J. Cryptology (1990) 2: 105–112

Journal of Cryptology

A Combinatorial Approach to Probabilistic Results on the Linear-Complexity Profile of Random Sequences[1]

Harald Niederreiter

Institute for Information Processing, Austrian Academy of Sciences, Sonnenfelsgasse 19, A-1010 Vienna, Austria

Abstract. The linear-complexity profile measures the extent to which the initial segments of a keystream sequence can be simulated by linear feedback shift-register sequences. To provide a benchmark for the assessment of keystream sequences, a probabilistic theory of the linear-complexity profile of random sequences is needed. For sequences of elements of a finite field we show probabilistic results that can be derived by a combinatorial method.

Key words. Stream cipher, Keystream sequence, Linear-complexity profile, Confidence interval.

1. Introduction

The linear-complexity profile was introduced by Rueppel [5], [6, Chapter 4] as a tool for assessing randomness properties of keystream sequences in the context of stream ciphers. Rueppel considered only binary keystream sequences, which is the case of greatest practical interest, but the linear-complexity profile can be defined for sequences of elements of any field (see [3]). To provide a benchmark for the assessment of keystream sequences, we have to investigate the linear-complexity profile of random sequences in a suitable probabilistic model. We use the probabilistic model set up by the author [4] which works for sequences of elements of any finite field. With this probabilistic model, various results on the linear-complexity profile of random sequences were established in [4]. The proofs of these results required rather heavy mathematical machinery from probability theory, topology, and the theory of dynamical systems. In this paper we study the connection between these results and the work of Rueppel [5], [6, Chapter 4] which is based on combinatorial arguments. In particular, we show that some interesting probabilistic results can also be obtained by the more elementary combinatorial method. In Section 2 we deal with infinite sequences, whereas in Section 3 we establish confidence intervals for finite strings when testing the deviation of the local linear complexity from the perfect linear complexity.

[1] Date received: March 16, 1989. Date revised: December 27, 1989.

We view a keystream sequence as a sequence of elements of a finite field. We denote by F_q the finite field with q elements, where q is an arbitrary prime power. The most important case of binary keystream sequences corresponds to $q = 2$. The following definitions are basic. A sequence $s_1, s_2, \ldots$ of elements of F_q is called a kth-*order* (*linear feedback*) *shift-register sequence* if there exist constant coefficients $a_{k-1}, \ldots, a_0 \in F_q$ such that

$$s_{i+k} = a_{k-1}s_{i+k-1} + \cdots + a_0 s_i \qquad \text{for} \quad i = 1, 2, \ldots.$$

The zero sequence $0, 0, \ldots$ is viewed as a shift-register sequence of order 0. Now let S be an arbitrary sequence $s_1, s_2, \ldots$ of elements of F_q and let n be a positive integer. Then the (*local*) *linear complexity* $L_n(S)$ is defined as the least k such that $s_1, s_2, \ldots, s_n$ form the first n terms of a kth-order shift-register sequence. The sequence $L_1(S)$, $L_2(S)$, ... of integers is called the *linear-complexity profile* of S. Thus the linear-complexity profile measures the extent to which the initial segments of a keystream sequence can be simulated by shift-register sequences. We clearly have $0 \leq L_n(S) \leq n$ and $L_n(S) \leq L_{n+1}(S)$ for all n and S.

A suitable probabilistic model was obtained in [4] by identifying sequences of elements of F_q with their generating functions, then furnishing the set of all generating functions with the structure of a compact abelian group and using the unique Haar measure on this compact abelian group as the probability measure. If we transfer this measure from the set of all generating functions to the set F_q^∞ of all sequences of elements of F_q, then we get a probability measure h on F_q^∞ which can be described as follows. For fixed $b_1, \ldots, b_m \in F_q$ define the cylinder set

$$C(b_1, \ldots, b_m) = \{S = (s_1, s_2, \ldots) \in F_q^\infty : s_i = b_i \text{ for } i = 1, 2, \ldots, m\}.$$

Then we have

$$h(C(b_1, \ldots, b_m)) = q^{-m} \tag{1}$$

for all positive integers m and all $b_1, \ldots, b_m \in F_q$. This follows from a comparison with formula (4) in [4]. The probability measure h is now obtained by the usual process of extension to the σ-algebra generated by all cylinder sets and subsequent completion (see Section 4 of [2]). Equivalently, h may be constructed by first considering the uniform probability measure μ on F_q which assigns the measure $1/q$ to each element of F_q and then letting h be the complete product measure on F_q^∞ induced by μ.

For a property P of sequences $S \in F_q^\infty$ we write $\text{Prob}(P)$ for the h-measure of the set of all $S \in F_q^\infty$ which have the property P. Of particular interest are those properties P for which $\text{Prob}(P) = 1$ since these can be viewed as typical properties of a random sequence of elements of F_q. We say that a property P holds *with probability 1* if $\text{Prob}(P) = 1$.

2. Probabilistic Laws for Linear Complexity

The basis of Rueppel's method is an explicit formula for the number of n-bit strings with a prescribed value of the linear complexity L_n. Such a formula for general F_q was established by Gustavson [1] and it is stated below; see also [5] and Chapter 4 of [6] for a proof in the case $q = 2$.

Lemma 1. *Let $N_n(L)$ be the number of strings of elements of F_q of length n and linear complexity $L_n = L$. Then*

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

By using the probability measure h on F_q^∞ constructed in Section 1, and in particular formula (1), we obtain, for fixed n and L, the identity

$$\text{Prob}(L_n(S) = L) = q^{-n} N_n(L). \tag{2}$$

The calculations by Rueppel [5], [6, Chapter 4] of the expected value and the variance of L_n may be interpreted in our model as the calculations of the integrals

$$E(L_n) = \int L_n(S)\, dh \quad \text{and} \quad \text{Var}(L_n) = \int (L_n(S) - E(L_n))^2\, dh$$

with respect to the measure h. This was carried out by Rueppel in the case $q = 2$, and these calculations were recently extended to general F_q by Smeets [7]. We recall that these results show that the expected value of L_n is close to $n/2$ and its variance is asymptotically a constant.

We now prove further probabilistic results for the linear complexity which can be derived from (2) and Lemma 1.

Theorem 1. *Let f be a nonnegative function on the positive integers with $\sum_{n=1}^{\infty} q^{-f(n)} < \infty$. Then with probability 1 we have*

$$\left| L_n(S) - \frac{n}{2} \right| \leq \tfrac{1}{2} f(n) \qquad \textit{for all sufficiently large } n.$$

Proof. For a fixed positive integer n let

$$D_n = \{S \in F_q^\infty : L_n(S) > \tfrac{1}{2}(n + f(n))\}.$$

If $k(n)$ is the least integer $> (n + f(n))/2$, then, for $S \in D_n$, we have $L_n(S) \geq k(n) \geq (n+1)/2$, and so $2n - 2L_n(S) < 2L_n(S) - 1$. Therefore from (2) and Lemma 1 we get, under the assumption that $k(n) \leq n$,

$$\begin{aligned} h(D_n) &= q^{-n} \sum_{L=k(n)}^{n} N_n(L) = q^{-n} \sum_{L=k(n)}^{n} (q-1)q^{2n-2L} \\ &= (q-1)q^n \sum_{L=k(n)}^{n} q^{-2L} = (q-1)q^{n-2k(n)} \sum_{L=0}^{n-k(n)} q^{-2L} \\ &< (q-1)q^{n-2k(n)} \sum_{L=0}^{\infty} q^{-2L} = \frac{1}{q+1} q^{2+n-2k(n)}. \end{aligned}$$

Since $k(n) > (n + f(n))/2$, it follows that

$$h(D_n) < \frac{1}{q+1} q^{2-f(n)}.$$

If $k(n) > n$, then $h(D_n) = 0$, and so the bound above holds in all cases. From the

hypothesis $\sum_{n=1}^{\infty} q^{-f(n)} < \infty$ we then obtain $\sum_{n=1}^{\infty} h(D_n) < \infty$. The Borel–Cantelli lemma [2, p. 228] now shows that the set of all S for which $S \in D_n$ for infinitely many n has h-measure 0. In other words, with probability 1 we have $S \in D_n$ for at most finitely many n. From the definition of D_n it follows then that with probability 1 we have

$$L_n(S) \leq \tfrac{1}{2}(n + f(n)) \qquad \text{for all sufficiently large } n. \tag{3}$$

By a similar method we get an analogous lower bound. For a fixed n let

$$E_n = \{S \in F_q^{\infty} : L_n(S) < \tfrac{1}{2}(n - f(n))\}.$$

If $m(n)$ is the largest integer $< (n - f(n))/2$, then, for $S \in E_n$, we have $L_n(S) \leq m(n) < n/2$, and so $2n - 2L_n(S) > 2L_n(S) - 1$. Therefore from (2) and Lemma 1 we get, under the assumption that $m(n) \geq 1$,

$$\begin{aligned} h(E_n) &= q^{-n} \sum_{L=0}^{m(n)} N_n(L) = q^{-n} + q^{-n} \sum_{L=1}^{m(n)} (q-1) q^{2L-1} \\ &= q^{-n} + (q-1) q^{1-n} \sum_{L=0}^{m(n)-1} q^{2L} \\ &= q^{-n} + \frac{1}{q+1} q^{1-n} (q^{2m(n)} - 1) < q^{-n} + \frac{1}{q+1} q^{1-n+2m(n)}. \end{aligned}$$

Since $m(n) < (n - f(n))/2$, we obtain

$$h(E_n) < q^{-n} + \frac{1}{q+1} q^{1-f(n)}.$$

We have $h(E_n) = q^{-n}$ if $m(n) = 0$ and $h(E_n) = 0$ if $m(n) < 0$, and so the bound above holds in all cases. It follows that $\sum_{n=1}^{\infty} h(E_n) < \infty$, and by applying the Borel–Cantelli lemma as before we deduce that with probability 1 we have

$$L_n(S) \geq \tfrac{1}{2}(n - f(n)) \qquad \text{for all sufficiently large } n.$$

Together with (3) this shows the desired result. □

Theorem 1 yields the result of Theorem 8 of [4] by a less involved method, and in fact the hypothesis in Theorem 1 is weaker since in Theorem 8 of [4] it was necessary for technical reasons to make the additional assumption that f is non-decreasing. On the other hand, the elementary method in the proof of Theorem 1 cannot be used to prove deeper results such as Theorem 9 of [4]. The reason is that the events $D_1, D_2, \ldots$ in the proof of Theorem 1 are not independent (also $E_1, E_2, \ldots$ are not independent), and so the required Borel zero-one law [2, p. 228] cannot be applied. Therefore, it is also impossible to derive the law of the logarithm for linear complexity [4, Theorem 10] by the elementary method above. However, the following weaker form of this law can be established as a consequence of Theorem 1.

Corollary 1. *With probability* 1 *we have*

$$\overline{\lim_{n \to \infty}} \frac{|L_n(S) - (n/2)|}{\log n} \leq \frac{1}{2 \log q}.$$

Proof. For a positive integer m we apply Theorem 1 with the function $f(n) = (1 + m^{-1})(\log n)/(\log q)$. Then with probability 1

$$\frac{|L_n(S) - (n/2)|}{\log n} \leq \frac{1 + m^{-1}}{2 \log q} \qquad \text{for all sufficiently large } n.$$

This property holds simultaneously for all m with probability 1 since the countable intersection of sets of h-measure 1 has again h-measure 1. The desired conclusion follows. □

Corollary 2. *With probability* 1 *we have*

$$\lim_{n\to\infty} \frac{L_n(S)}{n} = \tfrac{1}{2}.$$

The result of Corollary 2, derived here by elementary means, was shown in Theorem 7 of [4] by using methods from the theory of dynamical systems. Another asymptotic result is obtained on the basis of the following lemma. For real t we use the following standard notation: $\lfloor t \rfloor$ is the greatest integer $\leq t$ and $\lceil t \rceil$ is the least integer $\geq t$.

Lemma 2. *For any integers k, m, and n with $1 \leq m \leq n/2 \leq k \leq n$ we have*

$$\text{Prob}(m \leq L_n(S) \leq k) = 1 - \frac{1}{q+1}(q^{2m-n-1} + q^{n-2k}).$$

Proof. Using (2) and Lemma 1 and the standard convention that empty sums have the value 0, we get

$$\begin{aligned}
\text{Prob}(m \leq L_n(S) \leq k) &= q^{-n} \sum_{L=m}^{k} N_n(L) \\
&= q^{-n} \sum_{L=m}^{\lfloor n/2 \rfloor} N_n(L) + q^{-n} \sum_{L=\lfloor n/2 \rfloor + 1}^{k} N_n(L) \\
&= (q-1)q^{-n-1} \sum_{L=m}^{\lfloor n/2 \rfloor} q^{2L} + (q-1)q^{n} \sum_{L=\lfloor n/2 \rfloor + 1}^{k} q^{-2L} \\
&= (q-1)q^{2m-n-1} \sum_{L=0}^{\lfloor n/2 \rfloor - m} q^{2L} + (q-1)q^{n-2\lfloor n/2 \rfloor - 2} \sum_{L=0}^{k-\lfloor n/2 \rfloor - 1} q^{-2L} \\
&= \frac{1}{q+1} q^{2m-n-1}(q^{2\lfloor n/2 \rfloor - 2m + 2} - 1) \\
&\quad + \frac{1}{q+1} q^{n-2k}(q^{2k-2\lfloor n/2 \rfloor} - 1) \\
&= 1 - \frac{1}{q+1}(q^{2m-n-1} + q^{n-2k}).
\end{aligned}$$

□

Theorem 2. *If f and g are nonnegative functions on the positive integers with $\lim_{n\to\infty} f(n) = \lim_{n\to\infty} g(n) = \infty$, then*

$$\lim_{n\to\infty} \text{Prob}\left(-f(n) \leq L_n(S) - \frac{n}{2} \leq g(n) \right) = 1.$$

Proof. We take $n \geq 2$ so large that $f(n) \geq \frac{1}{2}$ and $g(n) \geq \frac{1}{2}$. Put

$$m(n) = \max\left(1, \left\lceil \frac{n}{2} - f(n) \right\rceil\right), \qquad k(n) = \min\left(n, \left\lfloor \frac{n}{2} + g(n) \right\rfloor\right).$$

Then $1 \leq m(n) \leq n/2 \leq k(n) \leq n$, and so Lemma 2 yields

$$\begin{aligned} \text{Prob}\left(-f(n) \leq L_n(S) - \frac{n}{2} \leq g(n)\right) &= \text{Prob}\left(\left\lceil \frac{n}{2} - f(n) \right\rceil \leq L_n(S) \leq \left\lfloor \frac{n}{2} + g(n) \right\rfloor\right) \\ &\geq \text{Prob}(m(n) \leq L_n(S) \leq k(n)) \\ &= 1 - \frac{1}{q+1}(q^{2m(n)-n-1} + q^{n-2k(n)}). \end{aligned}$$

Now

$$\lim_{n\to\infty} (2m(n) - n) = -2 \lim_{n\to\infty} \min\left(\frac{n}{2} - 1, \frac{n}{2} - \left\lceil \frac{n}{2} - f(n) \right\rceil\right) = -\infty,$$

$$\lim_{n\to\infty} (n - 2k(n)) = -2 \lim_{n\to\infty} \min\left(\frac{n}{2}, \left\lfloor \frac{n}{2} + g(n) \right\rfloor - \frac{n}{2}\right) = -\infty,$$

and so

$$\lim_{n\to\infty} \left(1 - \frac{1}{q+1}(q^{2m(n)-n-1} + q^{n-2k(n)})\right) = 1. \qquad \square$$

The result of Theorem 2 is best possible in the sense that if one of the conditions $\lim_{n\to\infty} f(n) = \infty$ and $\lim_{n\to\infty} g(n) = \infty$ is not satisfied, then the conclusion of the theorem cannot hold.

Theorem 3. *If f and g are nonnegative functions on the positive integers such that either $\underline{\lim}_{n\to\infty} f(n) < \infty$ or $\underline{\lim}_{n\to\infty} g(n) < \infty$, then*

$$\underline{\lim_{n\to\infty}} \text{Prob}\left(-f(n) \leq L_n(S) - \frac{n}{2} \leq g(n)\right) < 1.$$

Proof. We consider only the case $\underline{\lim}_{n\to\infty} g(n) < \infty$, the other case is treated similarly. This assumption implies that there exists an $M \geq 1$ such that $g(n) \leq M$ for infinitely many n. In particular, there exist infinitely many $n \geq 2M$ with $g(n) \leq M$. For such n we put $k(n) = \lceil (n/2) + M \rceil$ and we note that $k(n) \leq n$. We then get

$$\begin{aligned} \text{Prob}\left(-f(n) \leq L_n(S) - \frac{n}{2} \leq g(n)\right) &\leq \text{Prob}(0 \leq L_n(S) \leq k(n)) \\ &= \text{Prob}(L_n(S) = 0) + \text{Prob}(1 \leq L_n(S) \leq k(n)) \\ &= q^{-n} + 1 - \frac{1}{q+1}(q^{1-n} + q^{n-2k(n)}) \end{aligned}$$

by Lemma 2. Now $k(n) < (n/2) + M + 1$, thus $n - 2k(n) > -2M - 2$, and so

$$\text{Prob}\left(-f(n) \leq L_n(S) - \frac{n}{2} \leq g(n)\right) < q^{-n} + 1 - \frac{1}{q+1}(q^{1-n} + q^{-2M-2})$$

for infinitely many n. This implies

$$\varliminf_{n\to\infty} \text{Prob}\left(-f(n) \le L_n(S) - \frac{n}{2} \le g(n)\right) \le 1 - \frac{1}{q+1}q^{-2M-2} < 1. \qquad \square$$

3. Confidence Intervals for Finite Strings

In practical statistical testing based on linear complexity we cannot work with infinite sequences, but we have to deal with strings of elements of F_q of finite length n. One important aspect of this is to obtain confidence intervals for the deviation of $L_n(S)$ from the perfect linear complexity $\lfloor (n+1)/2 \rfloor$. For this reason we consider the following problem: for a given positive integer n and a given ε with $0 < \varepsilon < 1$ determine the least integer $d = d_n(\varepsilon)$ such that

$$\text{Prob}\left(\left|L_n(S) - \left\lfloor \frac{n+1}{2} \right\rfloor\right| \le d\right) \ge 1 - \varepsilon.$$

Note that because of (2), the probability on the left-hand side is also equal to q^{-n} times the number of strings of n elements of F_q for which the linear complexity deviates from $\lfloor (n+1)/2 \rfloor$ by at most d.

Lemma 3. *For any positive integer n and any integer d with $0 \le d \le (n-1)/2$ we have*

$$\text{Prob}\left(\left|L_n(S) - \left\lfloor \frac{n+1}{2} \right\rfloor\right| \le d\right) = 1 - q^{-2d-1}.$$

Proof. The result is easily checked for $d = 0$ by (2) and Lemma 1. For $d \ge 1$ we apply Lemma 2 and get

$$\begin{aligned}\text{Prob}\left(\left|L_n(S) - \left\lfloor \frac{n+1}{2} \right\rfloor\right| \le d\right) &= \text{Prob}\left(\left\lfloor \frac{n+1}{2} \right\rfloor - d \le L_n(S) \le \left\lfloor \frac{n+1}{2} \right\rfloor + d\right) \\ &= 1 - \frac{1}{q+1}q^{-2d-1}(q^{2\lfloor (n+1)/2 \rfloor - n} + q^{n+1-2\lfloor (n+1)/2 \rfloor}) \\ &= 1 - q^{-2d-1}. \qquad \square\end{aligned}$$

Theorem 4. *If $d_n(\varepsilon)$ is defined as above, then, for any positive integer n, we have*

$$d_n(\varepsilon) = \begin{cases} \left\lfloor \dfrac{n+1}{2} \right\rfloor & \text{for } 0 < \varepsilon < q^{1-2\lfloor (n+1)/2 \rfloor}, \\ \lceil -\frac{1}{2}(1 + \log_q \varepsilon) \rceil & \text{for } q^{1-2\lfloor (n+1)/2 \rfloor} \le \varepsilon < 1, \end{cases}$$

where $\log_q$ denotes the logarithm to the base q.

Proof. First let $0 < \varepsilon < q^{1-2\lfloor (n+1)/2 \rfloor}$. Then, for $0 \le d \le \lfloor (n-1)/2 \rfloor$, we have, by Lemma 3,

$$\begin{aligned}\text{Prob}\left(\left|L_n(S) - \left\lfloor \frac{n+1}{2} \right\rfloor\right| \le d\right) &= 1 - q^{-2d-1} \\ &\le 1 - q^{1-2\lfloor (n+1)/2 \rfloor} < 1 - \varepsilon,\end{aligned}$$

whereas, for $d = \lfloor (n+1)/2 \rfloor$, we trivially have

$$\text{Prob}\left(\left|L_n(S) - \left\lfloor \frac{n+1}{2} \right\rfloor\right| \leq d\right) = 1.$$

Therefore in this case $d_n(\varepsilon) = \lfloor (n+1)/2 \rfloor$. Now let $q^{1-2\lfloor (n+1)/2 \rfloor} \leq \varepsilon < 1$. Then, for $d = \lfloor (n-1)/2 \rfloor$, we get, by Lemma 3,

$$\text{Prob}\left(\left|L_n(S) - \left\lfloor \frac{n+1}{2} \right\rfloor\right| \leq d\right) = 1 - q^{1-2\lfloor (n+1)/2 \rfloor} \geq 1 - \varepsilon,$$

and so we have $0 \leq d_n(\varepsilon) \leq \lfloor (n-1)/2 \rfloor$. Using again Lemma 3, it follows that $d_n(\varepsilon)$ is the least integer d such that

$$1 - q^{-2d-1} \geq 1 - \varepsilon.$$

Since this inequality is equivalent to $d \geq -(1 + \log_q \varepsilon)/2$, the desired formula follows. □

References

[1] F. G. Gustavson, Analysis of the Berlekamp–Massey linear feedback shift-register synthesis algorithm, *IBM J. Res. Develop.*, vol. 20, 1976, pp. 204–212.

[2] M. Loève, *Probability Theory*, 3rd edn., Van Nostrand, New York, 1963.

[3] H. Niederreiter, Sequences with almost perfect linear complexity profile, *Advances in Cryptology—EUROCRYPT '87*, Lecture Notes in Computer Science, Vol. 304, Springer-Verlag, Berlin, 1988, pp. 37–51.

[4] H. Niederreiter, The probabilistic theory of linear complexity, *Advances in Cryptology—EUROCRYPT '88*, Lecture Notes in Computer Science, Vol. 330, Springer-Verlag, Berlin, 1988, pp. 191–209.

[5] R. A. Rueppel, Linear complexity and random sequences, *Advances in Cryptology—EUROCRYPT '85*, Lecture Notes in Computer Science, Vol. 219, Springer-Verlag, Berlin, 1986, pp. 167–188.

[6] R. A. Rueppel, *Analysis and Design of Stream Ciphers*, Springer-Verlag, Berlin, 1986.

[7] B. Smeets, The linear complexity profile and experimental results on a randomness test of sequences over the field F_q, Preprint, University of Lund, September 1988.