Technische Universität Graz   Institut für Mathematik

Sequences with subword complexity $2n$

Günter Rote

Report 204   April 1992

Technische Universität Graz, Steyrergasse 30, A-8010 Graz, Austria

Version 2a

This paper appeared in *Journal of Number Theory* **46** (1993), 196–213. Some typographical errors have been corrected.

Figure 2 of this report is not complete in this version because it should contain some hand-drawn additional arcs. The reader, however, is able work out the details for herself.

Title page processed by TEX on February 10, 1995

Sequences with subword complexity $2n$

Günter Rote

Abstract

We construct and discuss infinite 0-1-sequences which contain $2n$ different subwords of length n, for every n.

1 Introduction

For an infinite word $w = w_1w_2w_3\ldots$ over some finite alphabet, we denote by $L = \{ w_iw_{i+1}\ldots w_j : 1 \le i \le j \}$ the set of its finite subwords (factors). Let $L_n = \{ x \in L : |x| = n \}$ denote the set of subwords of length n. Then the function $P_w(n) := \#L_n$ which gives the cardinalities of the sets L_n is called the (*subword*) *complexity* of the sequence w. Usually, w is fixed by the context and we will write $P(n)$ for $P_w(n)$.

This paper considers sequences with complexity $P(n) = 2n$. We give more general and more specific methods for constructing them.

Infinite words (sequences) over some finite alphabet have been the study of research since the pioneering work of Thue [1906]. Areas of application include such diverse areas as iteration theory, ergodic theory and dynamical systems, formal languages, probability theory, and number theory (see Allouche [1987]).

If $P(n) \le n$ for some n, then the word is ultimately periodic, and $P(n)$ is in fact bounded. The lowest possible complexity for an interesting infinite word is thus $P(n) = n+1$. Sequences with complexity $P(n) \le n+1$ are called Sturmian sequences, and they are well understood (cf. Morse and Hedlund [1940] or Coven and Hedlund [1973]). There are several different ways by which one can construct any Sturmian sequence.

Recently, Arnoux and Rauzy [1991] went one step further and investigated sequences with complexity $2n+1$. They showed that, under certain conditions, such sequences can be represented in a geometric way like Sturmian sequences: as the orbit of a point under a simple one-to-one mapping of the unit circle into itself.

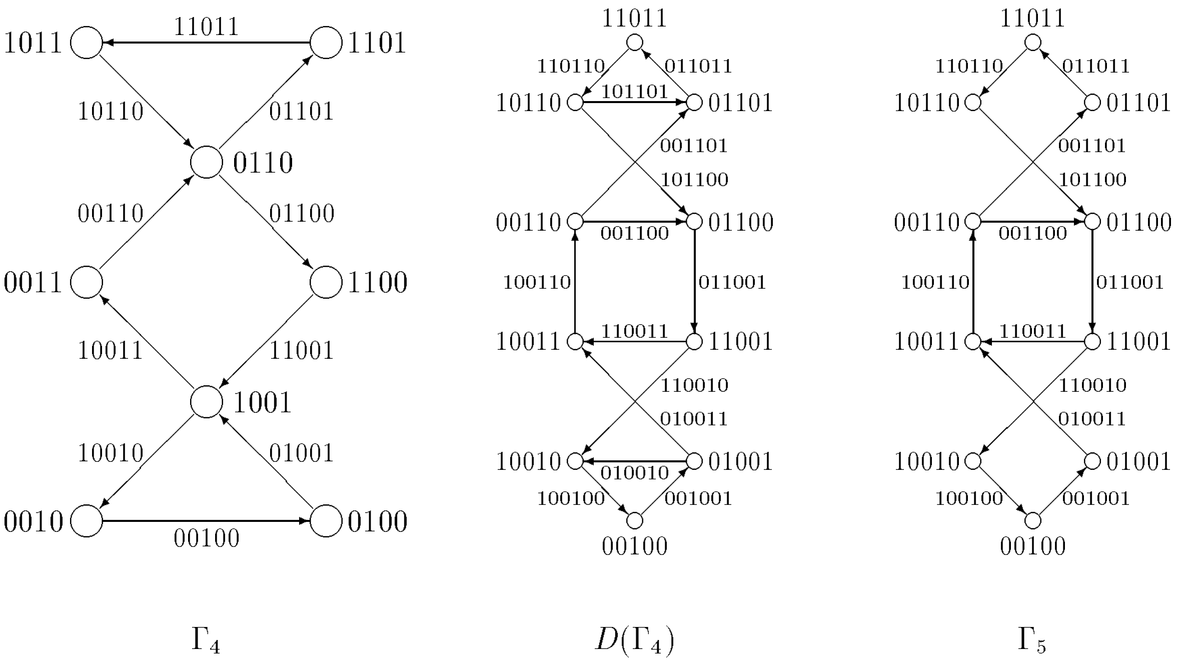
In this paper we consider an intermediate case: sequences with complexity $2n$. We show how one can in principle construct all such sequences (section 2). The tools that we use are purely combinatorial, similar to the methods of Arnoux and Rauzy [1991].

Since $P(1) = 2$, the ground alphabet has two symbols, and we will assume that our alphabet is $\{0, 1\}$.

In section 2, we give the main graph-theoretic construction. In section 3 we give three concrete examples of sequences with different properties that arise from this construction, and we give alternate methods for constructing special classes of sequences of complexity $2n$. The special class of sequences which are closed under complementation has a strong connection with the class of Sturmian sequences. This relation is explored in section 4. The last section mentions some open questions.

2 A construction using graphs

The main tool in this section is the set of directed graphs Γ_n which are related to the sets of n-letter subwords L_n. The vertices of Γ_n are the elements of L_n, and the arcs correspond to the elements of L_{n+1}: for each word $axb \in L_{n+1}$, where $a, b \in \{0,1\}$ and $x \in \{0,1\}^{n-1}$, the graph has an arc from ax to xb. For example, the graph Γ_4 of the word 0110011001001100110110 01... is shown in figure 1.



Γ_4 $D(\Gamma_4)$ Γ_5

Figure 1: The graph Γ_4, its line graph, and the graph Γ_5,
for the word 01100110010011001101100110010011...

The *line graph* $D(\Gamma_n)$ of a word graph Γ_n is defined as usual in graph theory: $D(\Gamma_n)$ has a vertex for each arc of Γ_n, and two vertices u and v of $D(\Gamma_n)$ are joined by an arc from u to v if the endpoint of the arc u in Γ_n coincides with the initial point of v. Figure 1 shows an example. Naturally, the vertices of $D(\Gamma_n)$ are labeled by the words of L_{n+1}, and its arcs are labeled by words of length $n+2$: for an arc (u,v), the final $n+1$ letters of u must coincide with the first $n+1$ letters of v, i. e., we can write $u = axb$ and $v = cyd$ with $xb = cy$. Then we label the arc (u,v) by $axbd = acyd$.

In our construction of an infinite word w with complexity $P(n) = 2n$ we will concentrate on successively constructing the graphs Γ_n, for $n = 1, 2, \ldots$. Let us therefore discuss the required properties of those graphs, and the relation between Γ_n and Γ_{n+1}.

The arcs of the line graph $D(\Gamma_n)$ are all possible words of length $n+2$ whose $(n+1)$-letter subwords belong to L_{n+1}. Clearly, L_{n+2} must be a subset of those words. In other words, the graph Γ_{n+1} can be formed from $D(\Gamma_n)$ by removing some edges (or possibly no edges).

Since we are interested in words with complexity $2n$, we want Γ_n to have $2n$ vertices and $2n+2$ arcs. In particular, Γ_1 has 2 vertices and 4 arcs, i. e., we have no choice for Γ_1; it must be the "complete" graph with vertex set $\{0,1\}$ and arc set $\{00, 01, 10, 11\}$.

Thus we can state the following algorithm:

```
Algorithm for constructing the sequence of graphs $\Gamma_n$, for $n = 1, 2, \ldots$
        Let $\Gamma_1$ be the graph with vertex set $\{0, 1\}$ and arc set $\{00, 01, 10, 11\}$.
        for $n := 1$ to infinity do
           Construct the line graph $D(\Gamma_n)$ of $\Gamma_n$.
           if $D(\Gamma_n)$ contains $2n + 4$ arcs
              then let $\Gamma_{n+1} := D(\Gamma_n)$.
           if $D(\Gamma_n)$ contains more than $2n + 4$ arcs
$(*)$         then remove the correct number of arcs from $D(\Gamma_n)$ to obtain $\Gamma_{n+1}$.
           if $D(\Gamma_n)$ contains less than $2n + 4$ arcs
$(**)$        then STOP.
        end for.
```

In the step marked $(*)$ we actually have a *choice* of the arcs which we remove. We will therefore derive further necessary properties of Γ_{n+1} and formulate a set of *rules* that will guide us in step $(*)$.

Let us consider the number of arcs of $D(\Gamma_n)$. This number decides about the course that the algorithm takes, and in particular it decides whether the algorithm may get stuck in step $(**)$. $D(\Gamma_n)$ has an arc for each pair consisting of an arc of Γ_n that leads to a vertex v and an arc leaving this vertex v. The numbers of these arcs of Γ_n are called the *indegree* $\delta^-(v)$ and the *outdegree* $\delta^+(v)$, respectively. Thus $D(\Gamma_n)$ has $\sum_{v \in L_n} \delta^-(v) \cdot \delta^+(v)$ arcs. Since our alphabet has only two symbols, there are only two possible arcs that can leave a vertex v, namely $v0$ and $v1$. Therefore, we always have $\delta^+(v) \le 2$, and similarly, $\delta^-(v) \le 2$.

Since the n-letter subsequences of w starting at positions $1, 2, 3, \ldots$ trace out an infinite path through Γ_n which visits every vertex and every arc, the graph must have a vertex from which every other vertex can be reached. We will even insist on the stronger requirement that all graphs Γ_n are strongly connected, i. e., every vertex can be reached from *every* other vertex.

Rule 1: Keep Γ_{n+1} strongly connected.

The following theorem states what this rule amounts to.

Theorem 1 *Let $w = w_1w_2w_3\ldots$ be an infinite word. Then the following are equivalent:*

(i) *All graphs Γ_n are strongly connected.*

(ii) *Every subword that occurs in w occurs at least twice.*

(iii) *Every subword that occurs in w occurs infinitely often.*

Proof. (i) implies (ii): Let $x = w_1w_2\ldots w_n$ be an initial subword of w. Since Γ_n has an arc entering x there must be a second occurrence of x in w. Thus (ii) holds for all initial subwords. Since any subword of w is contained in an initial subword, (ii) follows in general.

(ii) implies (iii): Let x be a subword of w. By (ii), there is a subword x' of w in which x occurs twice. By applying (ii) again, there is a subword x'' of w in which x'

occurs twice. This word x'' must contain at least three occurrences of x. Similarly, a word x''' which contains x'' twice contains at least four occurrences of x. Continuing inductively, we conclude that x is contained infinitely often in w.

(iii) implies (i): Consider any two subwords x and y of length n. After any occurrence of x there must still follow another occurrence of y in w, and hence there is a path from x to y in Γ_n. □

Rule 1 implies in particular that we will always have $\delta^-(v) \geq 1$ and $\delta^+(v) \geq 1$, for all v. Γ_n has $P(n) = 2n =: t$ vertices and $P(n+1) = t+2$ arcs. The possible topologies of strongly connected directed graphs with t vertices and $t+2$ arcs and with $\delta^-(v), \delta^+(v) \leq 2$ are listed in the middle column of figure 2. In these pictures, each curved arrow represents a path of one or more arcs. The cases are classified according to the set of degree pairs $(\delta^-(v), \delta^+(v))$ of Γ_n, which is listed in the leftmost column of the figure. The possible degree pairs are subject to the condition that the sum of the indegrees $\delta^-(v)$ and the sum of the outdegrees $\delta^+(v)$ of the t vertices v equals the number of arcs, which is $t+2$. This leaves only the three possible degree sequences given in the table.

The right column of figure 2 shows the topologies of the resulting line graphs. In these pictures, the curved arrows denote paths consisting of *zero* or more arcs, whereas the short straight arrows denote single arcs. The table states also in each case the number of arcs that have to be removed from $D(\Gamma_k)$ in order to get a graph Γ_{k+1} with $t+4$ arcs. In the bottom-most case, this number is zero, and since we do not have to make any choice in this case, we need not care about the possible topologies.

Since the table lists all possible degree sequences, we see that case $(**)$ cannot arise, and the algorithm can never get stuck.

In the line graphs, some arcs are marked by little lines crossing through them. One way of satisfying rule 1 is the following more specific rule:

> **Rule 1′:** Select the arcs that are to be removed from $D(\Gamma_n)$ only from among the marked arcs in figure 2

As can be checked from the pictures, this ensures that the graph L_{n+1} is strongly connected.

Lemma 1 *If a sequence of graphs Γ_n with vertex sets L_n is constructed according to rule 1, then every word of L_n is contained in some word of L_{n+1}, and hence, in some word of $L_{n'}$ for every $n' \geq n$.*

Proof. Every vertex x of Γ_n has at least one arc incident to it. This arc is a vertex of Γ_{n+1} and therefore a word of L_{n+1} containing x. □

So far, we have succeeded to construct sequences of word sets L_n with cardinality $\#L_n = 2n$ and the property that the language $L = \bigcup_{n=1}^{\infty} L_n$ is closed under taking subwords. However, we must still make sure that L is the set of subwords of an infinite sequence. Therefore, the sequence $L_1, L_2, \ldots$ that we will construct must have the following property:

> **Extension property:** There is a subsequence $L_{n_1}, L_{n_2}, L_{n_3}, \ldots$ ($n_1 < n_2 < n_3 < \cdots$) and a sequence of words $x_1, x_2, x_3, \ldots$ with $x_i \in L_{n_i}$ such that x_{i+1} starts with x_i and contains all words of L_{n_i}.

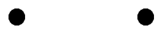
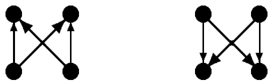
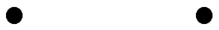
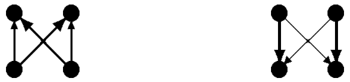
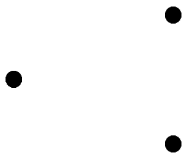
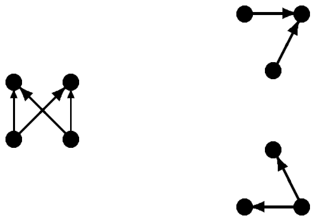
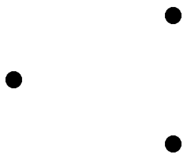
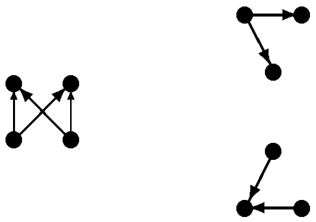
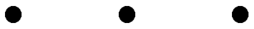
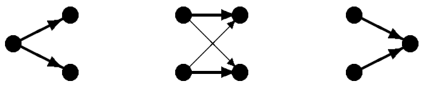
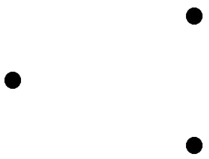
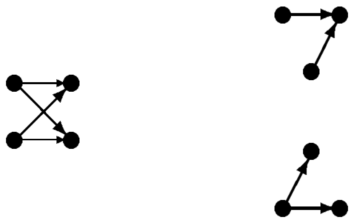
degree sequence	the graph Γ_n	the line graph $D(\Gamma_n)$
1 $(2,2)$ $(2,2)$ $(1,1)$ $(1,1)$ $\vdots$ 2 arcs are to be deleted from the line graph.	1A 	
	1B 	
2 $(2,2)$ $(2,1)$ $(1,2)$ $(1,1)$ $(1,1)$ $\vdots$ One arc is to be deleted from the line graph.	2A 	
	2B 	
	2C 	
	2D 	
3 $(2,1)$ $(2,1)$ $(1,2)$ $(1,2)$ $(1,1)$ $\vdots$	No arcs are to be deleted from the line graph in this case. Γ_{n+1} is unique and contains every arc of $D(\Gamma_n)$.	

Figure 2: The possible degree sequences and topologies of Γ_n, and the corresponding topologies of the line graphs $D(\Gamma_n)$ from which Γ_{n+1} is obtained.

The following elementary lemma states that this property is necessary and sufficient for our purposes.

Lemma 2 *A set L of finite words is the set of finite subwords of an infinite word $w_1w_2w_3\dots$ if and only if*

(i) *every word of length n in L is contained in some word of length $n+1$ in L; and*

(ii) *the sequence $L_1, L_2, \dots$, where $L_n = \{\, x \in L : |x| = n \,\}$ is the set of subwords of length n, fulfills the extension property.*

Proof. Note first that (i) is the property stated in lemma 1. The two properties are clearly necessary. They are also sufficient because the extension property ensures that the infinite word $w = w_1w_2w_3\dots$ which is the limit of the sequence $x_1, x_2, x_3, \dots$ contains every word of L_{n_1}, L_{n_2}, etc. Together with property (i) this implies that w contains every word of L. □

There are more marked arcs than we must remove; we will use this freedom of choice for achieving the extension property.

Lemma 3 *Let x and y be two different vertices in L_n, and let d be the distance from x to y. Then we can choose the correct number of arcs to be removed from the line graph $D(L_n)$ such that the resulting graph L_{n+1} contains two vertices xa and cy, (which correspond in L_n to an arc leaving x and an arc entering y) such that the distance from xa to cy in L_{n+1} is $d-1$.*

Proof. Let xa and cy be the first and last arc on the shortest path from x to y in L_n. Then the line graph contains a (shortest) path from xa to cy of length $d-1$.

Now we simply have to check for all possibilities in the table that for any specified shortest path in the line graph, we can always select the correct number of marked arcs to be removed and still leave this shortest path intact. If we have to remove p arcs from among q marked arcs, it suffices to show that no shortest path can contain more than $q-p$ marked arcs.

For example, in case 1A, where two arcs are to be removed, we have to show that a path using three of the four marked arcs cannot be a shortest path. Such a path would have to contain the two marked arcs in the left half of the picture or the two marked arcs in the right half of the picture. In any case, there is an arc from the starting point of the first marked arc to the endpoint of the second marked arc, and this would shortcut the path. □

We will now use this lemma to inductively construct $L_{n_{i+1}}$ and x_{i+1} from L_{n_i} and x_i so that the extension property is fulfilled. We call this transition from i to $i+1$ a *stage* of the algorithm.

Suppose we have already constructed a graph L_k with $k \geq n_i$ and a word $u \in L_k$ which starts with x_i and contains a certain number of other words of L_{n_i} as subwords. (At the start of a stage we have $k = n_i$ and $u = x_i$.) If u already contains all words of L_{n_i} as subwords, we can take $n_{i+1} = k$ and $x_{i+1} = u$ and we have completed the stage and can start the next one. Otherwise, select some word of L_{n_i} which is not contained in u and find a word v in L_k which contains it. (Such a word v exists by lemma 1.) Declare v to be the current *target word* and call u the current *start word.*

Let d be the distance from u to v. By the previous lemma, L_{k+1} contains a word ua whose distance to some word cv containing v is $d-1$ in $D(\Gamma_k)$. We declare ua to be the new start word and cv to be the new target word. We can ensure that the distance from ua to cv in Γ_{k+1} is $d-1$, if we obey the following rule:

> **Rule 2:** Do not remove any arcs that lie on a shortest path from the start vertex to the target vertex.

We iterate this process of declaring the new start word and the new target word and constructing L_{k+1} by rules 1 and 2. After d steps, we have a graph L_{k+d} and a word $uw \in L_{k+d}$ which contains v. This means that the current start word uw contains an additional word of L_{n_i}.

In this way, we can pick up one word of L_{n_i} after the other until we have a word containing all of them, and the stage is completed. We summarize the result of the preceding arguments in the following lemma.

Lemma 4 *By following rule 2 we can ensure that the extension property is fulfilled.* □

Together with lemma 2 this implies plainly that sequences of complexity $2n$ exist. Such sequences can be constructed by following rules 1 and 2. The rules leave some freedom of choice (except in case 3 of figure 2), and thus there are many sequences of complexity $2n$.

Note that we need not abide by rule 2 all the time. In fact, we can remove arbitrary arcs subject only to rule 1 as long as we like (but not indefinitely); there is always time for repentantly returning to rule 2 and finishing the stage.

3 Some particular sequences

3.1 A first sequence

By following rules 1 and 2 and choosing among the possibilities offered by these rules according to a particular pattern I constructed a couple of sequences of complexity $2n$. One such sequence w can be described in the following way. We start with the word consisting of the single symbol /, and repeatedly transform it by the set of substitutions given in figure 3a. When we apply the substitution, we apply it to all elements of a word in parallel. Since the replacement string for the start symbol / starts with /, the word that results from applying the substitution $k+1$ times is an extension of the word which we get after k substitution steps. This means that the sequence of words converges to an infinite sequence, which is shown in the upper part of figure 3c. The little marks above and below the sequence show how far the word extends after 0,1,2,... substitutions. To the infinite word over the four-letter alphabet {/, \, _, ¯} we finally apply the letter-to-letter substitution in figure 3b, to obtain the 0-1-sequence w, which is shown in the lower part of figure 3c. The graphs in figure 1 of section 2 are the graphs Γ_4 and Γ_5 of this sequence. The four symbols /, \, _, and ¯ correspond to the four arrows in case 1A of the middle column of figure 2. The up and down movement of the word indicates the transitions between the two vertices of the figure.

In the theory of formal languages, a system like the one above for generating an infinite word is called a tag system (cf. Cobham [1972]), or when it is viewed as a method

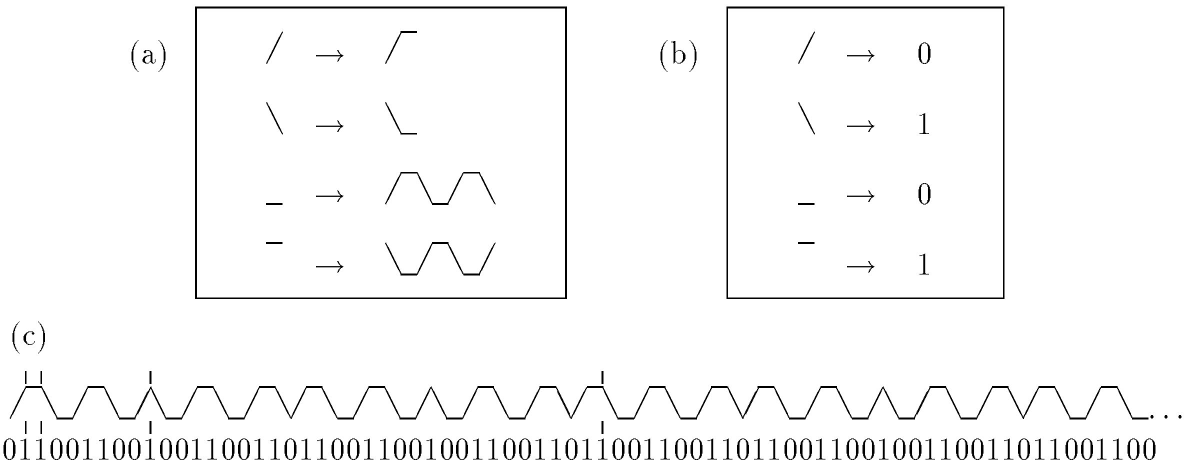


Figure 3: The substitutions and the infinite word w which they generate.

of generating a sequence of finite words, a CD0L system (cf. the book of Rozenberg and Salomaa [1980]).

We can generate the above sequence w by another mechanism, namely as the orbit of the mapping $x \mapsto (x+\theta) \bmod 2$, with $\theta = 1 - 1/\sqrt{5} \approx 0.5528$:

$$w_n = \begin{cases} 0, & \text{if } \ n\theta \bmod 2 \in [0,1), \\ 1, & \text{if } \ n\theta \bmod 2 \in [1,2). \end{cases} \tag{1}$$

This construction can be generalized by choosing a different breakpoint than the midpoint of the interval $[0,2)$ for selecting between the cases $w_n = 0$ and $w_n = 1$. For reasons which will become apparent in section 4, we have chosen the interval $[0,2)$ as the domain of our iteration mapping. In the following theorem and in the remainder of this section we take the more natural choice $[0,1)$.

Theorem 2 *Let c, φ, and θ be real numbers with $0 < \varphi < 1$, $0 < \theta < \min(\varphi, 1-\varphi)$, θ irrational, and $n\theta \not\equiv \varphi \pmod 1$, for all integers n. Then the sequence $w = w_1w_2w_3\ldots$ which is defined below has complexity $P(n) = 2n$:*

$$w_n = \begin{cases} 1, & \text{if } \ (c+n\theta) \bmod 1 \in [0,\varphi), \\ 0, & \text{if } \ (c+n\theta) \bmod 1 \in [\varphi,1). \end{cases}$$

Proof. Let us first see why the condition $\theta < \min(\varphi, 1-\varphi)$ is necessary: if $\theta \ge \varphi$ it is easy to check that the subword 11 cannot occur in w, and similarly, if $\theta \ge 1-\varphi$ the subword 00 cannot occur. Thus one of the four words 00, 01, 10, 11 of length 2 is missing and we cannot have $P(2) = 4$.

For the proof note that the subword $w_nw_{n+1}\ldots w_{n+l-1}$ depends only on the value $x_n := (c+n\theta) \bmod 1$. We can view the mapping $x \mapsto (x+\theta) \bmod 1$ that maps x_n to x_{n+1} as a rotation of the unit circle by the angle $\theta \cdot 2\pi$. The real numbers modulo 1, which are represented by the interval [0,1), correspond then to the points on the unit circle. The letter w_{n+d} depends on the position of $x_{n+d} = (x_n + d\theta) \bmod 1$ relative to the interval $[0,\varphi)$ on the unit circle. In other words, w_{n+d} depends on the relative position of x_n with respect to the two points $(0 - d\theta) \bmod 1$ and $(\varphi - d\theta) \bmod 1$. Thus, the $2l$ points $(-d\theta) \bmod 1$ and $(\varphi - d\theta) \bmod 1$, for $d = 0, 1, \ldots, l-1$, dissect the circle into at most $2l$ half-open circular intervals, and the subword $w_nw_{n+1}\ldots w_{n+l-1}$ of length

n depends on the interval into which x_n falls. It follows that $P(n) \leq 2n$. To show equality, we inductively prove the following claim:

> *Claim:* There are exactly $2l$ circular intervals; and they correspond to $2l$ different subwords of length l.

The first part of the claim is true because the irrationality assumptions on φ and θ guarantee that the $2l$ boundary points of the intervals are distinct. The second part is proved by induction on l. It is clearly true for $l = 1$. For the conclusion from l to $l+1$, note that the two "new" boundary points $(-l\theta) \bmod 1$ and $(\varphi - l\theta) \bmod 1$ fall into two different subintervals of the $2l$ subintervals so far, because they are for example separated by the two previous points $(-(l-1)\theta) \bmod 1$ and $(\varphi-(l-1)l\theta) \bmod 1$. (Here the assumption $0 < \theta < \min(\varphi, 1-\varphi)$ is used: the four points in question can be rotated to the points 0, φ, θ, and $\theta + \varphi$, for which the situation is clear.) This implies that for the 2 subintervals which are split by the new points, the corresponding subwords of length l are extended to length $l+1$ in two possible ways, whereas the remaining $2l-2$ subwords are extended in only one way. This clearly gives $2l+2$ *different* words of length $l+1$ for the $2l+2$ circular intervals. This concludes the proof of the claim.

Since the points x_n, $n = 1, 2, \ldots$, are dense in [0,1), every half-open circular subinterval contains such a point, and thus every word which correspond to a half-open circular subinterval actually occurs as a subword. □

Remark. Without the assumption $0 < \theta < \min(\varphi, 1-\varphi)$, we still have proved $P(n) \leq 2n$. In fact, there is some threshold n_0 such that for all $n > n_0$, even $P(n) = 2n$ is true; i. e., $P(n)$ is deficient only at the beginning.

The following lemma gives two further properties of the sequences which are generated as described above.

Lemma 5 *Let an infinite word of complexity $2n$ be constructed according to theorem 2.*

(a) *The parameter φ gives the approximate relative frequency of ones in long subwords of w. More precisely, let $\#_1(u)$ denote the number of ones in the finite subword u, and let $L_n(w)$ denote the number of n-letter subwords of w. Then*

$$\lim_{n\to\infty} \max\left\{ \frac{\#_1(u)}{n} : u \in L_n(w) \right\} = \lim_{n\to\infty} \min\left\{ \frac{\#_1(u)}{n} : u \in L_n(w) \right\} = \varphi.$$

(b) *The number of occurrences of the pattern* 01 *in a substring of length $n+1$ is either $\lfloor n\theta \rfloor$ or $\lceil n\theta \rceil$. Thus, θ gives the approximate relative frequency of the pattern* 01 *in long subwords of w.*

(c) *The number of occurrences of the pattern* 10 *in a substring of length $n+1$ is either $\lfloor n\theta \rfloor$ or $\lceil n\theta \rceil$. Thus, θ gives the approximate relative frequency of the pattern* 10 *in long subwords of w.*

Proof. (a): This follows from the classical fact that the sequence $x_n = (c+n\theta) \bmod 1$ is uniformly distributed in $[0, 1)$: $\#_1(w_l w_{l+1} \ldots w_{l+n-1})/n$ equals the proportion of the points $x_l, x_{l+1} \ldots x_{l+n-1}$ which lie in the interval $[0, \varphi)$, and this proportion converges to φ as n goes to ∞.

(b): The pattern 01 occurs at position l precisely if $y_l := c + l\theta \in [i-1, i)$ and $y_{l+1} = c+(l+1)\theta \in [i, i+1)$, for some integer i. In a sequence of $n+1$ successive values

$y_l, y_{l+1}, \ldots, y_{l+n}$ this happens $\lfloor y_{l+n} \rfloor - \lfloor y_l \rfloor = \lfloor y_l + n\theta \rfloor - \lfloor y_l \rfloor$ times. For any value of y_l, this number is equal to one of the two values stated in the lemma.

Part (c) is analogous to (b). □

The method of theorem 2 for generating an infinite sequence is very similar to a method for generating Sturmian sequences. There is in fact a very close relation between sequences with $\varphi = 1/2$ (like the sequence w of figure 3 and (1)) and Sturmian sequences, which will be investigated in section 4.

3.2 Another sequence

In this subsection give a different sequence which cannot be generated by theorem 2. It is again described by a substitution on a four-letter alphabet $\{/, \backslash, \mathsf{L}, {}^{-}\}$ (see figure 4a). The start symbol is /. The final homomorphism to the alphabet $\{0, 1\}$, which is shown in figure 4b, is now not just a letter-to-letter mapping, but maps different symbols to 0-1-words of different lengths. (In formal language theory this is called a HD0L system.) The resulting words are shown in figure 4c. In this sequence, the only case among the rows in figure 2 that arises is case 2D, apart from the trivial case 3.

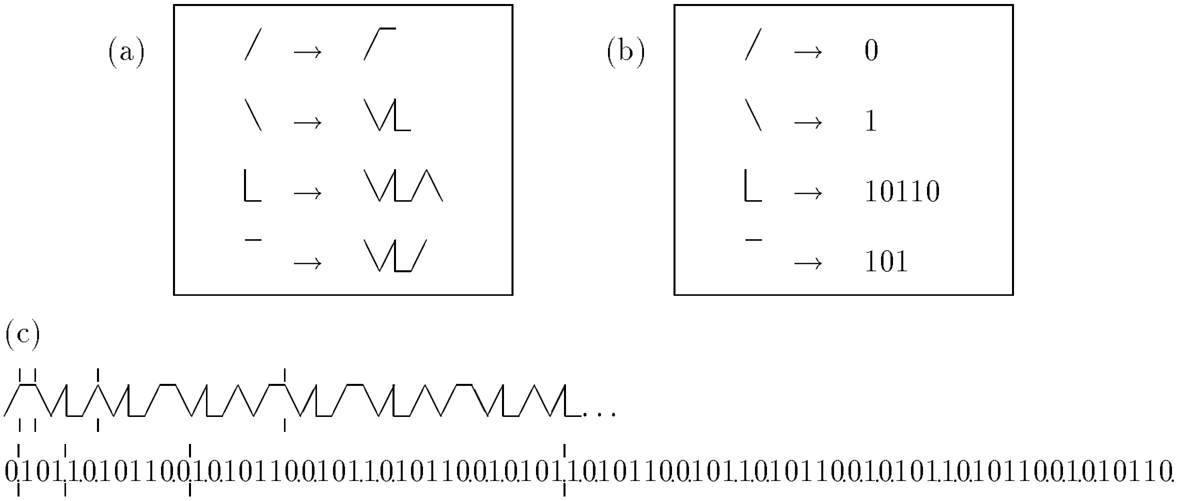


Figure 4: Another set of substitutions generating an infinite word with complexity $2n$. The dots in the last line are for orientation only; they separate the words that correspond to single symbols of the first line.

By studying the eigenvalues of the 4×4-matrix which specifies the number of symbols of each type in the replacement string for each symbol in figure 4a one can compute the limiting relative frequency φ of ones and the limiting relative frequency θ of the subword 01. The exact expression of these numbers by radicals is unwieldy and therefore not given; their approximate values are $\varphi \approx 0.5346$ and $\theta \approx 0.3737$. By lemma 5, if our sequence is constructed according to theorem 2, it must have these values of φ and θ. It is easily checked that such a sequence contains the subword 0100, which is not contained in the sequence of figure 4. On the other hand, the latter sequence contains 1100, which is contained in no sequence generated by theorem 2 with the given values of φ and θ.

3.3 A third sequence

All sequences considered so far have the property that the distance between any two adjacent occurrences of a given finite word is bounded. I conclude this section with a sequence of complexity $2n$ for which this is not true. Figure 5 shows the substitution and the resulting sequence. This substitution is particularly simple and works directly with the final alphabet $\{0, 1\}$. The start symbol is 0. Clearly, the sequence contains arbitrarily long blocks of ones, and thus there is no bound n_0 such that, for example, the subword 00 is contained in every subword of length n_0.

$$\boxed{\begin{array}{ccc} 0 & \rightarrow & 001 \\ 1 & \rightarrow & 111 \end{array}}$$

0|0.1|001.111|001001111.111111111|001001111001001111111111111.1111111111111111...

Figure 5: Another substitution generating an infinite word with complexity $2n$.

4 A relation with Sturmian sequences

According to Morse and Hedlund [1940], an infinite 0-1-sequence is called *Sturmian*, if the lengths of any two subwords which start and end with zero and contain the same number of zeros differs by at most one. Equivalently, the number of ones in any two subwords of the same length differ by at most one (see also Coven and Hedlund [1973], section 3).

We recall a few basic facts about Sturmian sequences from the papers cited above. Note that we deal only with one-way infinite sequences (rays), whereas the word "sequences" in the cited papers includes also doubly-infinite sequences (trajectories, series) and finite sequences (blocks).

A Sturmian sequence is either periodic from some point on, in which case the complexity is bounded, or it has complexity $P(n) = n + 1$. Any sequences with complexity $P(n) = n + 1$ is an aperiodic Sturmian sequence.

Such an aperiodic Sturmian sequence $\beta = \beta_1\beta_2\beta_3\ldots$ can be generated in the following way. For given real numbers c and θ between 0 and 1 with θ irrational, we either define β_n for $n = 1, 2, 3, \ldots$ as

$$\beta_n := \begin{cases} 1, & \text{if } (c + n\theta) \bmod 1 \in [0, \theta), \\ 0, & \text{if } (c + n\theta) \bmod 1 \in [\theta, 1). \end{cases} \tag{2}$$

or as

$$\beta_n := \begin{cases} 1, & \text{if } (c - n\theta) \bmod 1 \in [0, \theta), \\ 0, & \text{if } (c - n\theta) \bmod 1 \in [\theta, 1). \end{cases} \tag{2'}$$

Lemma 6 *An infinite sequence has complexity $P(n) = n + 1$ if and only if it can be constructed by (2) or by (2′) with some irrational θ.* □

These aperiodic sequences are called *irrational Sturmian sequences.* In such a sequence, the number of zeros in a subword of length n is either $\lfloor n\theta \rfloor$ or $\lceil n\theta \rceil$. It follows that the proportion of zeros in subwords of increasing length converges to θ. Similarly, the number of ones divided by the number of zeros tends to a limit $\alpha = \theta/(1-\theta)$ which is called the *frequency* of the sequence.

By setting $c = \theta = 1 - 1/\sqrt{5}$ and using (2) we get a Sturmian sequence whose close relation to the sequence w which was defined by (1) in the section 3.1 is obvious:

$$\beta_n := \begin{cases} 1, & \text{if } (n+1)\theta \bmod 1 \in [0,\theta), \\ 0, & \text{if } (n+1)\theta \bmod 1 \in [\theta,1). \end{cases}$$

In fact, we have

$$\beta_n = \begin{cases} 1, & \text{if } w_{n+1} \neq w_n, \\ 0, & \text{if } w_{n+1} = w_n. \end{cases}$$

In other words, $\beta_n = (w_{n+1} - w_n) \bmod 2$ is a kind of *difference sequence* or *derivative* of the sequence w. On the other hand, $w_n = (\beta_1 + \beta_2 + \cdots + \beta_{n-1}) \bmod 2$ can be obtained as the *partial sums sequence* of the sequence β.

The sequence w has a strong symmetry property with respect to complementation of its symbols (interchanging zeros and ones). For any finite subword of w, the complemented word also occurs in w. This can be easily seen from the definition in figure 3a–b, because the rules are completely symmetric with respect to exchanging / with \, _ with ‾, and 0 with 1. We call a 0-1-sequence with this property *complementation-symmetric.* We will show that any sequence with complexity $2n$ that fulfills this symmetry condition is obtainable from a Sturmian sequence in the way described above, namely as its partial sums sequence.

Theorem 3 *An infinite 0-1-sequence $w = w_1w_2w_3\ldots$ is a complementation-symmetric sequence with complexity $P(n) = 2n$ if and only if its difference sequence $\beta = \beta_1\beta_2\beta_3\ldots$, which is defined by $\beta_n = (w_{n+1} - w_n) \bmod 2$, is an irrational Sturmian sequence.*

Proof. "Only if": Consider the set L_{n+1} of subwords of w of length $n+1$. Every subword x of β of length n can be obtained as the "difference word" of some word $y \in L_{n+1}$. L_{n+1} contains $P_w(n+1) = 2n+2$ different words, but two words $y \in L_{n+1}$ which are complements of each other yield the same difference word x. This gives $P_\beta(n) = P_w(n+1)/2 = n+1$ different words x of length n.

"If": For an irrational Sturmian sequence β with complexity $n+1$, we have to show that its partial sums sequence w_n, which is defined by $w_n = (w_1 + \beta_1 + \beta_2 + \cdots + \beta_{n-1}) \bmod 2$, where w_1 can be arbitrarily set to 0 or 1, has complexity $2n$ and is complementation-symmetric. Each subword x of β of length n gives rise to one of two complementary subwords y of w of length $n+1$, depending on the position where it occurs: for $x = \beta_l\beta_{l+1}\beta_{l+2}\ldots\beta_{l+n-1}$, the word $y = w_lw_{l+1}w_{l+2}\ldots w_{l+n}$ is one of the two words whose difference word is x, depending on w_l. If we show that *both* of these words occur in L_{n+1}, for a given x, we have at once proved that w is complementation-symmetric and that is has the correct complexity: each of the $P_\beta(n) = n+1$ words x of length n gives 2 words y of length $n+1$, and thus $P_w(n+1) = 2(n+1)$.

To finish the proof of theorem 3, we will need a definition and an intermediate theorem.

Definition. Suppose that a subword x occurs in two different positions in an infinite word w, i. e., the word w can be written in the form $w = ux\ldots = uvx\ldots$, for some words u and v. Then we call v the *offset* between these two occurrences of x.

In other words, the offset is the subword between the "starting points" of the two occurrences. Note that the offset includes the first occurrence of x (or at least part of it if the two occurrences of x overlap).

Theorem 4 *Every subword x of an irrational Sturmian sequence has two occurrences with an offset containing an odd number of ones.*

Note that this is what is needed for the proof of theorem 3, because the two occurrences x give rise to both words y whose difference word is x.

Note also that the above definition of the offset is somewhat arbitrary because it depends on some "reference point" of x. (In our case this is the starting point.) However, it is easy to see that the parity of the number of ones does not change if we choose a different reference point in the definition of the offset, as long as the reference point lies inside x or at the boundary of x.

Proof. In w, each zero is followed either by $\lfloor\alpha\rfloor$ ones or by $\lceil\alpha\rceil$ ones. If the pattern x consists of at most $\lfloor\alpha\rfloor$ ones and no zeros, the theorem follows directly: we find in a block of $\lceil\alpha\rceil$ ones two occurrences of x whose offset is a single 1, and the theorem is proved.

Otherwise, we use a transformation which reduces w to another infinite sequence w' and the pattern x to a shorter pattern x'. (This is essentially the same as the process called derivation in Morse and Hedlund [1940] and transformation (B) in Flor [1962].) This reduction has the property that every occurrence of x' in w' corresponds to some occurrence of x in w.

We take the sequence w and cut it before each 0, thus decomposing it into blocks which consist of an initial zero followed by ones. A possible initial block of ones is discarded.

There are two types of blocks, with $\lfloor\alpha\rfloor$ ones and with $\lceil\alpha\rceil$ ones, respectively. Now we replace each block that has an even number of ones by a 0 and each block with an odd number of ones by a 1, and we call the resulting sequence w' the *reduced sequence*. According to Morse and Hedlund [1940, section 8] the reduced sequence is again irrational Sturmian, with frequency $\alpha' = (\alpha - \lfloor\alpha\rfloor)/(1 - (\alpha - \lfloor\alpha\rfloor))$ if $\lfloor\alpha\rfloor$ is even and $\alpha' = (1 - (\alpha - \lfloor\alpha\rfloor))/(\alpha - \lfloor\alpha\rfloor)$ if $\lfloor\alpha\rfloor$ is odd.

We also have to reduce the pattern x. Before replacing blocks that start with zero by single letters as above, we apply some "cosmetic" changes to x that do not affect the occurrences of x in w in an essential way. If x starts with $\lceil\alpha\rceil$ ones, we can prepend a zero to x: since any sequence of $\lceil\alpha\rceil$ ones in w is preceded by a zero, this does not change the occurrences of x and is therefore permitted. (An initial occurrence of x as the starting word of w might be lost.)

Now we are sure that x contains at least a zero. If x contains ones before the first zero, there can be at most $\lfloor\alpha\rfloor$ of them, and we omit them. Again, this does not introduce additional occurrences of x in w. (Remember that, before forming w', we have deleted initial ones from w; hence there will be no additional occurrence of the shortened x close to the beginning of w.)

If the last zero of x is followed by fewer than $\lceil\alpha\rceil$ ones, we discard this zero and the following ones. (It is possible that x becomes the empty word.) Now we apply the reduction procedure to x, as described above for w, yielding x'. Every occurrence of x' in w' corresponds uniquely to an occurrence of x in w. (This is true even if x' is the empty word; the empty word occurs at every position of w'.) Furthermore, the offset between two occurrences of x' in w' contains an odd number of ones if and only if this holds for the offset between the corresponding occurrences of x in w. Thus it is sufficient to prove the theorem for x' and w' in place of x and w.

In each reduction step, the pattern x will be shortened: if x contains ones, the ones will either be deleted or they will be merged with a preceding zero into a single letter; and if x consists only of zeros, the last zero will be canceled. It follows that the sequence of reductions eventually bottoms out in a pattern with at most $\lfloor \alpha \rfloor$ ones for which the direct proof works. (This case includes the empty pattern.) □

With this proof the proof of theorem 3 is also complete.

Note that the proof even proves the existence of two *adjacent* occurrences of x with an offset containing an odd number of ones (i. e., two occurrences with no other occurrences in between). Since there is nothing special about the symbol 1, we also know that there are adjacent occurrences of x with an offset containing an odd number of *zeros*. It is an easy matter to prove that every subword x has two (not necessarily adjacent) occurrences with an offset containing an even number of ones (or an even number of zeros).

The technique of the proof of theorem 3 has also been used by Allouche [1992] to relate the complexity of generalized Rudin–Shapiro sequences in the sense of Mendès France and Tenenbaum [1981] to the complexity of paperfolding sequences. (For a survey on various aspects of paperfolding sequences see for example Dekking, Mendès France, and van der Poorten [1982].) A generalized Rudin–Shapiro sequence w is defined as the partial sums sequence of a paperfolding sequence u, and the property of theorem 4 holds for every paperfolding sequence, which yields $P_w(n) = 2P_u(n-1)$.

Corollary *An infinite 0-1-sequence $w = w_1 w_2 w_3 \ldots$ is a complementation-symmetric sequence with complexity $P(n) = 2n$ if and only if it can be generated according to theorem 2 with $\varphi = 1/2$.*

Proof. It is easy to see that the sequences generated by theorem 2 with $\varphi = 1/2$ are complementation-symmetric. The other direction follows with the help of lemma 6. □

Note that the sequences generated by theorem 2 with $\varphi = 1/2$ were essentially already considered by Flor [1962], who investigated sequences of ±1's that can be written as

$$\operatorname{sgn} \sin 2\pi(c + n\theta).$$

Apart from the possible occurrence of zeros in these sequences, they coincide with the complementation-symmetric sequences considered in this section.

5 Future work

We have presented a general method and a couple of special methods for constructing sequences of complexity $2n$, and we have given several examples with different properties. However, the totality of these sequences is far from being well understood.

By a more careful analysis it should be possible to give a scheme for generating all words of complexity $2n$ using the expressive power of L-systems (see Rozenberg and

Salomaa [1980]), as in figures 3, 4, and 5. For Sturmian sequences such a scheme is known, see Rauzy [1985] or Arnoux and Rauzy [1991]. I will investigate this question in a future paper.

It seems less difficult to extend the methods of section 2 to complexity functions like $P(n) = 2n + k$, for a fixed positive k, or to relax the condition of strong connectedness (rule 1), cf. theorem 1. It would only be necessary to include in figure 2 cases with $\delta^-(v) = 3$, $\delta^-(v) = 4$, and $\delta^+(v) = 3$. It should also be no problem to include doubly infinite sequences.

References

1. J.-P. ALLOUCHE [1987],
   Automates finis en théorie des nombres, *Expositiones Math.* **5**, 239–266.
2. J.-P. ALLOUCHE [1992],
   The number of factors in a paperfolding sequence, *Bull. Austral. Math. Soc.* **46**, 23–32
3. P. ARNOUX AND G. RAUZY [1991],
   Représentation géométrique des suites the complexité $2n + 1$, *Bull. Soc. Math. France* **119**, 199–215.
4. A. COBHAM [1972],
   On uniform tag systems, *Math. Systems Theory* **6**, 164–192.
5. E. M. COVEN AND G. A. HEDLUND [1973],
   Sequences with minimal block growth, *Math. Systems Theory* **7**, 138–153.
6. M. DEKKING, M. MENDÈS FRANCE, AND A. VAN DER POORTEN [1982],
   FOLDS!, *Math. Intelligencer* **4**, 130–138, 173–181, and 190–195; see also L. AUTEURS, Corrigendum, *Math. Intelligencer* **5**, 2 (1983), p. 5.
7. P. FLOR [1962],
   Ein Verteilungsproblem für arithmetische Folgen, *Abh. Math. Sem. Univ. Hamburg* **25**, 62–70.
8. M. MENDÈS FRANCE AND G. TENENBAUM [1981],
   Dimension des courbes planes, papiers pliés et suites de Rudin–Shapiro, *Bull. Soc. Math. France* **109**, 207–215.
9. M. MORSE AND G. A. HEDLUND [1940],
   Symbolic dynamics II. Sturmian trajectories, *Amer. J. Math.* **62**, 1–42.
10. G. RAUZY [1985],
    Mots infinis en arithmétique, in: *Automata on Infinite Words, École de Printemps d'Informatique Théorique, Le Mont Dore, May 1984*, ed. M. Nivat and D. Perrin, Lecture Notes in Computer Science, vol. **192**, Springer-Verlag, Berlin etc., pp. 165–171.
11. G. ROZENBERG AND A. SALOMAA [1980],
    The Mathematical Theory of L Systems, Academic Press, New York etc.
12. A. THUE [1906],
    Über unendliche Zeichenreihen, *Norske Vid. Selsk. Skr. I, Math.-Nat. Kl.* **7**, 1–22.