$ may lead to $\Lambda(s^n) = L = \frac{n}{2}$, since $L = \frac{n}{2} = n-L'$ would require $L' = \frac{n}{2}$ which contradicts the assumption. We can now write the recursion for $N_n(L)$, the number of sequences of length n with linear complexity N, as

$$N_n(L) = \begin{cases} 2N_{n-1}(L) + N_{n-1}(n-L) & n \geq L > \frac{n}{2} \quad (3a) \\ 2N_{n-1}(L) & L = \frac{n}{2} \quad (3b) \\ N_{n-1}(L) & \frac{n}{2} > L \geq 0 \; . \quad (3c) \end{cases}$$

The initial conditions for the recursion (3) are $N_1(0) = N_1(1) = 1$. At any length n the total number of sequences is 2^n. In table 1, the values of $N_n(L)$ are listed for all positive $n \leq 10$.

L \ n	1	2	3	4	5	6	7	8	9	10
0	1	1	1	1	1	1	1	1	1	1
1	1	2	2	2	2	2	2	2	2	2
2		1	4	8	8	8	8	8	8	8
3			1	4	16	32	32	32	32	32
4				1	4	16	64	128	128	128
5					1	4	16	64	256	512
6						1	4	16	64	256
7							1	4	16	64
8								1	4	16
9									1	4
10										1

Table 1. Values of $N_n(L)$ for n = 1, ..., 10

The general form of $N_n(L)$ is easily guessed from table 1.

$$N_n(L) = \begin{cases} 2^{\min\{2n-2L,\,2L-1\}} & n \geq L > 0 \qquad (4a) \\ 1 & n > L = 0 \qquad (4b) \end{cases}$$

To show that this solution is correct, we first prove that the solution satisfies the recursion (3) for all $n > 1$.

Suppose $n \geq L > n/2$, then $N_n(L) = 2^{2n-2L}$, $N_{n-1}(L) = 2^{2n-2-2L}$ and $N_{n-1}(n-L) = 2^{2n-2L-1}$, since $n \leq 2L$ implies $2(n-L) < n-1$. These values satisfy recursion (3a) for all $n > 1$, as can be seen by substitution.

Suppose $L = n/2$, then $N_n(L) = 2^{2L-1}$ and $N_{n-1}(L) = 2^{2L-2}$, which satisfy recursion (3b) for all even $n > 1$.

Suppose $n/2 > L > 0$, then $N_n(L) = N_{n-1}(L) = 2^{2L-1}$ and the recursion (3c) is trivially satisfied for all $n > 1$.

By taking into account the initial conditions $N_1(0) = N_1(1) = 1$ the solution (4) is seen to yield the correct values for $n = 2$. Thus (4) is the solution to the recursion (3). We summarize the result in the following proposition.

Proposition 1. Distribution of $N_n(L)$

The number $N_n(L)$ of binary sequences $s^n = s_0, s_1, \ldots, s_{n-1}$ of length n having linear complexity exactly L is

$$N_n(L) = \begin{cases} 2^{\min\{2n-2L,\,2L-1\}} & n \geq L > 0 \\ 1 & n > L = 0 \quad . \end{cases}$$

The form of $N_n(L)$ for the general case of q-ary sequences may be found in (Gust 76) where the objective of that author was to evaluate the performance of the Berlekamp-Massey LFSR synthesis algorithm. Our interest is in

characterizing a "typical" random sequence by means of the associated linear complexity. Proposition 1 tells us that the vast majority of the possible binary sequences of length n will have linear complexity close to n/2. A quantity of independent interest, related to $N_n(L)$, is the number of semi-infinite sequences of linear complexity L or less, which we denote by N_L. For finite $L > 0$, Proposition 1. gives $N_\infty(L) = 2^{2L-1}$. Thus

$$N_L = 1 + \sum_{j=1}^{L} 2^{2j-1} \tag{5}$$

where the added 1 accounts for the allzero sequence, which has linear complexity L = 0. Evaluating the finite geometric series (5) yields

$$N_L = \frac{2}{3}\, 2^{2L} + \frac{1}{3}\,. \tag{6}$$

When we consider the tree corresponding to the set of all binary semi-infinite sequences, then at depth 2L every sequence of linear complexity L or less is characterized by the fact that the associated LFSR which may produce the sequence is unique. Hence the significance of (6) is that almost exactly 2/3 of all sequences of length 2L may be generated with an LFSR of length L or less. Both proposition 1. and the above argument on N_L suggest that any sequence of n randomly selected binary digits will "typically" have a linear complexity close to n/2. To obtain a precise characterization, we may compute the expected linear complexity of a sequence s^n of n independent binary random variables $s_0, s_1, \ldots, s_{n-1}$ (as emitted from a BSS).

$$E[\Lambda(s^n)] = \sum_{b^n} \Lambda(b^n)\ \ P(b^n) \tag{7}$$

where b^n denotes a particular realization of the coin tossing sequence s^n. Since each b^n is equally likely, the probability $P(s^n = b^n)$ is 2^{-n}. Therefore

$$E[\Lambda(s^n)] = 2^{-n} \sum_{b^n} \Lambda(b^n) = 2^{-n} L^*(n) \tag{8}$$

where we have introduced the symbol $L^*(n)$ for $2^n E[\Lambda(s^n)]$. The set of all b^n may be subdivided into equivalence classes according to the associated linear complexity. Thus we may rewrite the sum $L^*(n)$ in (8) as

$$L^*(n) = \sum_{L=1}^{n} \sum_{\{b^n:\Lambda(b^n)=L\}} L \quad . \tag{9}$$

The Lth equivalent class is easily identified to contain $N_n(L)$ elements. Thus

$$L^*(n) = \sum_{L=1}^{n} L\ N_n(L) \quad . \tag{10}$$

Replacing $N_n(L)$ by the solution given in proposition 1, we obtain

$$L^*(n) = \sum_{L=1}^{n} L2^{\min\{2n-2L,2L-1\}} \tag{11}$$

which may be subdivided into two sums according to the dominance of 2n-2L or 2L-1, which results in

$$L^*(n) = \sum_{L=1}^{\lfloor n/2 \rfloor} L2^{2L-1} + \sum_{L=\lceil \frac{n+1}{2} \rceil} L2^{2n-2L} \quad . \tag{12}$$

It is now possible to obtain a closed form expression for the finite sum in (12) by applying standard analytical methods. We illustrate the principle by evaluating

$$\sum_{j=1}^{m} j2^{2j-1} \quad . \tag{13}$$

First, we introduce a dummy variable I raised to the (j-1)st power,

$$\sum_{j=1}^{m} j\ I^{j-1} 2^{2j-1} \quad .$$

Now we integrate the sum with respect to I,

$$\sum_{j=1}^{m} I^j 2^{2j-1} \quad .$$

This is an ordinary geometric series whose sum is given by

$$2I \frac{I^m 2^{2m}-1}{I\ 2^2\ -1} \quad .$$

Differentiating this sum and setting I = 1, we obtain as the closed form solution for (13)

$$\sum_{j=1}^{m} j2^{2j-1} = \frac{(m+1)}{3} 2^{2m+1} - \frac{2}{9} (2^{2m+1} -1) \quad . \tag{14}$$

Because of the floor- and ceiling-functions in (12), it is convenient to distinguish between even and odd n. Let $L_e^*(n)$ and $L_0^*(n)$ denote the function L*(n) evaluated at even n and at odd n, respectively. Then by applying the standard techniques, as explained in the derivation of (14), to the individual sums in (12), we obtain for even n

$$L_e^*(n) = \{2^n (\frac{n}{3} - \frac{2}{9} + \frac{2}{9} 2^{-n})\} + \{2^n (\frac{n}{6} + \frac{4}{9} - 2^{-n} (\frac{n}{3} + \frac{4}{9}))\} \tag{15}$$

where the brackets {} enclose the values of the two distinct sums in (20).

In the case of odd n, we similarly obtain

$$L_0^*(n) = \{2^n (\frac{n}{6} - \frac{5}{18} + \frac{2}{9} 2^{-n})\} + \{2^n (\frac{n}{3} + \frac{5}{9} - 2^{-n} (\frac{n}{3} + \frac{4}{9}))\} . \tag{16}$$

Now it is straightforward to combine (8), (15) and (16) to obtain the desired expected linear complexity $E[\Lambda(s^n)]$. We summarize the result in the following proposition.

Proposition 2. $E[\Lambda(s^n)]$

The expected linear complexity of a sequence $s^n=s_0,s_1,\ldots,s_{n-1}$ of n independent and uniformly distributed binary random variables is given by

$$E[\Lambda(s^n)] = \frac{n}{2} + \frac{4+R_2(n)}{18} - 2^{-n}\left(\frac{n}{3}+\frac{2}{9}\right) \qquad (17)$$

where $R_2(n)$ denotes the remainder when n is divided by 2.

Proposition 2. confirms our suspicion that the linear complexity of a randomly selected sequence s^n can be expected close to n/2. Nevertheless, it is surprising how very close to half the sequence length that the expected linear complexity actually lies. For large values of n,

$$E[\Lambda(s^n)] \cong \frac{n}{2} + \frac{4+R_2(n)}{18} \qquad n \gg 1 \qquad (18)$$

which differs from n/2 by only an offset of 2/9 in the case of even n or 5/18 in the case of odd n. Besides the expectation, the variance of the linear complexity is a second key parameter suited for characterizing "typical" random sequences. The variance is defined as

$$\begin{aligned} Var[\Lambda(s^n)] &= E\left[\{\Lambda(s^n) - E[\Lambda(s^n)]\}^2\right] \\ &= E[\Lambda^2(s^n)] - E[\Lambda(s^n)]^2 \quad . \end{aligned} \qquad (19)$$

Following the same approach as for the derivation of $E[\Lambda(s^n)]$, the second moment $E[\Lambda^2(s^n)]$ is found to be (compare 12)

$$L^{2*}(n) = E[\Lambda^2(s^n)]2^n = \sum_{L=1}^{\lfloor n_2 \rfloor} L^2 2^{2L-1} + \sum_{L=\lceil \frac{n+1}{2} \rceil}^{n} L^2 2^{2n-2L} \qquad (20)$$

We apply again the standard technique of integration and differentation of the finite sums in (20) to obtain a closed form expression for $L^{2*}(n)$.

For analytical convenience, let $L_e^{2*}(n)$ and $L_o^{*2}(n)$ denote the function $L^{2*}(n)$ evaluated at even and odd n, respectively. We indicate the two distinct sums in (20) by enclosing them with brackets {}. In the case of even n, we obtain

$$L_e^{2*}(n) = \{2^{n+1}\left(\frac{1}{12}n^2 - \frac{1}{9}n + \frac{5}{27}\right) - \frac{10}{27}\}$$
$$+ \{2^n\left(\frac{1}{12}n^2 + \frac{4}{9}n + \frac{20}{27}\right) - \left(\frac{1}{3}n^2 + \frac{8}{9}n + \frac{20}{27}\right)\} \quad . \tag{21}$$

In the case of odd n, we obtain

$$L_o^{2*}(n) = \{2^n\left(\frac{1}{12}n^2 - \frac{5}{18}n + \frac{41}{108}\right) - \frac{10}{27}\}$$
$$+ \{2^n\left(\frac{1}{6}n^2 + \frac{5}{9}n + \frac{41}{54}\right) - \left(\frac{1}{3}n^2 + \frac{8}{9}n + \frac{20}{27}\right\} \quad . \tag{22}$$

Now it is straightforward to combine (20), (21), and (22) to obtain the desired closed form expression for the second moment of the linear complexity for all positive n:

$$E[\Lambda^2(s^n)] = \frac{1}{4}n^2 + \frac{4+R_2(n)}{18} + \frac{40+R_2(n)}{36}$$
$$- 2^{-n}\left(\frac{1}{3}n^2 + \frac{8}{9}n + \frac{10}{9}\right) \tag{23}$$

where $R_2(n)$ denotes the remainder when n is divided by 2. Finally, the first moment of the linear complexity (as shown in proposition 2.) together with the second moment as displayed in (23), allow the calculation of $\mathrm{Var}[\Lambda(s^n)]$, via (19). We summarize the result in the following proposition.

Proposition 3. $\mathrm{Var}[\Lambda(s^n)]$

The variance of the linear complexity of a sequence $s^n = s_0, s_1, \ldots, s_{n-1}$ of n independent and uniformly distributed binary random variables is given by

$$\mathrm{Var}[\Lambda(s^n)] = \frac{86}{81} - 2^{-n}\left(\frac{14-R_2(n)}{27}n + \frac{82-2R_2(n)}{81}\right) - 2^{-2n}\left(\frac{1}{9}n^2 + \frac{4}{27}n + \frac{4}{81}\right) \tag{24}$$

where $R_2(n)$ denotes the remainder when n is divided by 2. Moreover,

$$\lim_{n\to\infty} \mathrm{Var}[\Lambda(s^n)] = \frac{86}{81} \quad . \tag{25}$$

The variance is a measure of spread. If the variance is small then large deviations of the random variable under consideration from its mean are improbable. One might have expected that the spread of the linear complexity grows with increasing length n of the investigated sequence. Note that $\Lambda(s^n)$ may assume more and more values with increasing n. The interesting implication of proposition 3. is that the spread of the linear complexity $\Lambda(s^n)$ is virtually independent of the sequence length n. Regardless of how many sequence bits are processed, the fraction of sequences centered around the mean is virtually constant. We may make these intuitive statements more precise by invoking Chebyshev's inequality (Fell 68), which implies that, for any $k > 0$, the probability that the linear complexity of a random sequence s^n differs by an amount larger or equal than k from its mean is bounded from above by the variance of the linear complexity divided by k^2. Thus, for all n,

$$P\{\left|\Lambda(s^n) - E[\Lambda(s^n)]\right| \geq k\} \leq \frac{\mathrm{Var}[\Lambda(s^n)]}{k^2} \quad . \tag{26}$$

Suppose $k = 10$, then, for sufficiently large n, Chebychev's inequality provides a bound of $(86/81)10^{-2} = 0.0106$. Consequently, at least 99 % of all random sequences s^n have a linear complexity within the range $(n/2) \pm 10$. This is a surprisingly sharp characterization of random sequences by

means of their associated linear complexity. Moreover, Chebychev's inequality is known to yield fairly loose bounds in individual applications because of its universality, so we may expect an even closer scattering of the linear complexities around the mean.

A different approach which could help to characterize random sequences is to consider the growth process of the linear complexity as a special kind of random walk. In this interpretation, $\Lambda(s^n)$ gives the "position" of the "particle" at time n. We may define the n/2-line as the "origin" of the "particle", since at any time the expected location of the "particle" is about n/2 (compare proposition 2). Typically the "particle" would depart from the n/2-line to some position below the n/2-line, then jump above the n/2-line and walk back to the n/2 line. Fig. 3 illustrates such a typical section of the linear complexity profile of a binary sequence.

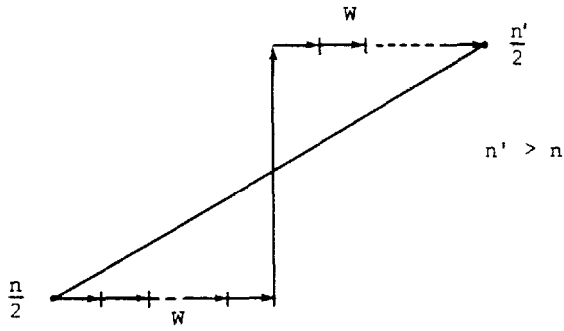

Fig. 3. A typical random walk segment of $\Lambda(s^n)$

Compare also the linear complexity profile of the swiss coin sequence (1) depicted in Fig. 1. The recursion (2) describing the growth of linear complexity forces $\Lambda(s^n)$ to retain its value, whenever that value is greater than n/2, until $\Lambda(s^{n'}) = n'/2$. From this point on, a change in linear complexity could occur at every step. In case of such a change, the jump of $\Lambda(s^n)$ is symmetrical with respect to the n/2-line, i.e. the "particle" $\Lambda(s^n)$ jumps from L to (n+1)-L. Without loss of essential generality, assume that $\Lambda(s^n) = n/2$. (Note that every nonzero sequence crosses at least once the n/2-line). Then the next jump will occur at time n+k, that is, after k time units, if

$$\delta_n = \delta_{n+1} = \ldots = \delta_{n+k-2} = 0 \; ; \; \delta_{n+k-1} = 1 \qquad (27)$$

causing the new linear complexity to be

$$\Lambda(s^{n+k}) = (n+k) - \Lambda(s^n) \quad . \qquad (28)$$

By the fact that the s_i are independent and fair coin tosses, the probability that the event (27) occurs is 2^{-k}. Let W be the random variable denoting the number of time units until the next length change occurs, given that at time n $\Lambda(s^n) = n/2$. The above observations then imply

$$E[W] = \sum_{k=1}^{\infty} k \, 2^{-k} = \sum_{k=0}^{\infty} 2^{-k} = 2 \quad . \qquad (29)$$

Thus, for the "particle" $\Lambda(s^n)$, the average return time to the origin (the n/2-line) will be $2E[W] = 4$; and the average jump height will be $E[\Delta L] = E[W]$, since $\Delta L = (n + W - (n/2)) - (n/2) = W$. The results obtained from the random walk interpretation of the linear complexity profile are summarized in the following proposition, where we have also generalized to an arbitrary starting point $\Lambda(s^n) = L$ to cover all possible sequences.

Proposition 4. Random walk setup

If $\tilde{s} = s_0, s_1, \ldots$ denotes a sequence of independent and uniformly distributed binary random variables and if $\Lambda(s^n) = L$, then the average number of sequence bits that have to be processed until the next length change occurs is given by

$$E[W \mid \Lambda(s^n) = L] = \begin{cases} 2 & \text{if } L \leq \frac{n}{2} \\ 2+2L-n & \text{if } L > \frac{n}{2} \end{cases} \quad . \qquad (30)$$

Moreover, the average length change is

$$E[\Delta L \mid \Lambda(s^n) = L] = \begin{cases} 2 & \text{if } L \geq \frac{n}{2} \\ n-2L+2 & \text{if } L < \frac{n}{2} \end{cases} \quad . \qquad (31)$$

The import of proposition 4. is that it provides information about the details of the linear complexity profile of random sequences.

Proposition 4. tells us that the linear complexity profile of a random sequence will look like an irregular staircase with an average step length of 4 time units and an average step height of 2 linear complexity units. A good illustration of this "typical" growth process is given by the linear complexity profile of the swiss coin sequence depicted in Fig. 1.

The various characterizations of binary random sequences by means of the associated linear complexity (as described in proposition 1. - 4.) might now suggest that we have only to put a "channel" of sufficient size around the n/2-line to separate the random looking sequences from the nonrandom looking sequences. But obviously enough, the probability that a random sequence $\Lambda(s^n)$ will leave this fictitious channel at least once goes to 1 as n goes to infinity. It is not even true that the sequences whose linear complexity profile stays very close to the n/2 line will always exhibit good statistical properties. An interesting example is provided by the sequence $\tilde{y}$ whose terms are defined as

$$y_j = \begin{cases} 1 & \text{if } j = 2^n - 1 \qquad n=0,1,2... \\ 0 & \text{otherwise.} \end{cases} \tag{32}$$

The sequence $\tilde{y}$ is highly "nonrandom", yet it has a linear complexity profile following the n/2-line as closely as is possible at least for $n \leq 127$ (and we conjecture for all n) (see Fig. 4). This conjecture was recently proven to be true by Zong-duo Dai (Dai 85).

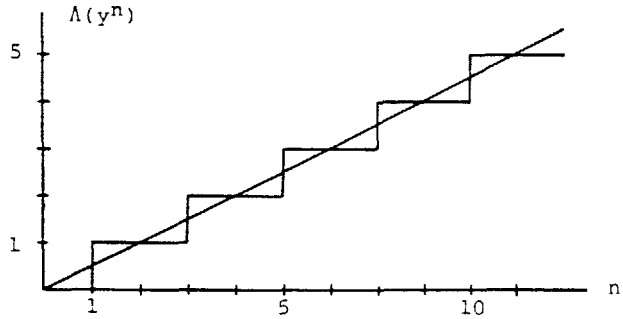

Fig. 4. The perfect staircase profile associated to the sequence (32)

This example suggests that too regular linear complexity profiles are incompatible with the randomness properties of the associated sequences. But note that the sequence $\tilde{y}$ as defined in (32) is not the only sequence with this perfect staircase profile. Whenever $\Lambda(s^n) > \frac{n}{2}$ then, independent of the choice for y_n, $\Lambda(y^{n+1})$ will be equal to $\Lambda(y^n)$. This indicates that there exist in fact many sequences which have associated the perfect staircase profil shown in Fig. 4. And undoubtedly, there will be some among them with good statistical properties. But remember that the perfect staircase profil would indeed pass randomness tests based on the expectation of linear complexity (proposition 2. and 3.), but it never would pass a randomness test based on the random walk setup (proposition 4.). Hence with the knowledge acquired so far on the linear complexity profile of random sequences, we would not accept as "random" a sequence with such a regular profile as that shown in Fig. 4.

From the practical standpoint in good stream cipher design, one important question remains to be answered. A deterministically generated key stream must necessarily be (ultimately) periodic. Thus, the question of what the linear complexity profile of a periodically repeated random bit string will look like is of considerable practical interest. Let $z^T = z_0, z_1, \ldots, z_{T-1}$ denote the first period of the semi-infinite sequence $\tilde{z}$, and assume z^T to be selected according to a fair coin tossing experiment. Then from the preceding analysis we may immediately deduce that $E[\Lambda(\tilde{z})]$ is at least T/2, since that result holds for the finite random sequence z^T. On the other hand z^T could be put into a pure cycling shift register of length T to produce $\tilde{z}$. Thus $\tilde{z}$ certainly satisfies the recursion $z_{T+j} = z_j$, which implies that $E[\Lambda(\tilde{z})]$ is at most T. But how likely is it that $\tilde{z}$ satisfies a linear recursion of order lower than T? And how would the linear complexity profile change from that point on where the first bits of z^T are repeated? Intuitively, one would expect the linear complexity to grow to close to the period length T, since the recursion which produces the second half of z^T from the first half is unlikely to have any similarities to the recursion that produces the first half of z^T from the second half (which is required by the periodic repetition). Now let $Z^*(D)$ denote the polynomial associated with the first period z^T of $\tilde{z}$. Then

$$Z(D) = \frac{Z^*(D)}{1+D^T} \tag{33}$$

$Z^*(D)$ may be interpreted as the polynomial associated with the initial state of a circulating shift register. The question of the expected linear complexity of $\tilde{z}$ now corresponds to asking for the expected degree m of the denominator polynomial in (33) after reduction by $\gcd(Z^*(D), 1+D^T)$. To every choice of $Z^*(D)$, there is a unique partial fraction expansion

$$Z(D) = \sum_{i=1}^{n} \sum_{k=1}^{m_i} \frac{P_{ik}(D)}{[C_i(D)]^k} \tag{34}$$

where $C_i(D)$, i=1,..,n, are the irreducible factors of $1 + D^T$ and m_i, i = 1,...,n are their multiplicities, and where $\deg(P_{ik}(D)) < \deg(C_i(D))$. Suppose now that the binary coefficients of the numerator polynomials $P_{ik}(D)$ are chosen independently from a uniform distribution. This induces a uniform probability distribution over the set of possible initial periods z^T, (or equivalently, over the set of possible $Z^*(D)$), since there exists a unique correspondence between initial periods $Z^*(D)$ and the choice of numerator polynomials in the partial fraction expansion (34). But a uniform probability measure over all z^T implies that each digit z_j, j=0,...,T-1, is an independent and uniformly distributed binary random variable. We conclude that the expected linear complexity of z may equivalently be computed as the expected degree of the minimal polynomial of $\tilde{z}$ given that all coefficients of the numerator polynomials $P_{ik}(D)$ are chosen independently from a uniform distribution. Unfortunately, there appears to be no simple solution to this problem since the irreducible factors $C_i(D)$ of $1+D^T$, as well as their multiplicities strongly depend on the value of T. We will demonstrate the solution for 2 extreme cases thereby obtaining results of some significance for the general case. Suppose first that T is equal to 2^n-1 with n a prime. Then the partial fraction expansion (34) takes on the special form

$$Z(D) = \frac{Z^*(D)}{1+D^{2^n-1}} = \frac{A}{1+D} + \sum_{i=1}^{M} \frac{P_i(D)}{C_i(D)} \tag{35}$$

where each $C_i(D)$ has prime degree n, and thus the number of such factors is $M = (2^n-2)/n$. When we randomly select A and the coefficients of $P_i(D)$, i=1,...,M, then the probability that A and $P_i(D)$ are zero is 2^{-1} and 2^{-n}, respectively. Therefore

$$P_k = P(\Lambda(\tilde{z}) = 2^n - 1 - kn) = P(\Lambda(\tilde{z}) = 2^n - 2 - kn)$$
$$= \frac{1}{2}\binom{M}{k} (1-2^{-n})^{M-k} (2^{-n})^k .$$

We obtain for large prime n and small k

$$P_k \approx \frac{1}{2k!n^k} e^{-\frac{1}{n}} \tag{37}$$

By considering the two choices of 2^n-1 and 2^n-2 for the linear complexity we may provide a rough lowerbound on the expected linear complexity of $\tilde{z}$,

$$E[\Lambda(\tilde{z})] \geq (2^n-1)P_o + (2^n-2)P_o$$
$$\geqslant \approx e^{-\frac{1}{n}} (2^n - \frac{3}{2}) \tag{38}$$

The significance of the bound (38) lies in the fact that, as n increases, it approaches the period T, thereby showing that the linear complexity of z can be expected to be very close to the period length for all prime n. A much better estimate of the actual $E[\Lambda(\tilde{z})]$ may be obtained when more than just the two largest choices for $\Lambda(\tilde{z})$, with their corresponding probabilities P_k as computed in (37) are taken into account. When T is chosen odd, then the minimal polynomial of $\tilde{z}$ does not contain any repeated factors (which is equivalent to saying that the minimal polynomial of $\tilde{z}$ has only simple roots). The other extreme may be found when the period length T is chosen to be a power of 2, i.e. $T = 2^n$. Then there exists only one root, namely 1, which occurs with multiplicity 2^n, and

$$Z(D) = \frac{Z^*(D)}{1+D^{2^n}} = \frac{Z^*(D)}{(1+D)^{2^n}} . \tag{39}$$

Then the partial fraction expansion (34) takes on the special form

$$Z(D) = \sum_{i=1}^{2^n} \frac{A_i}{(1+D)^i} . \tag{40}$$

When all the binary coefficients A_i are drawn independently from a uniform distribution, then half the sequences $\tilde{z}$ will have linear complexity 2^n, one forth of the $\tilde{z}$ will have linear complexity 2^n-1, one eighth will have $\Lambda(\tilde{z}) = 2^n-2$, and so on. Thus the probability distribution induced on $\Lambda(\tilde{z})$ is given by

$$P(\Lambda(\tilde{z}) = L) = 2^{L-2^n-1} \qquad L = 1,\ldots,2^n \tag{41}$$

With the help of this probability distribution, it is now easy to compute the expected linear complexity

$$E[\Lambda(\tilde{z})] = \sum_{L=1}^{2^n} L \cdot 2^{L-2^n-1} = 2^{-2^n-1} \sum_{L=1}^{2^n} L2^L \quad . \tag{42}$$

Invoking the integration/differentiation technique for sums (as demonstrated in the derivation of (14))results in

$$E[\Lambda(\tilde{z})] = 2^n - 1 + 2^{-2^n} \quad .$$

This result is summarized in the following proposition.

Proposition 5. Periodic repetition of random sequence

If the semi-infinite sequence $\tilde{z}$ is generated by periodically repeating a sequence $z^T = z_0,\ldots,z_{T-1}$ of T independent and uniformly distributed binary random variables, i.e. $\tilde{z} = z^T, z^T, \ldots$, and if $T = 2^n$, then the expected linear complexity of $\tilde{z}$ is

$$E[\Lambda(\tilde{z})] = 2^n - 1 + 2^{-2^n} \quad . \tag{43}$$

The two investigated cases of periodically repeating a finite sequence of random bits are extreme in the sense that, for a period $T = 2^n-1$, the minimal polynomial of $\tilde{z}$ is sure to contain only simple roots whose number then equals the linear complexity of $\tilde{z}$, and, for a period $T = 2^n$, the minimal polynomial of $\tilde{z}$ is sure to contain only one root whose multiplicity then equals the linear complexity of $\tilde{z}$. For both choices of the period we

were able to show that the expected linear complexity is almost equal to the period length.

Recapitulating, we may say that the linear complexity of a sequence provides a good measure of its unpredictability, expecially when the growth process of the linear complexity with respect to the number of considered sequence bits (which was termed the linear complexity profile) is taken into account. For true random sequences of length n, the expected linear complexity was shown to be about n/2. Moreover, the vast majority of these sequences were shown to have associated a linear complexity very close to n/2. The dynamic characterization of random sequences by means of linear complexity results in an average linear complexity increase of 2 after an average number of 4 considered sequence digits. When a random sequence of length $T = 2^n$ ($n \geq 0$) or $T = 2^n - 1$ (n prime) T is periodically repeated, then the expected linear complexity is close to the period length T and the associated linear complexity profile is not distinguishable from the linear complexity profile of a true random sequence up to T digits. Heuristic arguments suggest that the expected linear complexity will in general be close to the period length T and that in fact the associated linear complexity profile will not be distinguishable from the linear complexity profile of a true random sequence even up to 2T digits. (Compare also the swiss coin sequence example displayed in Fig. 4.1.). we conclude that a good random sequence generator should have linear complexity close to the period length, and also a linear complexity profile which follows closely, but "irregularly", the n/2-line (where n denotes the number of sequence digits) thereby exhibiting average step lengths and step heights of 4 and 2, respectively.

References:

Dai 85 Zong-duo Dai, "Proof of Rueppel's Linear Complexity Conjecture", submitted for publication in IEEE Trans. on Info. Th.

Fell 68 W. Feller, "An Introduction to Probability Theory and its Applications", Vol. 1, John Wiley, 1968.

Golo 67 S.W. Golomb, "Shift Register Sequences", Holden-Day, San Francisco, Calif., 1967.

Knut 81 D.E. Knuth, "The Art of Computer Programming, Vol. 2: Seminumberical Algorithms", Addison-Wesley, 1981.

Kolm 65 A.N. Kolmogorov, "Three Approaches to the Quantitative Definition of Information", Probl. Inform. Transmission, Vol. 1, 1965.

Lemp 76 A. Lempel, J. Ziv, "On the Complexity of Finite Sequences", IEEE Trans. on Info. Theory, IT-22, Jan. 1976.

Mart 66 P. Martin-Loef, "The Definition of Random Sequences", Information and Control, Vol. 9, 602-619, 1966.

Mass 69 J.L. Massey, "Shift-Register Synthesis and BCH Decoding", IEEE Trans. on Info. Theory, Vol. IT-15, Jan. 1969.

Solo 64 R.J.Solomonov, "A Formal Theory of Inductive Inference", Part I, Inform. Control 7, 1964.