Chapter 1

SEQUENCES OF LOW COMPLEXITY: AUTOMATIC AND STURMIAN SEQUENCES

Valérie BERTHÉ
Institut de Mathématiques de Luminy
CNRS-UPR 9016
Case 907, 163 avenue de Luminy
F-13288 Marseille Cedex 9
France

The complexity function is a classical measure of disorder for sequences with values in a finite alphabet: this function counts the number of factors of given length. We introduce here two characteristic families of sequences of low complexity function: automatic sequences and Sturmian sequences. We discuss their topological and measure-theoretic properties, by introducing some classical tools in combinatorics on words and in the study of symbolic dynamical systems.

1.1 Introduction

The aim of this course is to introduce two characteristic families of sequences of low "complexity": automatic sequences and Sturmian sequences (complexity is defined here as the combinatorial function which counts the number of factors of given length of a sequence over a finite alphabet). These sequences not only occur in many mathematical fields but also in various domains as theoretical computer science, biology, physics, crystallography...

We first define some classical tools in combinatorics on words and in the study of symbolic dynamical systems: the complexity function and frequencies of factors in connection with the notions of topological and measure-theoretic entropy (Sections 1.2 and 1.3), the graphs of words (Section 1.4), special and bispecial factors (Section 1.5). Then we study Sturmian sequences in Section 1.6: these sequences are defined as the sequences of minimal complexity among non-ultimately periodic sequences. This combinatorial definition has the particularity of being equivalent to the following simple geometrical representation: a Sturmian sequence codes the orbit of a point of the unit circle under a rotation by irrational angle α with respect to a partition of the unit circle into two intervals of lengths α and $1-\alpha$. Sturmian sequences have thus remarkable combinatorial and arithmetical properties. Then we introduce automatic sequences in Section 1.7: an automatic sequence is defined as the image by a letter-to-letter projection of a fixed point of a substitution of constant length or equivalently as a finite-state function of the representation of the index in a given basis. We emphasize on the connections with transcendence of formal power series with coefficients in a finite field. In particular, we will try to answer the following question: how to recognize if a sequence is automatic or not? We conclude this course by discussing the connections between sequences with a linear growth order for the complexity function, and substitutions.

1.2 Complexity Function

1.2.1 Definition

Let us introduce a combinatorial measure of disorder for sequences over a finite alphabet: this notion is called *(symbolic) complexity*. For more information on the subject, we refer the reader to the surveys [8, 43] and to the course [59].

In all that follows we restrict ourselves to sequences over a finite alphabet indexed by the set $\mathbb{N}$ of non-negative integers. A *factor* of the infinite sequence $u=(u_n)_{n\in\mathbb{N}}$ is a finite block w of consecutive letters of u, say $w=u_{n+1}\cdots u_{n+l}$; l is called the *length* of w, denoted by $|w|$. Let $p(n)$ denote the *complexity function* of sequence u with values in a finite alphabet: it counts the number of distinct factors of length n of the sequence u. The complexity function is obviously non-decreasing and for any integer n, one has $1\leq p(n)\leq d^n$, where d denotes the cardinality of the alphabet.

This function can be considered to *measure the predictability* of a sequence. The first difference of the complexity function counts the number of possible extensions in the sequence of factors of given length. We call *right extension* (respectively *left extension*) of a factor w a letter x such that wx (respectively xw) is a factor of the sequence. Let w^+ (respectively w^-) denotes the number of right (respectively left) extensions of w. (One may have $w^- = 0$ but always $w^+ \geq 1$.) We have

$$p(n+1) = \sum_{|w|=n} w^+ = \sum_{|w|=n} w^-,$$

and thus

$$p(n+1) - p(n) = \sum_{|w|=n} (w^+ - 1) = \sum_{|w|=n} (w^- - 1).$$

Exercise 1.2.1 (see [31, 54]) Prove that a sequence is *ultimately periodic* (i.e., periodic from a certain index on) if and only if its complexity function satisfies

$$\exists n,\ \ p(n) \leq n \Longleftrightarrow \exists C,\ \forall n\ \ p(n) \leq C.$$

What happens in the case of a sequence defined over $\mathbb{Z}$?

The complexity function is a measure of disorder connected to the topological entropy: the *topological entropy* [1] is defined as the exponential growth rate of the complexity as the length increases

$$H_{top}(u) = \lim_{n \to +\infty} \frac{\log_d(p(n))}{n}.$$

It is easy to check that this limit exists because of the subadditivity of the function $n \mapsto \log_d(p(n))$. Note that the word *entropy* is used here as a measure of randomness or disorder. For a survey on the connections between entropy and sequences, see [13].

The study of the complexity is mainly concerned with the following three questions.

- How to compute the complexity of a sequence?
- Which functions can be obtained as the complexity function of some sequence?
- Can one deduce from the complexity a geometrical representation of sequences?

We will see how to answer the first question by introducing special and bispecial factors, in some particular cases of substitutive sequences (Section 1.5). The second question is still very much in progress and far from being solved (in particular in the case of positive entropy): for a survey on the question, see [24, 43]. Although the complexity function is in general not sufficient to describe a sequence, we will see in Section 1.6 that much can be said on the geometrical properties in the case of lowest complexity, i.e., in the case of *Sturmian sequences*: these sequences are defined to have exactly $n+1$ factors of length n, for any integer n. By Exercise 1.2.1 a sequence with complexity satisfying $p(n) \leq n$ for some n is ultimately periodic. Sturmian sequences have thus the minimal complexity among all sequences that are not ultimately periodic.

Exercise 1.2.2 Deduce from Exercise 1.2.1 that every prefix of a Sturmian sequence appears at least two times in the sequence. Deduce that the factors of every Sturmian sequence appear infinitely often (such a sequence is called *recurrent*).

1.2.2 Frequencies and Measure-Theoretic Entropy

The purpose of this section is to introduce a more "precise" (in a sense that we will see in Section 1.2.3) measure of disorder of sequences, connected with frequencies of factors. The *frequency* $f(B)$ of a factor B of a sequence (called *density* in Host's course) is defined as the limit, if it exists, of the number of occurrences of this block in the first k terms of the sequence divided by k.

Exercise 1.2.3 Construct a sequence for which the frequencies of letters do not exist.

Let us first introduce *the block entropies* for sequences with values in a finite alphabet in order to define the notion of *measure-theoretic entropy*. These sequences of block entropies were first introduced by Shannon in information theory, to measure the entropy of the English language (see [65]).

Let u be a sequence with values in the alphabet $\mathcal{A} = \{1, \cdots, d\}$. We suppose that all the block frequencies exist for u. Let

$$P(x|x_1 \cdots x_n) = \frac{f(x_1 \cdots x_n x)}{f(x_1 \cdots x_n)},$$

where $x_1 \cdots x_n$ is a block of non-zero frequency and x a letter. Intuitively $P(x|x_1 \cdots x_n)$ is the conditional probability that the letter x follows the block $x_1 \cdots x_n$ in the sequence u. We are going to associate with the sequence u two sequences of block entropies $(H_n)_{n \in \mathbb{N}}$ and $(V_n)_{n \in \mathbb{N}}$.

For all $n \geq 1$, let

$$V_n = \sum L(f(x_1 \cdots x_n)),$$

where the sum is over all the factors of length n and $L(x) = -x \log_d(x)$, for all $x \neq 0$ and $L(0) = 0$. We put $V_0 = 0$.

For all $n \geq 1$, let

$$H_n = \sum{}' f(x_1 \cdots x_n) H(x_1 \cdots x_n), \tag{1.1}$$

where the sum is over all the blocks of length n of non-zero frequency and

$$H(x_1\cdots x_n)=\sum_{x\in\mathcal{A}}L(P(x/x_1\cdots x_n)).$$

We put $H_0=V_1$. The sequence $(H_n)_{n\in\mathbb{N}}$ measures in some way the properties of predictability of the initial sequence u.

Exercise 1.2.4 Prove that: $\forall n\in\mathbb{N},\ H_n=V_{n+1}-V_n$. (This classical property in information theory is called the *chain-rule.*)

Thus, $(H_n)_{n\in\mathbb{N}}$ is the discrete derivative of $(V_n)_{n\in\mathbb{N}}$. Note that $(V_n)_{n\in\mathbb{N}}$ is a non-decreasing sequence, since $H_n\geq 0$ for all n.

It can be shown that $(H_n)_{n\in\mathbb{N}}$ is a monotonic non-increasing sequence of n (see, for instance [16]). The intuitive meaning of this is that the uncertainty about the choice of the next symbol decreases when the number of known preceding symbols increases. From the non-increasing behaviour of the positive sequence $(H_n)_{n\in\mathbb{N}}$, we deduce the existence of the limit $\lim_{n\to+\infty}H_n$. We have: $\forall n,\ H_n=V_{n+1}-V_n$ and $\sum_{k=0}^{n-1}H_k=V_n$. By taking Cesàro means, we obtain:

$$\lim_{n\to+\infty}H_n=\lim_{n\to+\infty}\frac{V_n}{n}.$$

This limit is called the *measure-theoretic entropy* of the sequence u, it is the limit of the entropy per symbol of the choice of a block of length n, when n tends to infinity.

1.2.3 Variational Principle

What is the relation between the sequences $(H_n)_{n\in\mathbb{N}}$ and $(V_n)_{n\in\mathbb{N}}$? We have:

$$\forall n,\ nH_n\leq\sum_{k=0}^{n-1}H_k=V_n=\sum L(P(x_1\cdots x_n)).$$

By concavity of the function L we get: $\forall n\geq 1,\ V_n\leq\log_d p(n)$. Hence the following proposition:

Proposition 1.2.5 *We have* $H_n\leq\frac{\log_d(p(n))}{n}$, *for all* $n\geq 1$.

We hence get:

$$\lim_{n\to+\infty}H_n=\lim_{n\to+\infty}\frac{V_n}{n}=H(u)\leq H_{top}(u)=\lim_{n\to+\infty}\frac{\log_d(p(n))}{n}.$$

This inequality is a particular case of a basic relationship between topological entropy and measure-theoretic entropy called the *variational principle* (for a proof see [53]).

The two limits $\lim\limits_{n\to+\infty} H_n$ and $\lim\limits_{n\to+\infty} \frac{\log_d(p(n))}{n}$ are distinct in general and the notion of measure-theoretic entropy for a sequence is more precise. But the sequences we are mostly dealing with here are *deterministic*, i.e., sequences with zero entropy. Therefore neither the metrical nor the topological entropy can distinguish between these sequences.

1.3 Symbolic Dynamical Systems

Recall some basic notions on symbolic dynamical systems. For a detailed introduction to the subject, see [57]. Let $\mathcal{A}$ denote a finite alphabet; here we work with the space $\mathcal{A}^{\mathbb{N}}$, whereas in Host's course it is $\mathcal{A}^{\mathbb{Z}}$.

Endow the set $\mathcal{A}^{\mathbb{N}}$ of all sequences with values in the finite set $\mathcal{A}$ with the product of discrete topologies on $\mathcal{A}$. This set is thus a compact space. The topology defined on $\mathcal{A}^{\mathbb{N}}$ is equivalent to the topology defined by the following metrics: for $x, y \in \mathcal{A}^{\mathbb{N}}$

$$d(x,y) = (1 + \inf\{k \geq 0;\ x_k \neq y_k\})^{-1}.$$

Two sequences are thus close to each other if their first terms coincide. The *cylinder* $[w]$, where $w = w_1 \ldots w_n$ belongs to $\mathcal{A}^n$, is the set of sequences of the form

$$[w] = \{x \in \mathcal{A}^{\mathbb{N}} |\ x_0 = w_1,\ x_1 = w_2, \ldots, x_{n-1} = w_n\}.$$

Cylinders are closed and open sets and span the topology.

The space $\mathcal{A}^{\mathbb{N}}$ is complete as a metric compact space. Let us deduce from this the existence of fixed points of substitutions. A *substitution* defined on the finite alphabet $\mathcal{A}$ is a map from $\mathcal{A}$ to the set of words defined on $\mathcal{A}$, denoted by $\mathcal{A}^*$, extended to $\mathcal{A}^*$ by concatenation, or in other words, a homomorphism of the free monoid $\mathcal{A}^*$ (see also [49] for a precise study of substitution dynamical systems).

Exercise 1.3.1 Let σ be a substitution and a be a letter such that $\sigma(a)$ begins by a and $|\sigma(a)| \geq 2$. Prove that there exists a unique sequence beginning with a satisfying $\sigma(u) = u$. This sequence is called a *fixed point* of the substitution.

For instance, the Fibonacci sequence is defined as the fixed point beginning with 1 of the following substitution

$$\sigma(1) = 10,\ \sigma(0) = 1.$$

Let T denote the following map defined on $\mathcal{A}^{\mathbb{N}}$, called the *one-sided shift*:

$$T((u_n)_{n\in\mathbb{N}}) = (u_{n+1})_{n\in\mathbb{N}}.$$

The map T is uniformly continuous, onto but not necessarily one-to-one on $\mathcal{A}^{\mathbb{N}}$.

Exercise 1.3.2 Recall that a sequence is said to be *recurrent* if every factor appears at least two times, or equivalently if every factor appears an infinite number of times in this sequence.

Prove that a sequence u is recurrent if and only if there exists a strictly increasing sequence $(n_k)_{k\in\mathbb{N}}$ such that

$$u = \lim_{k\to+\infty} T^{n_k}u.$$

Let u be a sequence with values in $\mathcal{A}$. Define $\overline{\mathcal{O}}(u)$ as the positive orbit closure of the sequence u under the action of the shift T, i.e., the closure of the set $\mathcal{O}(u) = \{T^n(u),\ n \geq 0\}$. The set $\overline{\mathcal{O}(u)}$ is a compact metric space, and thus complete. It is also T-*invariant*: $T(\overline{\mathcal{O}(u)}) \subset \overline{\mathcal{O}(u)}$. In other words T may be considered as acting on $\overline{\mathcal{O}}(u)$.

Exercise 1.3.3

1. Prove that
$$\overline{\mathcal{O}}(u) = \{x \in \mathcal{A}^{\mathbb{N}},\ L(x) \subset L(u)\},$$
where $L(x)$ denotes the set of factors of the sequence x.

2. Prove that u is recurrent if and only if T is onto on $\overline{\mathcal{O}}(u)$.

Let X be a non-empty compact metric space and T be a continuous map from X to X. The system (X, T) is called a *topological dynamical system.* For instance, $(\overline{\mathcal{O}(u)}, T)$ is a topological dynamical system. A topological dynamical system is called *minimal* if every closed T-invariant set E is either equal to the full set X or to the empty set.

Exercise 1.3.4

- Prove that (X, T) is minimal if and only if $X = \overline{\mathcal{O}(x)}$, for every element x of X.

- A sequence is said to be *uniformly recurrent* if every factor appears infinitely often and with bounded gaps (or, equivalently, if for every integer n, there exists an integer m such that every factor of u of length m contains every factor of length n). Prove that a sequence u is uniformly recurrent if and only if $(\overline{\mathcal{O}(u)}, T)$ is minimal. (If w is a factor of u, write
$$\overline{\mathcal{O}(u)} = \bigcup_{n\in\mathbb{N}} T^{-n}[w],$$
and conclude by a compactness argument.)

The following special case of the Daniell-Kolmogorov consistency theorem (see for instance [73]) establishes the existence of a certain probability measure on $(\overline{\mathcal{O}(u)}, T)$. A Borel probability measure μ defined on $(\overline{\mathcal{O}(u)}, T)$ is called T-*invariant* if $\mu(T^{-1}(B)) = \mu(B)$, for any Borel set B.

Theorem 1.3.5 *Let u be a sequence on $\mathcal{A} = \{1, \ldots, d\}$. Consider a family of maps $(p_n)_{n \geq 1}$, where p_n is a map from $\mathcal{A}^n$ to $\mathbb{R}$, such that*

- *for any word w in $\mathcal{A}^n$, $p_n(w) \geq 0$,*
- $\displaystyle\sum_{i=1}^{d} p_1(i) = 1,$
- *for any word $w = w_1 \ldots w_n$ in $\mathcal{A}^n$, $p_n(w) = \displaystyle\sum_{i=1}^{d} p_{n+1}(w_1 \ldots w_n i)$.*

Then there exists a unique probability measure μ on $\mathcal{A}^{\mathbb{N}}$ defined on the cylinders by $\mu([w_1 \ldots w_n]) = p_n(w_1 \ldots w_n)$.

Furthermore, if for any n and for any word $w = w_1 \ldots w_n$ in $\mathcal{A}^n$,

$$p_n(w) = \sum_{i=1}^{d} p_{n+1}(i w_1 \ldots w_n),$$

then this measure is T-invariant.

In particular, if all frequencies exist, then there exists a unique T-invariant probability measure which assigns to each cylinder the frequency of the corresponding factor. Moreover suppose the symbolic dynamical system $(\overline{\mathcal{O}(u)}, T)$ is *uniquely ergodic*, i.e., there exists a unique T-invariant probability measure μ on this dynamical system. Thus a precise knowledge of the frequencies allows a complete description of the measure μ. For instance, a symbolic dynamical system obtained via the fixed point of a primitive substitution [49, 57], or via a Sturmian sequence is uniquely ergodic.

1.4 The Graph of Words

The Rauzy graph Γ_n of words of length n of a sequence on a finite alphabet $\mathcal{A}$ (of cardinality d) is an oriented graph (see, for instance, [58]), which is a subgraph of the de Bruijn graph of words[1] (see [32]). Its vertices are the

[1] The de Bruijn graph of words corresponds to the graph of words of a sequence of maximal complexity ($\forall n, p(n) = d^n$) and was introduced by de Bruijn in order to construct circular finite sequences of length d^n with values in $\{0, 1, \ldots, d-1\}$ such that every factor of length n appears once and only once: such a sequence corresponds to a Hamiltonian closed path in de Bruijn graph.

factors of length n of the sequence and the edges are defined as follows: there is an edge from U to V if V follows U in the sequence, i.e., if there exists a word W and two letters x and y such that $U = xW$, $V = Wy$ and xWy is a factor of the sequence. There are $p(n+1)$ edges and $p(n)$ vertices, where $p(n)$ denotes the complexity function.

Exercise 1.4.1 Prove that the graphs of words of a sequence are always connected. Prove the following equivalence (see [61]):

- the sequence u is recurrent,
- every factor of u appears at least twice,
- the graphs of words are strongly connected.

Let U be a vertex of the graph. Denote by U^+ the number of edges of Γ_n with origin U and by U^- the number of edges of Γ_n with end vertex U. In other words, U^+ (respectively U^-) counts the number of right (respectively left) extensions of U. Recall that

$$p(n+1) = \sum_{|U|=n} U^+ = \sum_{|U|=n} U^-,$$

and thus

$$p(n+1) - p(n) = \sum_{|U|=n} (U^+ - 1) = \sum_{|U|=n} (U^- - 1).$$

Exercise 1.4.2 Recall that a Sturmian sequence is defined as a sequence of complexity function $p(n) = n + 1$, for every positive integer n, and that it is recurrent (Exercise 1.2.2).

- For any positive integer n, prove that there exists a unique factor of length n having two right (respectively left) extensions: such a factor is called a *right* (respectively *left*) *special factor* (or also *expansive* factor) and is denoted from now on by R_n (respectively L_n).
- Prove that the graph of words Γ_n of a Sturmian sequence has the two possible forms given in figure 1.1.

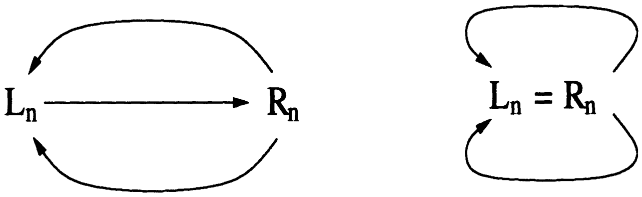

Figure 1.1:

- Deduce from the morphology of the graph of words Γ_n that every Sturmian sequence is uniformly recurrent. One can first prove that every factor of a Sturmian sequence is a subfactor of a factor of the form R_n and then deduce from the morphology of the graph Γ_n that R_n appears with bounded gaps.

Exercise 1.4.3

- Prove that if the sequence u is uniformly recurrent and non-constant, then the graph Γ_n has not edge of the form $U \to U$, for n large enough.

- Suppose that the sequence u is uniformly recurrent. Prove that if the graph of words Γ_{n+1} is Hamiltonian (i.e., there exists a closed oriented path passing exactly once through every vertex), then the graph Γ_n is Eulerian (there exists a closed path passing exactly once through every edge) and that $U^+ = U^-$, for every vertex of Γ_n. Is the converse true?

1.4.1 The Line Graph

The *line graph* $D(\Gamma_n)$ of the graph of words Γ_n is defined as follows: its vertices are the edges of Γ_n (i.e., the factors of length $n+1$); given two vertices u and v in $D(\Gamma_n)$, there is an edge from u to v if the end point of the edge labelled u in Γ_n is the origin of the edge labelled v. It is easily seen that the edges of the line graph correspond to words of length $n+2$ such that their prefix and their suffix of length $n+1$ are factors of the sequence u. The line graph of Γ_n is thus a subgraph of Γ_{n+1}.

Exercise 1.4.4 Study the evolution of the graph of words from Γ_n to Γ_{n+1} for a Sturmian sequence by using the line graph. (Distinguish between the two possible forms of the graph).

Remark 1.4.5 In [61] Rote uses the graph of words and the line graph for the study of sequences of complexity $p(n) = 2n$, for every n (see also [41]).