CHAPTER 1

Finite and Infinite Words

1.0. Introduction

The aim of this chapter is to provide an introduction to several concepts used elsewhere in the book. It fixes the general notation on words used elsewhere. It also introduces more specialized notions of general interest. For instance, the notion of a uniformly recurrent word used in several other chapters is introduced here.

We start with the notation concerning finite and infinite words. We also describe the Cantor space topology on the space of infinite words.

We provide a basic introduction to the theory of automata. It covers the determinisation algorithm, part of Kleene's theorem, syntactic monoids and basic facts about transducers. These concepts are illustrated on the classical combinatorial examples of the de Bruijn graph, and the Morse-Hedlund theorem.

We also consider the relationship with generating series, as a useful tool for the enumeration of words.

We introduce some basic concepts of symbolic dynamical systems, in relation with automata. We prove the equivalence between the notions of minimality and uniform recurrence. Entropy is considered, and we show how to compute it for a sofic system.

We also present a more specialized subject, namely unavoidable sets. This notion is easy to define but leads to interesting and significant results. In this sense, the last section of this chapter is an *avant-goût* of the rest of the book.

1.1. Semigroups

As usual, $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ denote the sets of nonnegative integers, integers, rational, real and complex numbers respectively. We denote by $\operatorname{Card} X$ the cardinality of the set X.

A *semigroup* is a set equipped with a binary associative operation. The set of words over a given alphabet has an obvious semigroup structure for the concatenation of words. A subsemigroup is a subset closed under the operation. A semigroup *morphism* from a semigroup S into a semigroup T is a mapping $f : S \to T$ such that $f(uv) = f(u)f(v)$ for all $u, v \in S$.

A *monoid* M is a semigroup with a neutral element, i.e. an element ε such that $m\varepsilon = \varepsilon m = m$ for all $m \in M$. A *submonoid* of a monoid M is a subset of M closed under the operation and containing the neutral element of M. A monoid *morphism* $f : M \to N$ is a semigroup morphism such that $f(\varepsilon_M) = \varepsilon_N$.

Given two semigroups S and T, the set $S \times T$ is canonically equipped with a semigroup operation by setting $(s,t)(s',t') = (ss',tt')$. The semigroup $S \times T$ is the *direct product* of S and T. A subset of $S \times T$ is called a *relations* between (or over) S and T.

Let X and Y be two subsets of a semigroup S. The *product* of X and Y is the set

$$XY = \{xy \mid x \in X, y \in Y\}$$

Given a set $X \subset S$, we denote by X^+ the subsemigroup generated by X, that is

$$X^+ = \{x_1 \cdots x_n \mid n \geq 1,\ x_i \in X\}$$

The operation $X \mapsto X^+$ is called the *plus operation.* This unary operation should not be confused with the (binary) disjoint union. If S is a monoid, we also define

$$X^* = X^+ \cup \{\varepsilon\} \tag{1.1.1}$$

which is the submonoid generated by X. The operation $X \mapsto X^*$ is called the *star operation.*

A subset X of a semigroup S is *rational* if it can be obtained from the finite subsets of S by a finite number of the operations of union, product, and plus.

In a monoid M, the family of rational sets is closed under the star operation because of Formula 1.1.1. Actually, this family is also generated by the operations of union, product and star, because $X^+ = XX^*$.

A special case deserves a mention. A rational subset of a product semigroup is called a *rational relation.*

EXAMPLE 1.1.1. For any set Q, the set $2^{Q\times Q}$ of binary relations on Q is a monoid for the composition of relations. The identity relation is the neutral element. The set of partial functions from Q into Q is a submonoid of $2^{Q\times Q}$. The set of permutations of Q is a submonoid of the latter.

EXAMPLE 1.1.2. For any finitely generated semigroup S, the set $D = \{(s,s) \mid s \in S\}$ is a rational relation called the *diagonal.*

1.2. Words

In this section, we first describe the (ordinary) finite words, before going to infinite one-sided and then two-sided words.

1.2.1. Finite words

We briefly introduce the basic terminology on words. A more detailed presentation can be found in Lothaire 1983. Let A be a set usually called the *alphabet*. We denote as usual by A^* the set of words over A and by ε the empty word. For a word w, we denote by $|w|$ the length of w. We use the notation $A^+ = A^* - \{\varepsilon\}$.

The set A^* is a monoid. Indeed, the concatenation of words is associative, and the empty word is a neutral element for concatenation. The set A^+ is usually called the *free semigroup* over A, while A^* is usually called the *free monoid*.

A word w is called a *factor* (resp. a *prefix*, resp. a *suffix*) of a word u if there exist words x, y such that $u = xwy$ (resp. $u = wy$, resp. $u = xw$). The factor (resp. the prefix, resp. the suffix) is *proper* if $xy \neq \varepsilon$ (resp. $y \neq \varepsilon$, resp. $x \neq \varepsilon$). The set of words over a finite alphabet A can be conveniently seen as a

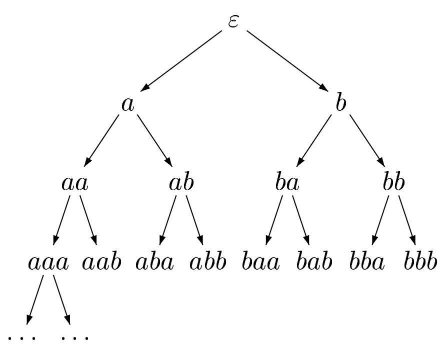

Figure 1.1. The tree of the free monoid.

tree. Figure 1.1 represents $\{a, b\}$ as a binary tree. The vertices are the elements of A^*. The root is the empty word ε. The sons of a node x are the words xa for $a \in A$. Every word x can also be viewed as the path from leading from the root to the node x. A word x is a prefix of a word y of it is an ancestor in the tree. We denote by $\mathrm{alph}\, w$ the set of letters having at least one occurrence in the word w.

The set of factors of a word x is denoted $F(x)$. We denote by $F(X)$ the set of factors of words in $X \subset A^*$. The *reversal* of a word $w = a_1 a_2 \cdots a_n$, where $a_1, \ldots, a_n$ are letters, is the word $\tilde{w} = a_n a_{n-1} \cdots a_1$. Similarly, for $X \subset A^*$, we denote $\widetilde{X} = \{\tilde{x} \mid x \in X\}$. A *palindrome word* is a word w such that $w = \tilde{w}$. If $|w|$ is even, then w is a palindrome if and only if $w = x\tilde{x}$ for some word x. Otherwise w is a palindrome if and only if $w = xa\tilde{x}$ for some word x and some letter a.

An integer $p \geq 1$ is a *period* of a word $w = a_1 a_2 \cdots a_n$ where $a_i \in A$ if $a_i = a_{i+p}$ for $i = 1, \ldots, n-p$. The smallest period of w is called *the* period of w.

A word $w \in A^+$ is *primitive* if $w = u^n$ for $u \in A^+$ implies $n = 1$.

Two words x, y are *conjugate* if there exist words u, v such that $x = uv$ and $y = vu$. Thus conjugate words are just cyclic shifts of one another. Conjugacy is thus an equivalence relation. The conjugacy class of a word of length n and period p has p elements if p divides n and has n elements otherwise. In particular, a primitive word of length n has n distinct conjugates.

There are three order relations frequently used on words. We give the definition of each of them.

The *prefix order* is the partial order defined by $x \leq y$ if x is a prefix of y.

Two other orders, the *radix order* and the are refinements of the prefix order which are defined for words over an ordered alphabet A. Both are total orders.

The *radix order* is defined by $x \leq y$ if $|x| < |y|$ or $|x| = |y|$ and $x = uax'$ and $y = uby'$ with a, b letters and $a \leq b$. If integers are represented in base k without

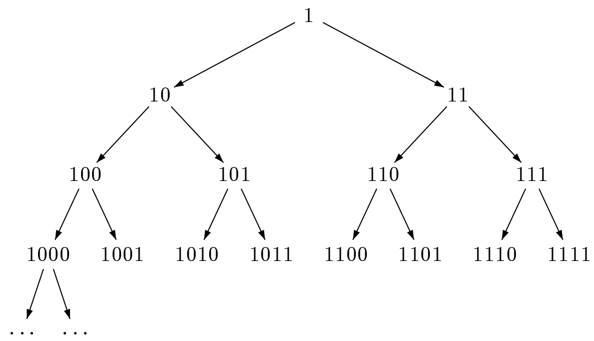

Figure 1.2. The tree of integers in binary notation.

leading zeroes, then the radix order on their representations corresponds to the natural ordering of the integers. If we allow leading zeroes, the same holds provided the words have the same length (which always can be achieved by padding).

For $k = 2$, the tree of words without leading zeroes is given in Figure 1.2. The radix order corresponds to the order in which the vertices are met in a breadth-first traversal. The index of a word in the radix order is equal to the number represented by the word in base 2.

The *lexicographic order*, also called *alphabetic order*, is defined as follows. Given two words x, y, we have $x < y$ if x is a proper prefix of y or if there exist factorizations $x = uax'$ and $y = uby'$ with a, b letters and $a < b$. This is the usual order in a dictionary. Note that $x < y$ in the radix order if $|x| < |y|$ or $|x| = |y|$ and $x < y$ in the lexicographic order.

A *Lyndon word* is a primitive word which is minimal for the lexicographic order in its conjugacy class. Thus, each nonempty word is a conjugate of a power of some Lyndon word.

The following result is known as Fine and Wilf's Theorem (see Proposition 1.3.5 in Lothaire 1983 and Theorem 8.1.4). As usual, $\gcd(n,m)$ denotes the greatest common divisor of n and m.

PROPOSITION 1.2.1 (Fine and Wilf). *Let x, y be words, let $n = |x|$, $m = |y|$, $d = \gcd(n,m)$. If two powers x^p and y^q of x and y have a common prefix of length at least $n + m - d$, then x and y are powers of the same word.*

For $X, Y \subset A^*$, we say that the union of X and Y is *unambiguous* if $X \cap Y = \emptyset$. In this case, we write $X + Y$ as a notation equivalent to $X \cup Y$.

The *product* of X and Y is, as in any semigroup, the set $XY = \{xy \mid x \in X, y \in Y\}$. The product is said to be *unambiguous* if for each $z \in XY$ there is exactly one pair $(x,y) \in X \times Y$ such that $z = xy$. In particular, we define

$$X^0 = \{\varepsilon\}, \quad X^{n+1} = X^n X \ (n \geq 0)\,.$$

Given a set $X \subset A^*$, the *star* of X is, as in any monoid, the set

$$X^* = \{x_1 \cdots x_n \mid n \geq 0,\ x_i \in X\} = \bigcup_{n \geq 0} X^n$$

PROPOSITION 1.2.2. *Any submonoid M of A^* has a unique minimal generating set*

$$X = (M - \varepsilon) - (M - \varepsilon)^2\,.$$

Proof. First, we show that X generates M. Let $w \in M - \varepsilon$. If $w \notin (M - \varepsilon)^2$, then $w \in X$. Otherwise, $w = w'w''$, with $w', w'' \in M - \varepsilon$. By induction, $w', w'' \in X^*$, and thus $w \in X^*$. This shows that $M \subset X^*$. The converse inclusion is clear since $X \subset M$.

Let Y be a set of generators of M. Clearly, $X \subset Y^*$. Since no word of X is a product of two nonempty words of M, we have actually $X \subset Y$. This shows that X is minimal. ∎

We say that a subset X of A^+ is a *code* if there is no relation among the elements of X, i.e.

$$x_1 \cdots x_n = y_1 \cdots y_m$$

with $x_1, \ldots, x_n, y_1, \ldots, y_m \in X$ implies $n = m$ and $x_i = y_i$ for $i = 1, \ldots, n$. In this case, one has

$$X^* = \sum_{n \geq 0} X^n \tag{1.2.1}$$

since the sets X^n are pairwise disjoint. We say that the star operation is *unambiguous* on the set X if X is a code.

A *prefix code* is a set X of words such that none of its elements is a prefix of another one. A prefix code is clearly a code.

A set of words is *prefix-closed* if it contains the prefixes of all its elements. A set of words is *factor-closed* or *factorial* if it contains the factors of all its elements.

A function $f : A^* \to B^*$ is a *morphism* (also called a *substitution*) if $f(xy) = f(x)f(y)$ for all $x, y \in A^*$. A morphism is uniquely determined by its value on the alphabet. A morphism is *literal* if the image of a letter is a letter. It is *nonerasing* if the image of a letter is always a nonempty word.

A morphism $f : A^* \to B^*$ is injective if and only if f is injective on A and if $f(A)$ is a code (see Proposition 6.1.3 for a proof).

1.2.2. Infinite words

We denote by $A^{\mathbb{N}}$ the set of (right) infinite words. It is the set of sequences of symbols in A indexed by nonnegative integers. We also denote

$$A^\infty = A^* \cup A^{\mathbb{N}}$$

the set of finite or infinite words.

For $x \in A^*$ and $y \in A^{\mathbb{N}}$, the product xy is well-defined. This defines a left action of the semigroup A^* on the set $A^{\mathbb{N}}$ since $x(yz) = (xy)z$ for all $x, y \in A^*$ and $z \in A^{\mathbb{N}}$.

A finite word w is a *factor* of an infinite word x if $x = uwy$. The set of factors of x is denoted by $F(x)$ and the set of factors of length n is denoted by $F_n(x)$. For a subset X of $A^{\mathbb{N}}$, we denote by $F(X)$ the set of factors of words in X. An infinite word $s \in A^{\mathbb{N}}$ is a *suffix* of $x \in A^{\mathbb{N}}$ if there is a word p in A^* such that $x = ps$. The suffix is called *proper* if $p \neq \varepsilon$.

The *lexicographic order* has a simple expression for infinite words over an ordered alphabet, since $x < y$ if and only if $x = uax'$, $y = uby'$ for some word $u \in A^*$, some letters $a, b \in A$ with $a < b$ and $x', y' \in A^{\mathbb{N}}$.

For a set $X \subset A^*$, we denote by X^ω the set of all $x = x_0x_1x_2\cdots$ with $x_i \in X - \varepsilon$. In particular, A^ω is the same as $A^{\mathbb{N}}$ and we use both notations indistinctly. We also use the notation

$$X^\infty = X^* \cup X^\omega$$

The *shift* function is the function $\sigma : A^{\mathbb{N}} \to A^{\mathbb{N}}$ defined by $\sigma(x_0x_1\cdots) = x_1x_2\cdots$.

Consider a finite or infinite word x over the alphabet A. The *complexity function* of x is the function that counts, for each integer $n \geq 0$, the number $P(x, n)$ of factors of length n in x:

$$P(x, n) = \mathrm{Card}(F_n(x)).$$

Clearly, $P(x, 0) = 1$ and $P(x, 1)$ is the number of letters appearing in x. If x is infinite, every factor can be extended to the right, whence $P(x, n) \leq P(x, n+1)$. Moreover,

$$P(x, n + m) \leq P(x, n)P(x, m)$$

since indeed $F_{n+m}(x) \subset F_n(x)F_m(x)$.

The set $A^{\mathbb{N}}$ is equipped with a distance defined as follows. For $x, y \in A^\omega$, we have $d(x, y) = 2^{-n}$ with

$$n = \min\{k \geq 0 \mid x_k \neq y_k\}$$

and the convention that $n = \infty$ and thus $d(x, y) = 0$ if $x = y$.

With respect to this distance, the set A^ω becomes a topological space, often called the *Cantor space*. A sequence $x^{(n)}$ of infinite words converges to y in this topology

$$y = \lim_{n\to\infty} x^{(n)}$$

if, for each index $i \in \mathbb{N}$, one has $x_i^{(n)} = y_i$ for large enough n.

For example, the sequence $x^{(n)} = a^n b^\omega$ converges to $y = a^\omega$.

A consequence of the definitions is that if a word w is a factor of $\lim_{n\to\infty} x^{(n)}$, then w is a factor of all but a finite number of the $x^{(n)}$.

A set of infinite words is (topologically) *closed* if it contains the limits of each convergent sequence of its elements.

The open sets are the complements of the closed sets. They happen to be also the sets of the form XA^ω for $X \subset A^*$ (Problem 1.2.3).

The following result is known as König's Lemma.

PROPOSITION 1.2.3 (König's Lemma). *If X is an infinite prefix-closed set of words over a finite alphabet A, there is an infinite word x having all its prefixes in X.*

Proof. There is a letter a_1 which is a prefix of an infinite number of elements of X. Similarly, there is a letter a_2 such that a_1a_2 is a prefix of an infinite number of elements of X. Continuing this way, one obtains an infinite word $a_1a_2\cdots$ having all its prefixes in X. ■

A set of words is *compact* if any sequence $x^{(n)}$ of infinite words of the set has a convergent subsequence.

PROPOSITION 1.2.4. *For a finite alphabet A, the space A^ω is compact.*

Proof. Consider a sequence $x^{(n)}$ of infinite words. Let X be the set of all prefixes of the words $x^{(n)}$. By König's Lemma, there is an infinite word x which has all its prefixes in X.

For each $i > 0$, let u_i be the prefix of length i of x. Since $u_i \in X$, there is an integer n_i such that u_i is a prefix of $x^{(n_i)}$. Clearly, the sequence $x^{(n_i)}$ converges to x. ■

A closed subset of $A^{\mathbb{N}}$ is also compact, as one may check, provided A is finite as above.

If $X_1 \supset X_2 \supset \cdots \supset X_n \supset \cdots$ is a decreasing sequence of nonempty closed subsets of $A^{\mathbb{N}}$ where A is finite, then their intersection X is nonempty. Consider indeed, for each n, a word x_n in X_n. By compactness, there is a subsequence converging to an infinite word x. Since the X_n are closed, x is in all of them, thus x is in X.

For any set $X \subset A^\infty$, we denote

$$X_* = X \cap A^*,\ X_\omega = X \cap A^\omega$$

A *binoid* over A is a subset M of A^∞ such that

$$M_* M \subset M, \quad (M_*)^\omega \subset M$$

Observe that in particular M_* is a submonoid of A^*. It is convenient to denote, for every set $X \subset A^\infty$

$$X^\infty = X_*^\infty \cup (X_*)^* X_\omega = X_*^* \cup X_*^\omega \cup (X_*)^* X_\omega$$

With this notation, a set $M \subset A^\infty$ is a binoid if and only if $M^\infty = M$. For any set $X \subset A^\infty$, the set X^∞ is a binoid called the binoid generated by X. It is also the intersection of all binoids containing X. A binoid M such that $M = X^\infty$ for some $X \subset A^*$ is called *finitary*.

EXAMPLE 1.2.5. Set $X = a \cup b^\omega$. Then $X^\infty = a^\infty \cup a^* b^\omega$. This binoid is not finitary.

PROPOSITION 1.2.6. *Any binoid M has a unique minimal generating set $X = (M - \varepsilon) - (M_* - \varepsilon)(M - \varepsilon)$.*

Proof. Observe first that X_* is the minimal generating set of M_* by Proposition 1.2.2.

To prove that X also generates M_ω, let $y \in M_\omega$. If $y \notin (M_* - \varepsilon)M$, then $y \in X$. Otherwise, $y = x_1 y_1$ for some $x_1 \in X_* - \varepsilon$ and $y_1 \in M_\omega$. Again, if $y_1 \notin (M_* - \varepsilon)M$, then $y_1 \in X$ and thus $y \in X^\infty$. Otherwise, $y_1 = x_2 y_2$ for some $x_2 \in X_* - \varepsilon$ and $y_2 \in M_\omega$. Thus $y = x_1 x_2 y_2$ is in X^∞. Continuing in this way, either we eventually obtain $y = x_1 x_2 \cdots x_n y_n \in X^\infty$, or $y = x_1 x_2 \cdots x_n \cdots \in (X_*)^\omega$. This proves that X generates M.

Let Y be a generating set of M. We know that $X_* \subset Y_*$. Let $x \in X_\omega$. Since $Y^\infty = Y_* Y \cup Y$, one has $x = yy'$ for some $y \in Y_* \cup \varepsilon$ and $y' \in Y$. By the definition of X, we have $y = \varepsilon$ and thus $x \in Y$. Thus $X \subset Y$. ∎

COROLLARY 1.2.7. *The minimal generating set of a finitary binoid M is the minimal generating set of the monoid M_*.* ∎

A word $x \in A^\omega$ is *periodic* if it is of the form $x = z^\omega$ for some $z \in A^+$. A word $x \in A^\omega$ is *eventually periodic* or *ultimately periodic* if it is of the form $x = yz^\omega$ for some $y, z \in A^+$. A word $x \in A^\omega$ is *aperiodic* if it is not eventually periodic.

A word x is periodic if and only if it is a proper suffix of itself or equivalently if $x = \sigma^p(x)$ for some $p > 0$.

A nonerasing morphism $f : A^* \to B^*$ defines a function, also called a morphism, from $A^\mathbb{N}$ to $B^\mathbb{N}$ by $f(a_0 a_1 \cdots a_n \cdots) = f(a_0) f(a_1) \cdots f(a_n) \cdots$.

A sequence $(u_n)_{n \geq 0}$ of finite words over an alphabet A *converges* to an infinite word x if every prefix of x is a prefix of all but a finite number of the words u_n. This word x is unique and is denoted by

$$x = \lim_{n \to \infty} u_n$$

This definition can be related to the topology considered above (Problem 1.2.4).

As an example, the sequence $a^n b^n$ converges to a^ω. An important special case arises when every u_n is a prefix of u_{n+1}. Then the sequence converges, provided the lengths of the words u_n is unbounded. A special case of this is described in the following statement.

PROPOSITION 1.2.8. *Let h be a nonerasing morphism from A^* into itself, and let a be a letter such that $h(a) = as$ for some nonempty word s. Set for $n \geq 0$,*

$$u_n = h^n(a), \quad v_n = h^n(s)$$

Then

1. *$u_{n+1} = u_n v_n$, and in particular, u_n is a prefix of u_{n+1} for all $n \geq 0$.*
2. *$u_{n+1} = a v_0 v_1 v_2 \cdots v_n$.*
3. *The infinite word*
$$x = ash(s)h^2(s) \cdots h^n(s) \cdots \tag{1.2.2}$$
is the limit of the words u_n and x is a fixed point of h. Moreover, it is the unique fixed point of h starting with the letter a.

Proof. (1) $u_{n+1} = h^{n+1}(a) = h^n(h(a)) = h^n(as) = u_n v_n$.

(2) holds for $n = 0$, and by induction $u_{n+1} = u_n v_n = a v_0 v_1 v_2 \cdots v_{n-1} v_n$.

(3) It is clear that x is the limit. Moreover,

$$h(x) = h(a)h(s)h^2(s) \cdots = x\,. \qquad \blacksquare$$

The word x of the proposition is also denoted by

$$x = h^\omega(a)$$

A word x obtained in this way is a *morphic* word.

We now develop two examples which will occur throughout the book.

EXAMPLE 1.2.9. The *Thue-Morse* infinite word t over the alphabet $A = \{0, 1\}$ is defined as the limit

$$t = \lim_{n \to \infty} u_n$$

where the sequences of words $(u_n)_{n \geq 0}$ and $(v_n)_{n \geq 0}$ are defined by

$$\begin{array}{lll} u_0 = 0 & v_0 = 1 & \\ u_{n+1} = u_n v_n & v_{n+1} = v_n u_n & n \geq 0 \end{array}$$

The first letters of t are

$$t = 0110100110010110100101100110100110010110011010010110100110\cdots$$

The word t is actually a morphic word since $u_n = \mu^n(0)$ where μ is the morphism

$$\mu : \begin{array}{l} 0 \mapsto 01 \\ 1 \mapsto 10 \end{array}$$

The decomposition of x corresponding to Equation (1.2.2) is

$$t = 0\ 1\ 10\ 1001\ 10010110\ 1001011001101001 \cdots$$

EXAMPLE 1.2.10. Let $A = \{a, b\}$. Let φ be the morphism defined by

$$\varphi : \begin{array}{l} a \mapsto ab \\ b \mapsto a \end{array}$$

The *Fibonacci* word is the infinite word $f = \varphi^\omega(a)$. The first letters of f are

$$f = abaababaabaababaababaabaababaabaab\cdots$$

One has also $f = \lim_{n\to\infty} f_n$ where the sequence of words $f_n = \varphi^n(a)$ can also be defined by

$$f_0 = a,\ f_1 = ab, \quad f_{n+2} = f_{n+1}f_n$$

The sequence of lengths of the words f_n is the traditional sequence of Fibonacci numbers (see Example 1.4.2).

1.2.3. Two-sided infinite words

We denote by $A^{\mathbb{Z}}$ the set of two-sided infinite words on A, which is the set of sequences of symbols of A indexed by integers. For $x \in A^{\mathbb{Z}}$, the *shift* function is the function $\sigma : A^{\mathbb{Z}} \to A^{\mathbb{Z}}$ defined by $\sigma(x) = y$ with $y_n = x_{n-1}$ for $n \in \mathbb{Z}$. Observe that, contrary to the one-sided case, the shift is a one-to-one transformation on $A^{\mathbb{Z}}$. The *period* of $x \in A^{\mathbb{Z}}$ is the greatest common divisor of the integers n such that $\sigma^n(x) = x$. It is an integer or ∞. The terminology used for words or one-sided infinite words carries over. In particular, $F(x)$ denotes the set of (finite) factors of a word $x \in A^{\mathbb{Z}}$, and if $X \subset A^{\mathbb{Z}}$, we denote by $F(X)$ the set of factors of words in X.

The set $A^{\mathbb{Z}}$ is also equipped with a distance defined in a way quite analogous to the distance of A^ω. We define $d(x, y) = 2^{-n}$ where

$$n = \min\{k \geq 0 \mid x_k \neq y_k \text{ or } x_{-k} \neq y_{-k}\}$$

with the convention that $d(x, y) = 0$ when $x = y$.

This distance defines a topology on the set $A^{\mathbb{Z}}$ as in the one-sided case. There exists a two-sided version of König's lemma.

PROPOSITION 1.2.11. *For any infinite factorial set X of words over a finite alphabet, there exists a two-sided infinite word having all its factors in X.*

Proof. It is similar to the one-sided case. ∎

Again, the space $A^{\mathbb{Z}}$ is compact when A is finite.

For a set $X \subset A^+$, we denote by X^ζ the closure under the shift of the set of all $w = (a_n)_{n\in\mathbb{Z}} \in A^{\mathbb{Z}}$ such that

$$\cdots a_{-1}a_0 = \cdots x_{-1}x_0, \quad a_1a_2\cdots = x_1x_2\cdots$$

with $x_n \in X$ for all $n \in \mathbb{Z}$. Observe that A^ζ coincides with $A^{\mathbb{Z}}$. For a single word $x = a_1a_2\ldots a_n \in A^+$, the set x^ζ is composed of the sequences of the form

$\cdots a_n(a_1a_2\cdots a_n)a_1a_2\cdots$. Each word in x^ζ has a period which divides n. The period is n if and only if the word x is primitive.

A literal morphism $f : A^* \to B^*$ defines a function, also called a literal morphism, from $A^{\mathbb{Z}}$ to $B^{\mathbb{Z}}$. It maps the word $x \in A^{\mathbb{Z}}$ to the word $y = f(x) \in B^{\mathbb{Z}}$ defined by $y_i = f(x_i)$.

1.3. Automata

1.3.1. Definitions

An *automaton* over the alphabet A is composed of a set Q of *states*, a set $E \subset Q \times A \times Q$ of *edges* or *transitions* and two sets $I, T \subset Q$ of *initial* and *terminal* states. For an edge $e = (p, a, q)$, the state p is the *origin*, a is the *label*, and q is the *end*.

The automaton is often denoted $\mathcal{A} = (Q, E, I, T)$, or also (Q, I, T) when E is understood, or even $\mathcal{A} = (Q, E)$ if $Q = I = T$.

A *path* in the automaton $\mathcal{A}$ is a sequence

$$(p_0, a_1, p_1), (p_1, a_2, p_2), \ldots, (p_{n-1}, a_n, p_n)$$

of consecutive edges. Its label is the word $x = a_1a_2\cdots a_n$. The path *starts* at p_0 and *ends* at p_n. The path is often denoted

$$p_0 \xrightarrow{x} p_n$$

A path is *successful* if it starts in an initial state and ends in a terminal state. The set *recognized* by the automaton is the set of labels of its successful paths.

A state p is *accessible* if there is a path starting in an initial state and ending in p. It is *coaccessible* if there is a path starting in p and ending in a terminal state. An automaton is *trim* if every state is accessible and coaccessible.

An automaton is *unambiguous* if, for each pair of states p, q, and for each word w, there is at most one path from p to q labeled with w.

An automaton is *deterministic* if, for each state p and each letter a, there is at most one edge which starts at p and is labeled by a. This state is denoted by $p \cdot a$. Clearly, a deterministic automaton is unambiguous.

Given an automaton $\mathcal{A}$ and a state q of $\mathcal{A}$, the set of *first returns* to q is the set of labels of paths from q to q which do not pass another time through q. If $\mathcal{A}$ is unambiguous, then the set of first returns to a state q is a code. If $\mathcal{A}$ is deterministic, it is a prefix code.

EXAMPLE 1.3.1. Let $\mathcal{A}$ be the automaton given in Figure 1.3. The set of first returns to state 1 is the prefix code $X = \{b, ab\}$ of Example 1.4.2.

An automaton is *finite* if its set of states is finite. Since the alphabet is usually assumed to be finite, this means that the set of edges is finite.

A set of words X over A is *recognizable* if it can be recognized by a finite automaton.

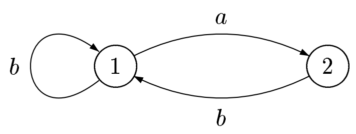

Figure 1.3. Golden mean automaton.

Proposition 1.3.2. *Every recognizable set can be recognized by a finite trim deterministic automaton having a unique initial state.*

Proof. Let $\mathcal{A} = (Q, E, I, T)$ be a finite automaton over A recognizing a set X. Let $\mathcal{B} = (\mathcal{R}, F, \{I\}, \mathcal{T})$ be the automaton defined as follows. Its states are the subsets

$$Q(u) = \{q \in Q \mid i \xrightarrow{u} q \text{ for some } i \in I\}$$

for all u in A^*. Since Q is finite, there is a finite number of subsets $Q(u)$. The edges of $\mathcal{B}$ are all triples

$$(Q(u), a, Q(ua)) .$$

The set of terminal states is

$$\mathcal{T} = \{U \in \mathcal{R} \mid U \cap T \neq \emptyset\} .$$

It is easy to verify that $\mathcal{B}$ is trim, deterministic, and recognizes X. ∎

Let X be a subset of A^*. For $w \in A^*$, we define the *left quotient*

$$w^{-1}X = \{u \in A^* \mid wu \in X\}, \quad Xw^{-1} = \{u \in A^* \mid uw \in X\}$$

The following relations hold for words v, w and a letter a

$$(vw)^{-1}X = w^{-1}(v^{-1}X), \quad a^{-1}(XY) = (a^{-1}X)Y \cup (X \cap \varepsilon)a^{-1}Y$$

The notation is extended as usual to sets by

$$X^{-1}Y = \bigcup_{x \in X} x^{-1}Y$$

To every set $X \subset A^*$ is associated a deterministic automaton $\mathcal{A}(X)$ as follows. Its set of states $Q(X)$ is

$$Q(X) = \{w^{-1}X \mid w \in A^*\}$$

Its initial state is X, its set of final states is $T(X) = \{S \in Q(X) \mid \varepsilon \in S\}$. Its transitions are defined for $S \in Q(X)$ and $a \in A$ by $S \cdot a = a^{-1}S$. The automaton $\mathcal{A}(X)$ is called the *minimal automaton* of X. It recognizes X because indeed

$$X \cdot w \in T(X) \iff \varepsilon \in w^{-1}X \iff x \in X .$$

An equivalence relation on A^* is *right regular* if $u \equiv v$ implies $ux \equiv vx$ for all $u, v, x \in A^*$.

PROPOSITION 1.3.3. *Let X be a subset of A^*. The following conditions are equivalent.*

(i) *X is recognizable,*
(ii) *the automaton $\mathcal{A}(X)$ is finite,*
(iii) *there exists a right regular equivalence relation of finite index on A^* for which X is a union of equivalence classes.*

Proof $(i) \Leftrightarrow (ii)$ First $\mathcal{A}(X)$ recognizes X. Conversely, let $\mathcal{A} = (Q, E, i, T)$ be a finite trim deterministic automaton recognizing X. For each state $q \in Q$, let $X_q = \{u \in A^* \mid q \cdot u \in T\}$. The set $Q(X)$ is equal to the set of the X_q for $q \in Q$. Indeed, for $w \in A^*$, we have $w^{-1}X = X_{i \cdot w}$.

$(ii) \Leftrightarrow (iii)$ The equivalence relation defined by $u \equiv v$ if and only if $u^{-1}X = v^{-1}X$ satisfies the conditions. Conversely, if (iii) is satisfied, the set $Q(X)$ is finite. Indeed, $u \equiv v$ implies $u^{-1}X = v^{-1}X$, and thus the elements of $Q(X)$ are unions of equivalence classes. ■

It can be shown (Problem 1.3.1) that $\mathcal{A}(X)$ is the (unique) smallest deterministic automaton recognizing X,

PROPOSITION 1.3.4. *A set of words X over the alphabet A is recognizable if and only if there exists a morphism $f : A^* \to S$ from A^* into a finite semigroup S such that $X = f^{-1}(f(X))$.*

Proof. Let $\mathcal{A} = (Q, E, I, T)$ be a finite automaton over A recognizing a set X. The set S of all binary relations over Q is a semigroup for the composition of relations. Define for each word w the relation $f(w)$ by

$$f(w) = \{(p, q) \in Q \times Q \mid p \xrightarrow{w} q\}$$

It is easy to check that f is a semigroup morphism and that $X = f^{-1}(U)$, where $U = \{s \in S \mid s \cap I \times T \neq \emptyset\}$. Thus $X = f^{-1}(f(X))$.

Conversely, let $f : A^* \to S$ be a semigroup morphism satisfying the conditions of the statement. Define an automaton $\mathcal{A} = (S, E, f(\varepsilon), f(X))$ where $E = \{p, a, q) \in S \times A \times S \mid pf(a) = q\}$. It can be verified that this automaton recognizes the set X. ■

A semigroup S is said to *recognize* a set X if there exists a morphism $f : A^* \to S$ such that $X = f^{-1}(f(X))$.

Let X be a set of words. The set of contexts of a word w is the set

$$C(w) = \{(x, y) \in A^* \times A^* \mid xwy \in X\}$$

The *syntactic equivalence* of X is defined by $u \equiv v$ if and only if $C(u) = C(v)$. The syntactic equivalence is compatible with concatenation of words, and thus the quotient $A^*/\equiv$ is a semigroup. It is called the *syntactic semigroup* of X. It can be shown (Problem 1.3.2) that the syntactic semigroup of X is the smallest semigroup recognizing X. In particular, X is recognizable if and only if its syntactic semigroup is finite.

A subset X of A^* is *rational* if it can be obtained from the finite subsets of A by a finite number of the operations of union, product, and star.

A subset X of A^* is *unambiguously rational* if it can be obtained from the finite subsets of A by a finite number of the operations of unambiguous union, product, and star.

A well-known theorem of Kleene (see Notes) asserts that, over a finite alphabet, a set is rational if and only if it is recognizable. We prove here one direction of the equivalence in a slightly stronger form.

PROPOSITION 1.3.5. *Any recognizable set is unambiguously rational.*

Proof. Let $\mathcal{A} = (Q, i, T)$ be a finite deterministic automaton recognizing a set X with a unique initial state. For $p, q \in Q$, let $X_{p,q}$ be the set of nonempty words recognized by the automaton (Q, p, q). Then

$$X = \sum_{t \in T} X_{i,t} + \Delta_{i,T}$$

where $\Delta_{i,T} = \{\varepsilon\}$ if $i \in T$, and $\Delta_{i,T} = \emptyset$ otherwise. It is therefore enough to prove that each $X_{i,j}$ is unambiguously rational.

For $P \subset Q$ and $p, q \in Q$, we denote by $X_{p,P,q}$ the set of nonempty words that are labels of paths from p to q and which only pass through states in P (except perhaps at the beginning and the end). We prove that each $X_{p,P,q}$ is unambiguously rational by induction on the size of P. If $P = \emptyset$, then $X_{p,P,q}$ is a subset of the alphabet A. Set $Y_{p,q} = X_{p,P,q}$ and $Z_{p,q} = X_{p,P\cup\{r\},q}$ for some state $r \notin P$. We have the formula

$$Z_{p,q} = Y_{p,q} + Y_{p,r}(Y_{r,r}^*)Y_{r,q}$$

The operations used in this formula are unambiguous. This proves the property by induction. ■

1.3.2. Automata on infinite words

In this section, we introduce acceptance of infinite words by finite automata in Büchi's sense.

Let $\mathcal{A} = (Q, E, I, T)$ be a finite automaton over A. An infinite path is an infinite sequence

$$(p_0, a_0, p_1), (p_1, a_1, p_2), \ldots$$

of consecutive edges. Its label is the infinite word $x = a_0 a_1 \cdots$. The path is *successful* if $p_0 \in I$ and if $p_n \in T$ for infinitely many indices n.

The set of infinite words *recognized* by the automaton is the set of labels of successful infinite paths. An automaton used to recognize infinite words in this sense is frequently called a *Büchi automaton*.

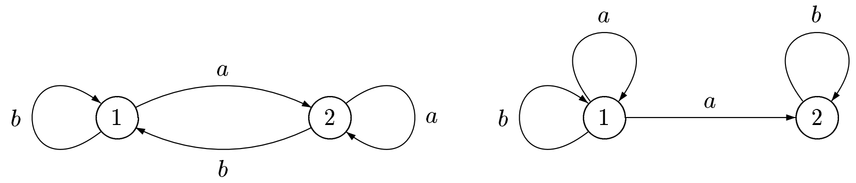

Figure 1.4. Büchi automata.

EXAMPLE 1.3.6. Consider first the automaton given in the left part of Figure 1.4 with $I = \{1\}$ and $T = \{2\}$. It recognizes the set X of words having an infinite number of occurrences of a. The second automaton, given on the right, again with $I = \{1\}$ and $T = \{2\}$, recognizes the complement of X, namely the set of words with a finite number of occurrences of a.

Observe that the complement of X is *not* obtained by simply complementing the set of terminal states in the first automaton.

1.3.3. Transducers

A *transducer* over the monoid $A^* \times B^*$ is composed of a set Q of *states*, a set $E \subset Q \times A^* \times B^* \times Q$ of *edges* and two sets $I, T \subset Q$ of *initial* and *terminal* states. For an edge $e = (p, x, y, q)$, the state p is the *origin*, x is the *input label*, y is the *output label*, and q is the *end*. Thus, a transducer is the same object as an automaton, except that the labels of the edges are pairs of words instead of letters.

A transducer is often denoted $\mathcal{A} = (Q, E, I, T)$, or also (Q, I, T) when E is understood, or even $\mathcal{A} = (Q, E)$ if $Q = I = T$.

A *path* in the transducer $\mathcal{A}$ is a sequence

$$(p_0, x_1, y_1, p_1), (p_1, x_2, y_2, p_2, \ldots, (p_{n-1}, x_n, y_n, p_n)$$

of consecutive edges. Its input label is the word $x = x_1 x_2 \cdots x_n$, its output label is the word $y = y_1 y_2 \cdots y_n$. The path *starts* at p_0 and *ends* at p_n. The path is often denoted

$$p_0 \xrightarrow{x/y} p_n$$

A path is *successful* if it starts in an initial state and ends in a terminal state. The set *recognized* by the transducer is the set of labels of its successful paths, which is actually a relation $R \subset A^* \times B^*$. The *function computed* by the transducer is the function f from A^* into the set of subsets of B^* associated to the relation R:

$$f(x) = \{y \in B^* \mid (x, y) \in R\}$$

Thus a transducer can be seen as a machine computing nondeterministically output words from input words.

A transducer is *finite* if its set of states is finite. It can be shown that a subset of $A^* \times B^*$ is a rational relation if and only if it is the set recognized by a finite transducer.

EXAMPLE 1.3.7. The automaton of Figure 1.5 computes the identity function on $\{a, b\}^*$.

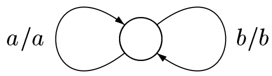

Figure 1.5. The transducer for the diagonal over $\{a, b\} \times \{a, b\}$.

EXAMPLE 1.3.8. The subset R of $a^* \times \{b, c\}^*$ defined as

$$R = (a^2, b^2)^* \cup (a^2, c^2)^*(a, c)$$

is a rational relation. Its elements have the form (a^n, d^n), with $d = b$ if n is even, and $d = c$ otherwise. The automaton of Figure 1.6 recognizes the relation R.

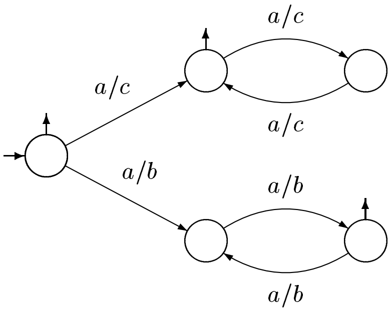

Figure 1.6. A transducer for the relation $(a^2, b^2)^* \cup (a^2, c^2)^*(a, c)$.

Let $\mathcal{A}$ be a transducer such that its edges are labeled by elements of $A \times B^*$. The *underlying input automaton* of $\mathcal{A}$ is obtained by omitting the output label of each edge.

The transducer $\mathcal{A}$ is *sequential* if the following conditions are satisfied.

(i) it has a unique initial state,

(ii) the underlying input automaton is deterministic,

(iii) every state is final.

These conditions ensure that for each word $x \in A^*$, there is at most one word $y \in B^*$ such that (x, y) is recognized by $\mathcal{A}$. Thus, the function computed by $\mathcal{A}$ is a partial function from A^* into B^*.

Example 1.3.9. The transducer of Example 1.3.7 is sequential. On the contrary, the transducer of Example 1.3.8 is not sequential. Actually, the function computed by this transducer is not computable by a sequential transducer. Indeed, one may verify that if f is a function computable by some sequential transducer, and if $f(xy)$ is defined, then $f(x)$ is a prefix of $f(xy)$.

A function f is *left sequential* (or *sequential* for short) if there is a sequential transducer which computes f. A function f is *right sequential* if the function $\tilde{f}$ defined by $y = \tilde{f}(x)$ if $\tilde{y} = f(\tilde{x})$ is left sequential. Thus, a right sequential function is a function computed by a sequential transducer operating from right to left.

A *subsequential transducer* $(\mathcal{A}, \omega)$ over $A^* \times B^*$ is a pair composed of a sequential transducer $\mathcal{A}$ over $A^* \times B^*$ with set of states Q, and of a function $\omega : Q \to B^*$. The function f computed by $(\mathcal{A}, \omega)$ is defined as follows. Let x be a word in A^*. The value $f(x)$ is defined if and only if there is a path $i \xrightarrow{x,y} q$ in $\mathcal{A}$ with input label x and starting in the initial state i. In this case, $f(x) = y\omega(q)$. Thus, the function ω is used to append a word to the output at the end of the computation.

A function computed by a subsequential transducer is a *left subsequential function*. *Right* subsequential functions are obtained by reversal. The following example shows that the successor of an integer in base 2 is a right subsequential function.

Example 1.3.10. Let $A = \{0, 1\}$. For every word x in A^*, let $f(x)$ be the binary expansion of the successor of the integer represented by x in base 2, (most significant bit first). The pair $(\mathcal{A}, \omega)$, with $\mathcal{A}$ given in Figure 1.7, and

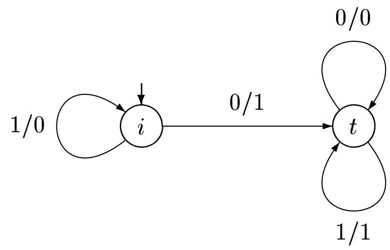

Figure 1.7. A transducer adding 1 in binary.

$\omega(i) = 1$, $\omega(t) = \varepsilon$ computes the function $\tilde{f}$. Since $\mathcal{A}$ is sequential, f is right subsequential.

Any sequential transducer $\mathcal{A}$ over $A^* \times B^*$ defines actually a function from A^∞ to B^∞. Indeed, for each x in A^∞, there is at most one y in B^∞ such that there is an infinite path starting in the initial state and labeled (x, y). A function $f : A^\infty \to B^\infty$ is *left sequential* if it can be computed by a sequential transducer.

1.3.4. Factor graphs

We now present a special family of automata which allows one to identify the occurrences of elements of a set X of words as factors of a given word.

For each $n \geq 1$, we define the de Bruijn graph of order n as the following labelled graph. The set of vertices is A^n and the set of edges is

$$E = \{(bs, a, sa) \mid a, b \in A, s \in A^{n-1}\}$$

A word x is the label of a path from u to v if and only if v is the suffix of length n of ux. The de Bruijn graph of order 2 is given in Figure 1.8. For each two-sided infinite word x , there is a unique infinite path labelled by x. The set of vertices occurring in the path is the set $F_n(x)$ of factors of length n of x. The set of edges occurring in the path corresponds to the set $F_{n+1}(x)$.

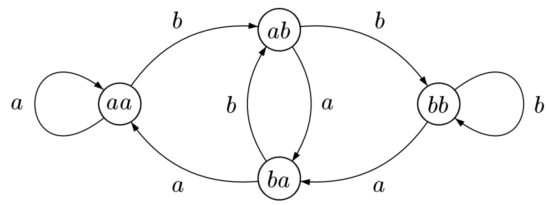

Figure 1.8. De Bruijn graph of order $n = 2$.

More generally, consider an infinite word x. The *factor graph* $G_n(x)$ of order n is the labelled graph with vertex set $F_n(x)$ and edge set

$$E = \{(bs, a, sa) \mid a, b \in A, bsa \in F_{n+1}(x)\}$$

A word y is the label of a path from u to v in $G_n(x)$ if and only if $F_{n+1}(uy) \subset F_{n+1}(x)$ and v is the suffix of length n of uy. The de Bruijn graph is a particular case of a factor graph corresponding to a word x such that $F_{n+1}(x) = A^{n+1}$.

A factor p is called *conservative* if there is exactly one edge leaving p, it is *right special* otherwise.

Example 1.3.11. Let t be the infinite Thue-Morse word (see Example 1.2.9). It is easily checked that 000 and 111 are not factors of t. The factor graph $G_3(t)$ is given in Figure 1.9.

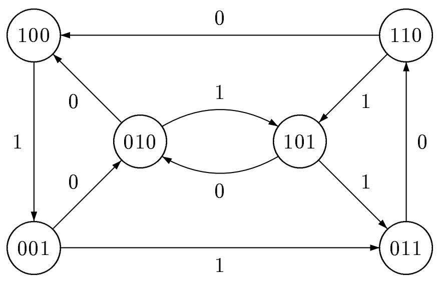

Figure 1.9. Factor graph of order 3 for the Thue-Morse word.

PROPOSITION 1.3.12. *A word y is the label of an infinite path starting at vertex p in $G_n(x)$ if and only if $F_{n+1}(py) \subset F(x)$.* ∎

Recall that the complexity function $P(x, n)$ of an infinite word x is defined by $P(x, n) = \mathrm{Card}(F_n(x))$. The following result is a gap theorem. It shows that the complexity is either bounded or more than linear.

THEOREM 1.3.13. *Let x be an infinite word. The following are equivalent:*
(i) *x is eventually periodic,*
(ii) *$P(x, n) = P(x, n + 1)$ for some n,*
(iii) *$P(x, n) < n + k - 1$ for some $n \geq 1$, where k is the number of letters appearing in x.*
(iv) *$P(x, n)$ is bounded.*

Proof. (i)⇒(iv). Observe that if $x = uy^\omega$, then $P(x, n) \leq |uy|$.

(iv)⇒(iii) is clear.

(iii)⇒(ii). Assume $P(x, m - 1) < P(x, m)$ for $m = 0, \ldots, n$, then $P(x, n) \geq n - 1 + P(x, 1) = n - 1 + k$, a contradiction since $P(x, n)$ is a non decreasing function of n..

(ii)⇒(i). Consider the factor graph $G_n(x)$. Since every factor of length n is a prefix of a factor of length $n + 1$, there is at least one edge starting at each vertex. Since $P(x, n) = P(x, n + 1)$, there is exactly one edge leaving each vertex. This implies that the strongly connected components of the graph are simple circuits. Thus any infinite path will loop through a fixed circuit after a while, and consequently its label is eventually periodic. Since x is the label of a path, the claim is proved. ∎

There is a slightly stronger result that is sometimes useful to show that an infinite word is eventually periodic.

PROPOSITION 1.3.14. *Let x be an infinite word, let $n \geq 1$ and let c be the number of conservative factors of length n in x. If x has a factor of length $n + c$ whose factors of length n are all conservative, then x is eventually periodic.*

Proof. Let $w = a_1 a_2 \cdots a_{n+c}$ be a factor of length $n + c$ whose factors of length n are all conservative, and set $p_i = a_i \cdots a_{i+n-1}$ for $i = 0, \ldots, c$. In the factor graph $G_n(x)$, the path $\pi = (p_0, \ldots, p_c)$ is part of the path of x. Since there are only c conservative vertices, the path π contains a circuit, and since each vertex p_i has a unique outgoing edge, the path of x must stay in this circuit indefinitely. Thus x is eventually periodic. ■

1.4. Generating series

For any set $X \subset A^*$, the *generating function* or *generating series* of X is the formal series

$$f_X(z) = \sum_{n \geq 0} u_n z^n$$

where

$$u_n = \mathrm{Card}(X \cap A^n)$$

PROPOSITION 1.4.1. *If X is a code, then $f_{X^*} = 1/(1 - f_X)$.*

Proof. If X is a code, every word in X^* has a unique decomposition as a product of words in X. This implies that

$$f_{X^n} = (f_X)^n$$

and thus, by Equation 1.2.1,

$$f_{X^*} = 1 + f_X + \cdots + f_{X^n} + \cdots = 1/(1 - f_X)\,.$$ ■

EXAMPLE 1.4.2. The set $X = \{b, ab\}$ is a prefix code. The series f_{X^*} is

$$f_{X^*}(z) = \frac{1}{1 - z - z^2}$$

Let $(F_n)_{n \geq 0}$ be the sequence of Fibonacci numbers defined by $F_0 = 0$, $F_1 = 1$, and $F_{n+2} = F_{n+1} + F_n$. It follows from the recurrence relation that

$$\frac{z}{1 - z - z^2} = \sum_{n \geq 0} F_n z^n$$

Consequently, $f_{X^*}(z) = \sum_{n \geq 0} F_{n+1} z^n$. It can also be proved by a combinatorial argument that the number of words of length n in X^* is F_{n+1}.

For any set $X \subset A^*$, we denote by ρ_X the radius of convergence of f_X. Since the coefficients u_n of f_X are bounded by $\mathrm{Card}(A)^n$, one has $\rho_X \geq 1/\mathrm{Card}(A)$.

PROPOSITION 1.4.3. *For any rational set X, the generating function $f_X(z)$ is a rational fraction.*

Proof. Let $X, Y \subset A^*$. If the union of X and Y is unambiguous, then $f_{X+Y} = f_X + f_Y$. Similarly, if the product of X and Y is unambiguous, then $f_{XY} = f_X f_Y$. Finally, if X is a code, then $f_{X^*} = 1/(1-f_X)$ by Proposition 1.4.1. ■

The following lemma gives a method to compute the radius of convergence of an unambiguous star.

LEMMA 1.4.4. *Let $X \subset A^+$ be a code. If there is a real number α with $0 < \alpha < \rho_X$ such that $f_X(\alpha) = 1$, then $\rho_{X^*} = \alpha$ and f_{X^*} is unbounded on $]0, \alpha[$.*

Proof. The function $r \mapsto f_X(r)$ is continuous and increasing on $[0, \alpha]$. Thus $f_X(r)$ takes all values in $[0, 1]$ for $0 \leq r \leq \alpha$. Thus $f_{X^*}(r) = \sum (f_X(r))^n$ converges for $0 < r < \alpha$ and diverges for $r = \alpha$. In particular, $\rho_{X^*} = \alpha$ and f_{X^*} is unbounded on $]0, \alpha[$. ■

The next lemma shows that the hypothesis of Lemma 1.4.4 is satisfied when the code is rational.

LEMMA 1.4.5. *Let $X \subset A^+$ be a nonempty rational set. For each real number $\alpha > 0$, there exist an real number β with $0 < \beta < \rho_X$ such that $f_X(\beta) = \alpha$.*

Proof. It suffices to prove that f_X is unbounded on the interval $]0, \rho_X[$. Indeed, the function $\beta \mapsto f_X(\beta)$ is a continuous increasing function on the interval $]0, \rho_X[$. By Proposition 1.3.5, any recognizable set is unambiguously rational. The proof is by induction on an unambiguous rational expression for X. The conclusion is clear if X is finite, i.e. if f_X is a polynomial. Assume that f_X is unbounded on $]0, \rho_X[$ and similarly f_Y is unbounded on $]0, \rho_Y[$. If the product XY is unambiguous, then $\rho_{XY} = \min(\rho_X, \rho_Y)$. Thus f_{XY} is unbounded on $]0, \rho_{XY}[$. Similarly, if the sum $X+Y$ is unambiguous, then $\rho_{X+Y} = \min(\rho_X, \rho_Y)$. Thus f_{X+Y} is unbounded on $]0, \rho_{X+Y}[$.

Finally, let X be a code such that f_X is unbounded on $]0, \rho_X[$. Let $0 < \beta < \rho_X$ be a real number such that $f_X(\beta) = 1$. The conclusion follows by Lemma 1.4.4. ■

The following example shows that for a code X which is not rational, there may not exist any solution of the equation $f_X(r) = 1$.

EXAMPLE 1.4.6. The set of words on $A = \{a, b\}$ having an equal number of occurrences of a and b is a submonoid of A^* generated by a prefix code D. Since any word of D^* of length $2n$ is obtained by choosing n positions among $2n$, we have

$$f_{D^*}(z) = \sum_{n \geq 0} \binom{2n}{n} z^{2n}.$$

By a simple application of the binomial formula, we obtain

$$f_{D^*}(z) = (1 - 4z^2)^{-\frac{1}{2}}.$$

This follows indeed, using the simple identity

$$\binom{-\frac{1}{2}}{n} = \frac{1}{(-4)^n}\binom{2n}{n}.$$

We have $f_D(z) = 1 - 1/f_{D^*}(z)$ and thus

$$f_D(z) = 1 - \sqrt{1-4z^2}.$$

Let D_a be the set of words of D which start with a and let D_b be the set of those which start with b. Then $D = D_a + D_b$ and $f_{D_a} = f_{D_b}$. Thus the prefix code $X = D_a$ satisfies

$$f_X = \frac{1-\sqrt{1-4z^2}}{2}.$$

Since $z = 1/2$ is a singularity of $f_X(z)$, we have $\rho_X = 1/2$. However $f_X(1/2) = 1/2$. Thus, $f_X([0, \rho_X[) = [0, \frac{1}{2}[$.

The following result that the set of factors of a rational set X is not much "larger" than X itself.

Proposition 1.4.7. *Let X be a rational set. Then*

$$\rho_X = \rho_{F(X)}$$

Proof. Let

$$f_X(z) = \sum_{n\geq 0} a_n z^n, \qquad f_{F(X)}(z) = \sum_{n\geq 0} b_n z^n$$

Since $X \subset F(X)$, we have $a_n \leq b_n$.

Let $\mathcal{A}$ be a finite automaton recognizing X with set of states Q. For each state q, there are words u_q and v_q and an initial state i_q and a terminal state t_q and a path $i_q \xrightarrow{u_q} q \xrightarrow{v_q} t_q$. Let w be a word of length n in $F(X)$. There exist two words u_q and v_p such that $u_q w v_p$ belongs to X. Thus

$$F(X) \subset \bigcup_{p,q\in Q} u_q^{-1} X v_p^{-1}$$

It follows that

$$a_n \leq b_n \leq a_n + a_{n+1} + \cdots + a_{n+k+\ell}$$

where k is the maximal length of the words u_q and ℓ is the maximal length of the words v_p. This shows that the series $f_X(z)$ and $f_{F(X)}(z)$ have the same radius of convergence. ∎

The following example shows that Proposition 1.4.7 may be false when X is not rational.

Example 1.4.8. Let $A = \{a, b\}$ and let $X = \{a^n bw \mid |w| = n\}$. This is a prefix code and $f_X(z) = \sum_{n\geq 1} 2^n z^{2n+1} = \frac{z}{1-2z^2}$ Thus $\rho_X = \sqrt{2}/2$. However, $F(X) = A^*$ and thus $\rho_{F(X)} = 1/2 < \rho_X$.

1.5. Symbolic dynamical systems

In this section, we present some basic concepts of symbolic dynamics. The alphabets considered in this section are finite.

1.5.1. Definitions

A two sided-infinite word $z \in A^{\mathbb{Z}}$ *avoids* a set of words $X \subset A^*$ if no factor of z is in X. We denote by S_X the set of all $y \in A^{\mathbb{Z}}$ which avoid X.

A *symbolic dynamical system* is a subset S of $A^{\mathbb{Z}}$ of the form S_X for some $X \subset A^+$.

A symbolic dynamical system S is defined by the set X of words that it avoids. Since $F(S)$ is, by definition, the complement of X in A^*, the set S is also determined by the set $F(S)$ of its factors. In particular, if S and T are two dynamical systems such that $F(S) = F(T)$, then $S = T$.

PROPOSITION 1.5.1. *A subset of $A^{\mathbb{Z}}$ is a symbolic dynamical system if and only if it is closed for the topology and invariant under the shift.*

Proof. It is clear that a symbolic dynamical system is both closed and invariant. Conversely, let $S \subset A^{\mathbb{Z}}$ be a closed and invariant set. Let $X = A^+ - F(S)$ be the set of words that do not appear as a factor in any of the words of S. We prove that $S = S_X$. It follows from the definition of X that $y \in S_X$ if and only if $F(y) \subset F(S)$. This shows that $S \subset S_X$. Conversely, let $y \in S_X$. For each integer n, let $w_n = y_{-n} \cdots y_{n-1}y_n$. Since $w_n \in F(y)$, there is a word $y^{(n)} \in S$ such that $w_n \in F(y^{(n)})$. Since S is shift-invariant, we can suppose that $w_n = y_{-n}^{(n)} \cdots y_n^{(n)}$. This implies that the sequence $y^{(n)}$ converges to y. Since S is closed, we obtain $y \in S$ and this concludes the proof. ∎

The system is denoted S or (S, σ) to emphasize the role of the shift σ and it is also called a *subshift*.

For example, $(A^{\mathbb{Z}}, \sigma)$ itself is a symbolic dynamical system, often called the *full shift* .

As a less trivial example, let us consider the following subshift.

EXAMPLE 1.5.2. Let S be the set of two-sided infinite words on $A = \{a, b\}$ such that a symbol a is always followed by a symbol b. Since $S = S_{\{aa\}}$, it is a subshift often called the *golden mean* subshift.

Let $h : A^{\mathbb{Z}} \to B^{\mathbb{Z}}$ be a literal morphism, with A finite. For any subshift S of $A^{\mathbb{Z}}$, the set $T = h(S)$ is a subshift of $B^{\mathbb{Z}}$. Indeed, T is clearly shift-invariant. It is also closed: Consider a sequence (y_n) of elements of T converging to some $y \in B^{\mathbb{Z}}$. Let (x_n) be a sequence of elements of S such that $y_n = h(x_n)$. Since S is compact, there is a subsequence (x_{n_i}) of the (x_n) which converges to some x in S. Then $y = h(x)$ and thus y is in T.

Conversely, if T is a subshift of $B^{\mathbb{Z}}$, then it is easy to see that $h^{-1}(T)$ is a subshift of $A^{\mathbb{Z}}$, even if A is infinite.

A subshift $S \subset A^{\mathbb{Z}}$ is of *finite type* if $S = S_X$ for some *finite* set $X \subset A^+$. As an example, the golden mean subshift is of finite type.

Let S be a subshift, and let $I(S) = A^+ - F(S)$ be the set of words avoided by S. Let $X(S)$ be the set of elements of $I(S)$ which are minimal for the factor ordering (i.e. which have no proper factor in $I(S)$). Then $S = S_{X(S)}$ and S is of finite type if and only if $X(S)$ is finite.

EXAMPLE 1.5.3. Let $G = (Q, E)$ be a finite graph. The set of two-sided infinite paths in G is a subshift of finite type. Indeed, the set $X(S)$ consists of the set of pairs of non consecutive edges. This subshift is called the *edge-shift* of G.

A subshift S is *sofic* if $S = S_X$ for some *rational* set $X \subset A^+$. As above, a subshift S is sofic if and only if $X(S)$ is a rational set.

It is clear that a subshift of finite type is sofic. The converse is not true, as shown by the following examples.

EXAMPLE 1.5.4. Let $S \subset A^{\mathbb{Z}}$ be the set of two-sided infinite words on $A = \{a, b\}$ that contain at most one b. Then $X(S) = ba^*b$. Thus S is sofic. Since $X(S)$ is infinite, S is not a subshift of finite type.

EXAMPLE 1.5.5. Let $S \subset A^{\mathbb{Z}}$ be the set of two-sided infinite words on $A = \{a, b\}$ such that the number of occurrences of a between two consecutive b is even. S is called the *even subshift*. This system is sofic, since $S = S_X$ for $X = ba(aa)^*b$. Since every proper factor of an element of X is in $F(S)$, we have $X = X(S)$. Since X is infinite, S is not a subshift of finite type.

Let $\mathcal{A} = (Q, E)$ be a finite automaton. Let $S \subset A^{\mathbb{Z}}$ be the set of labels of all two-sided infinite paths in $\mathcal{A}$. We say that it is the subshift *recognized* by $\mathcal{A}$. Any sofic subshift is obtained in this way:

PROPOSITION 1.5.6. *Let S be a subshift. The following conditions are equivalent:*

(i) *S is sofic;*
(ii) *S is recognizable by a finite automaton;*
(iii) *$F(S)$ is recognizable.*

Proof. $(i) \Rightarrow (ii)$. Let X be a rational set such that $S = S_X$. Let $\mathcal{A}$ be a finite trim automaton recognizing the rational set $A^* - A^*XA^*$. Let S' be the subshift recognized by $\mathcal{A}$. We claim that $F(S') = F(S)$. Indeed, a word $w \in F(S')$ is the label of some path in $\mathcal{A}$. Since $\mathcal{A}$ is trim, there exist words u, v such that uwv is recognized by $\mathcal{A}$. Thus, $w \in F(S)$. Conversely, by compactness, any label of a path in $\mathcal{A}$ is a factor of the label of a two-sided infinite path. Thus $S = S'$. The implications $(ii) \Rightarrow (iii)$ and $(iii) \Rightarrow (i)$ are clear. ∎

EXAMPLE 1.5.7. The golden mean subshift of Example 1.5.2 is recognized by the golden mean automaton given in Figure 1.3. The subshift of Example 1.5.4 is recognized by the automaton given in Figure 1.10.

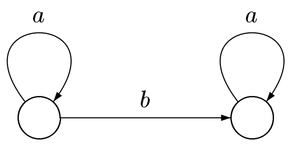

Figure 1.10. One b automaton.

The subshift of Example 1.5.5 is recognized by the even a automaton given in Figure 1.11.

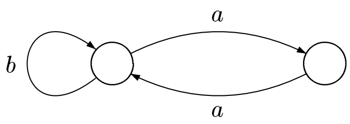

Figure 1.11. Even a automaton.

A consequence of Proposition 1.5.6 is that a subshift S is sofic if and only if the set $F(S)$ of its factors is rational. Indeed, if S is sofic, $F(S)$ is recognized by any automaton recognizing S.

The class of sofic subshifts is closed under image and inverse image by a literal morphism h. Indeed, if S is a sofic system and if $T = h(S)$, then $F(T) = h(F(S))$, and thus $F(T)$ is a rational set. Conversely, if T is a sofic system, then $S = h^{-1}(T)$, then $F(S) = h^{-1}(F(T))$. Thus again, $F(S)$ is a rational set.

As an example, let S be a sofic subshift recognized by an automaton $\mathcal{A}$. The set S is the image of the edge-shift of $\mathcal{A}$ under the literal morphism that maps each edge to its label.

A subshift S is *irreducible* if for all $u, v \in F(S)$, there is a word $w \in F(S)$ such that uwv is in $F(S)$.

The subshift of Example 1.5.4 is not irreducible since b appears at most once in a word. On the contrary, the subshift of Example 1.5.5 and the golden mean system are irreducible, as a consequence of the following result.

PROPOSITION 1.5.8. *A sofic subshift is irreducible if and only if it can be recognized by a strongly connected automaton.*

Proof. To prove that the condition is sufficient, let $\mathcal{A}$ be a strongly connected automaton recognizing S. Let u, v be words in $F(S)$, and consider two paths $p \xrightarrow{u} q$ and $r \xrightarrow{v} s$ in $\mathcal{A}$. Since $\mathcal{A}$ is strongly connected, there is a path from q to r labeled by some word w. Thus, uwv is the label of some path in $\mathcal{A}$ and $uwv \in F(S)$.

Conversely, we consider a trim deterministic automaton recognizing the set $F(S)$. Let $\mathcal{A}'$ be an automaton that is a strongly connected component of $\mathcal{A}$

without edges leaving this component. We prove that $\mathcal{A}'$ recognizes S. It suffices to prove that any word in $F(S)$ is the label of some path in $\mathcal{A}'$. Let $w \in F(S)$. Let u be the label of a path from the initial state i of $\mathcal{A}$ to some state p of $\mathcal{A}'$. Since S is irreducible, there is a word v such that $uvw \in F(S)$. Since $\mathcal{A}$ is deterministic, there is a path starting from p labeled with vw. This path is in $\mathcal{A}'$. Thus, w is the label of a path in $\mathcal{A}'$. ■

The notions introduced above can also be formulated in the context of one-sided infinite words. A one-sided symbolic system, or one-sided subshift, is a set $S \subset A^{\mathbb{N}}$ which is both closed and invariant. Equivalently, it is the set of right infinite sequences that appear in a subshift. We shall usually work with two-sided subshifts because two-sided shifts take into account both the past and the future. An exception will be made in Section 1.5.2 concerning the notion of recurrence.

1.5.2. Recurrence and minimality

In this section, we concentrate on a special kind of symbolic dynamical systems: the smallest system containing a given infinite word. It is more appropriate to present it in the one-sided case. We define $S(x) = \{y \in A^{\mathbb{N}} \mid F(y) \subset F(x)\}$ where $F(x)$ denotes the set of factors of x. The set $S(x)$ is the smallest subshift containing x. Indeed, $S(x) = S_{A^*-F(x)}$. This shows that $S(x)$ is a subshift. Moreover, if $x \in S_X$ for some $X \subset A^+$, then $X \subset A^* - F(x)$ and thus $S_X \supset S_{A^*-F(x)} = S(x)$.

A one-sided infinite word $x \in A^{\mathbb{N}}$ is said to be *recurrent* if any factor occurring in x has an infinite number of occurrences. It obviously suffices for x to be recurrent, that any prefix of x has a second occurrence in x.

It is easy to verify that x is recurrent if and only if the subshift $S(x)$ is irreducible. Indeed, if $S(x)$ is irreducible, then for any prefix u of x there is a v such that $uvu \in F(x)$ and thus u has a second occurrence. Conversely, if x is recurrent then for any $u, v \in F(x)$, v has an occurrence following any occurrence of u and thus there is a word w such that $uwv \in F(x)$.

A word $x \in A^{\mathbb{N}}$ is said to be *uniformly recurrent* if every block of x appears infinitely often at bounded distance, in other terms if, for every word $w \in F(x)$, there exists an integer r such that w is a factor of every word in $F_r(x)$.

A periodic word is obviously uniformly recurrent. We shall see another example below (Example 1.5.10).

These notions are strongly related to that of a *minimal subshift*, i.e. a subshift $S \subset A^{\mathbb{N}}$ such that $T \subset S$ implies $T = \emptyset$ or $T = S$.

The following result is one of the earliest in symbolic dynamics.

THEOREM 1.5.9. *Let $x \in A^{\mathbb{N}}$ be a one-sided infinite word. The following conditions are equivalent.*

1. *x is uniformly recurrent.*
2. *$S(x)$ is minimal.*

Proof. $1 \Rightarrow 2$. Let $S \subset S(x)$ be a subshift and let $y \in S$. Then $S(y) \subset S$. Since $y \in S(x)$, we have $F(y) \subset F(x)$ by the definition of $S(x)$. Let $w \in F(x)$. Since x is uniformly recurrent, w must appear in every long enough factor of x. If v is a factor of y of this length, then since it is also a factor of x, it admits w as a factor. Hence $w \in F(y)$. This shows that $F(x) = F(y)$ and this implies that $S(y) = S = S(x)$.

$2 \Rightarrow 1$. For every $y \in S(x)$, one has $F(x) = F(y)$, since $S(x)$ is minimal. For a given block w of x, we define $i_w(y)$ to be the function assigning to $y \in S(x)$ the least integer i such that $y = uwz$ with $|u| = i$. Since i_w is continuous and $S(x)$ is compact, i_w is bounded. Let w be a block of $x = uwy$. Since $y \in S(x)$, $w \in F(y)$ and thus w has a second occurrence in x at a distance equal to $|w| + i_w(y)$, hence bounded. ∎

EXAMPLE 1.5.10. The word of Thue-Morse $t = \mu^\omega(0)$ is uniformly recurrent. Indeed, 000 or 111 are not in $F(t)$. Thus successive occurrences of 0 or 1 are separated by at most two symbols. It follows that any block of t appears at bounded distance since it has to appear in some $\mu^k(0)$ or $\mu^k(1)$, and successive occurrences of blocks are again separated by at most two blocks. The system $S(t)$ is known as the *Morse minimal set*.

We used in the proof of Theorem 1.5.9 a possible variant of the definition of a uniformly recurrent word: for all $n > 0$ there is an $m > n$ such that any factor of length n appears in any factor of length m. This condition can be used as a definition for a uniformly recurrent two-sided infinite word. It also leads to the definition of a function $r_x(n)$ called the *recurrence index* of x. We let $r_x(n) = m$ if m is the smallest possible integer such that any factor of length n appears in any factor of length m. It is well defined for all integers n if and only if x is uniformly recurrent.

THEOREM 1.5.11. *Every nonempty subshift contains a uniformly recurrent word.*

Proof. Let S be a nonempty subshift. We define a decreasing sequence (H_n) of subshifts of S as follows. Let $H_0 = S$. Suppose that H_{n-1} is already defined. Let H_n be the set of elements of H_{n-1} which have a minimal number of factors of length n. Each H_n is a subshift. Let H be the intersection of the H_n. Since $A^{\mathbb{N}}$ is compact, any decreasing sequence of closed subsets has a nonempty intersection. Thus H is nonempty. Let x be an infinite word in H and let $S(x)$ be the smallest subshift containing x. Then $S(x)$ is clearly minimal and thus, x is uniformly recurrent. ∎

As an application of this result, we mention

PROPOSITION 1.5.12. *For any infinite set L of words over a finite alphabet, there is a uniformly recurrent infinite word x such that $F(x) \subset F(L)$.*

Proof. Let S be the subshift avoiding $A^* - F(L)$. Then $F(S) \subset F(L)$ and $F(S)$ is not empty by König's Lemma. Any uniformly recurrent word x in S satisfies $F(x) \subset F(L)$. ∎

1.5.3. Entropy

Let S be a nonempty subshift, and let $F = F(S)$ be the set of factors of S. The *entropy* of S is the number

$$h(S) = -\log(\rho_F)$$

Clearly, $0 \leq h(S) \leq \log k$, where $k = \operatorname{Card} A$. It is also clear that, for any subshifts S, T, if $S \subset T$, then $h(S) \leq h(T)$.

A subshift S is said to be *coded* if there is a prefix code X such that $F(S) = F(X^*)$. In this case, we say that X is a code for S. Any sofic system is coded. Indeed, if $\mathcal{A}$ is a deterministic strongly connected automaton recognizing S, the code of first returns to some state of $\mathcal{A}$ is a code for S. The following example shows that the notion of a coded subshift is more general than the notion of a sofic system.

Example 1.5.13. Let $X = \{a^n b^n \mid n \geq 1\}$. The system S coded by X is not sofic. Let us indeed suppose the contrary and let $\mathcal{A}$ be a finite automaton recognizing S. For each $n \geq 1$, since $(a^n b^n)^\zeta$ is included in S , there is in $\mathcal{A}$ a cycle labeled by some power of $a^n b^n$. These cycles cannot be all disjoint since the automaton is finite. This gives clearly infinite words which are not in S.

The following theorem gives a method to compute the entropy of a sofic system.

Theorem 1.5.14. *Let S be a sofic subshift and let X be a rational code for S. Then*

$$h(S) = -\log r$$

where r is the unique positive solution of the equation $f_X(z) = 1$.

Proof. By Proposition 1.4.7, we have $\rho_{X^*} = \rho_{F(X^*}$. By Proposition 1.4.4, $\rho_{X^*} = r$. ∎

An alternative method to compute r is to use the fact that $1/r$ is the maximal eigenvalue of the matrix associated to any unambiguous automaton recognizing S. Let in fact, M be a matrix with real coefficients. The spectral radius ρ of M is the maximal modulus of its eigenvalues. One has (see Gantmacher 1960 for example)

$$\rho = \limsup \sqrt[n]{\|M^n\|}$$

where $\|M\|$ is any norm of the matrix M.

Let now S be a sofic system, and let $\mathcal{A} = (Q, E)$ be an unambiguous automaton recognizing S. Let M be the $Q \times Q$-matrix with integer coefficients defined by

$$M_{p,q} = \mathrm{Card}\{a \in A \mid (p, a, q) \in E\}$$

Let us choose the particular norm equal to the sum of modulus of all coefficients. Then the number s_n of factors of length n appearing in S ratifies $s_n \leq \|M^n\| \leq c \cdot s_n$ for some constant c. Thus the entropy of S is $\log \rho$.

EXAMPLE 1.5.15. Let S be the even subshift recognized by the automaton of Figure 1.11. We have $X = \{aa, b\}$ and $f_X(z) = z + z^2$. Thus $r = 1/\tau$ where τ is the golden mean. Accordingly, the maximal eigenvalue of the matrix

$$M = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$$

is τ and the entropy of S is $h(S) = \log \tau$.

EXAMPLE 1.5.16. Let $X = \{a^n b^n \mid n \geq 1\}$ and let S be the subshift coded by X, as in Example 1.5.13. We have $F(X^*) = QX^*P$ where P (Q) is the set of proper left (right) factors of words of X. Actually $P = a^+X \cup \varepsilon$ since the nonempty words of P have the form $a^n b^m$ for $0 \leq m < n$. Similarly, $Q = Xb{+}\cup\varepsilon$. Thus $\rho_{F(X^*)} = \min(\rho_Q, \rho_{X^*}, \rho_P) = \rho_{X^*}$. Since $f_X(z) = \sum_{n \geq 1} z^{2n}$ we have $f_X(z) = \frac{z^2}{1-z^2}$. The equation $f_X(r) = 1$ has the solution $r = \sqrt{2}/2$. Hence

$$h(S) = \log \sqrt{2}.$$

1.6. Unavoidable sets

Unavoidable sets are sets of words X such that any infinite word has a factor in X. The purpose of this section is to present several properties of unavoidable sets. The main result is that, for each integer k, there is an explicit description of the unavoidable sets of cardinality k.

We start with several equivalent definitions of unavoidable sets and some elementary properties.

In this section, all alphabets are supposed to be finite and to contain at least two letters.

1.6.1. Definitions and elementary properties

Recall from Section 1.5 that a two-sided infinite word $z \in A^{\mathbb{Z}}$ *avoids* a set of words $X \subset A^*$ if no factor of z is in X. The set of all $y \in A^{\mathbb{Z}}$ which avoid X is denoted by S_X.

A set X of words over an alphabet A is called *unavoidable* (over A) if the set S_X is empty.

EXAMPLE 1.6.1. The set A^n is unavoidable for all $n \geq 0$.

EXAMPLE 1.6.2. The set $X = \{a, bb\}$ is unavoidable over $\{a, b\}$. Indeed, any two-sided infinite word over $\{a, b\}$ either contains an a or is reduced to b^ζ.

PROPOSITION 1.6.3. *A set X of words over A is unavoidable if and only if the set $A^* - A^*XA^*$ is finite.*

Proof. Assume first that X is unavoidable. Arguing by contradiction, suppose that $Y = A^* - A^*XA^*$ is infinite. By König's Lemma, there is a two-sided infinite word y with all its factors in Y. Consequently, y is in S_X, a contradiction.

Conversely, if Y is finite, any two-sided infinite word has a factor in X. ∎

PROPOSITION 1.6.4. *Any unavoidable set X contains a finite unavoidable set.*

Proof. Let d be the maximal length of the words in the finite set $Y = A^* - A^*XA^*$. Let Z be the set of words in X of length at most $d+1$. Every word of length $d+1$ has a factor in X which actually is in Z. Thus Z is unavoidable. ∎

A set containing an unavoidable set is again unavoidable. It is therefore natural to consider minimal unavoidable sets.

Minimal unavoidable sets contained in a given unavoidable set are not necessarily unique. Indeed the set $\{aa, ab, ba, bb\}$ contains both $\{aa, ab, bb\}$ and $\{aa, ba, bb\}$, which both are easily seen to be unavoidable and minimal.

The following example shows the existence of minimal unavoidable sets of arbitrary size $n \geq k$ on an alphabet with $k \geq 2$ letters.

EXAMPLE 1.6.5. Let first $A = \{a, b\}$. For each $n \geq 2$, the set

$$X = \{aa, aba, abba, \ldots, ab^{n-2}a, b^{n-1}\}$$

is a minimal unavoidable set with n elements. Indeed, any infinite word avoiding b^{n-1} has a block of the form ab^ia with $i < n-1$. This shows that X is unavoidable. For each $0 \leq i < n$, the infinite word $(ab^i)^\zeta$ has only one factor in X, namely ab^ia for $i < n-1$ and b^{n-1} for $i = n-1$. This shows that X is minimal.

Let now A be an alphabet with $k \geq 3$ letters. We use two symbols $a, b \in A$ to build as above a minimal unavoidable set X having size $n - k + 2$. The set $X \cup A - \{a, b\}$ is a minimal unavoidable set of size n.

It is worth observing that if X is a finite unavoidable set over A, then A has to be finite. Indeed, for each letter $a \in A$, some a^n is in X and thus $\mathrm{Card}(X) \geq \mathrm{Card}(A)$.

The following result gives an equivalent formulation of the definition of finite unavoidable sets which will be used in the sequel.

PROPOSITION 1.6.6. *A finite set X of words is unavoidable if and only if every periodic two-sided infinite word has a factor in X.*

Proof. The condition is clearly necessary. For the converse, we argue by contradiction. Let x be a two-sided infinite word avoiding X. Let n be the maximal length of the words in X. Consider two disjoint occurrences of a factor u of length n in x. Then there is a factor $y = uzu$ of x for some word z. Each word in $(uz)^\zeta$ is a periodic two-sided infinite word avoiding X. ∎

It is worth observing that the proposition becomes false when X is infinite. Indeed, the set of squares over a three-letter alphabet is a counter-example because it is avoidable, but every periodic word contains a square.

To check in practice that a given finite set X is unavoidable, there are two possible algorithms.

The first one consists in computing a graph $G = (P, E)$, where P is the set of prefixes of X and E is the set of pairs (p, s) for which there is a letter $a \in A$ such that s is the longest suffix of pa which is in P.

PROPOSITION 1.6.7. *A finite set X is unavoidable if and only if every cycle in G contains a vertex in X. Proof. For each integer $n \geq 0$, and vertices $u, v \in P$, there is a path of length n from u to v if and only if there exists a word y of length n such that v is the longest suffix of uy in P. This can be proved by induction on n. It follows that there is a path of length n from ε to a vertex $x \in X$ if and only if $AX \cap A^n \neq \emptyset$.* ∎

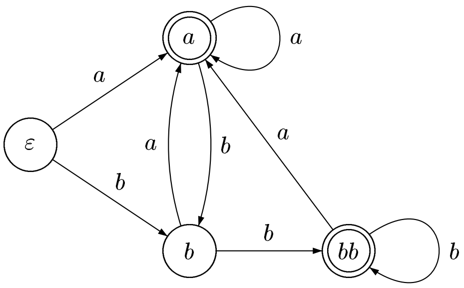

Figure 1.12. The graph for $X = \{a, bb\}$.

EXAMPLE 1.6.8. For $X = \{a, bb\}$, the word graph is given in Figure 1.12. By inspection, the set X is unavoidable.

The second algorithm is sometimes easier to write down by hand. Say that a set Y of words is obtained from a finite set of words X by an *elementary derivation* if

(i) either there exist words $u, v \in X$ such that u is a proper factor of v, and

$$Y = X - v$$

(ii) or there exists a word $x = ya \in X$ with $a \in A$ such that, for each letter $b \in A$ there is a suffix z of y such that $zb \in X$, and

$$Y = X - x + y$$

A *derivation* is a sequence of elementary derivations. We say that Y is derived from X if Y is obtained from X by a derivation.

EXAMPLE 1.6.9. Let $X = \{aaa, b\}$. Then we have the derivations

$$X \to \{aa, b\} \to \{a, b\} \to \{\varepsilon, b\} \to \{\varepsilon\}$$

where the first three arrows follow case (ii) and the last one case (i).

The following result shows in particular that if Y is derived from X, then X is unavoidable if and only if Y is unavoidable.

PROPOSITION 1.6.10. *If Y is derived from X, then $S_X = S_Y$.*

Proof. It is enough to consider the case of an elementary derivation. In the first case where $Y = X - v$, where v has a factor in X, then clearly $S_X = S_Y$. In the second case, we clearly have $S_Y \subset S_X$ since Y is obtained by replacing an element of X by one of its factors. Conversely, assume by contradiction the existence of some $s \in S_X - S_Y$. The only possible factor of s in Y is y. Let b be the letter following y in s. Then s has a factor in X, namely zb where z is the suffix of y such that $zb \in X$ whose existence is granted by the definition of a derivation. This is a contradiction. ∎

The notion of a derivation gives a practical method to check whether a set is unavoidable. We have indeed the following result.

PROPOSITION 1.6.11. *A finite set X is unavoidable if and only if there is a derivation from X to the set $\{\varepsilon\}$.*

Proof. Let $X \neq \{\varepsilon\}$ be unavoidable. We prove the existence of a derivation to $\{\varepsilon\}$ by induction on the sum $l(X)$ of the lengths of words in X. If $\varepsilon \in X$, we may derive $\{\varepsilon\}$ from X. Thus assume $\varepsilon \notin X$, and let w be a word of maximal length avoiding X. For each $b \in A$ there is a word $x_b = zb \in X$ which is a suffix of wb. Let $x_a = ya$ be the longest of the words x_b. Then the hypotheses of case (ii) are satisfied and thus there is a derivation from X to a set Y with $l(Y) < l(X)$. The converse is clear by Proposition 1.6.10. ∎

In practice, there is a shortcut which is useful to perform derivations. It is described in the following transformation from X to Y.

(iii) there is a word y such that $ya \in X$ for each $a \in A$ and

$$Y = X - \sum_{a \in A} ya + y$$

It is clear that such a set Y can be derived from X and thus, we do not change the definition of derivations by adding case (iii) to the definition of elementary derivations. We use this new definition in the following example.

EXAMPLE 1.6.12. Let $X = \{aaa, aba, abb, bbb\}$. We have the following sequence of derivations (with the symbol a in the word $x = ya$ underlined at each step)

$$\begin{aligned}\{aaa, ab\underline{a}, ab\underline{b}, bbb\} &\to \{aa\underline{a}, ab, bbb\}\\ &\to \{a\underline{a}, a\underline{b}, bbb\} \to \{a, bb\underline{b}\} \to \{a, b\underline{b}\}\\ &\to \{\underline{a}, \underline{b}\} \to \{\varepsilon\}\end{aligned}$$

Derivations could of course be performed on the left rather than on the right (see Problem 1.5.3).

1.6.2. The structure theorem

We will now see how one can describe the unavoidable sets with a fixed number of elements. Our aim is to give a description of this family in a parametric form $\{x_1^{k_1}, \ldots, x_m^{k_m}\}$ for some words $x_1, \ldots, x_m$ and integers $k_1, \ldots, k_m$.

To avoid confusions, we call a *family* any set of subsets of A^*. We will thus speak of the family of unavoidable sets. An n-subset is a set with n elements.

An *n-section* of a family $\mathcal{F} = \{X_1, \ldots, X_m\}$ is a set X of n words of A^* containing at least one element of each $F(X_i)$, for $i = 1, \ldots, m$. We denote by $\sec_n(X_1, \ldots, X_m)$ the family of n-sections of $\mathcal{F}$.

EXAMPLE 1.6.13. The family of 2-sections of the family $\mathcal{F} = \{\{a\}, b^*\}$ is composed of the sets $\{\varepsilon, w\}$ for $w \in A^+$, and of the sets $\{a, b^n\}$ for $n \geq 1$.

A subset X of A^* is *simple* if it is of the form $X = \{u\}$ or $X = u^*$ for some word u. A family $\mathcal{F}$ is *simple* if it is composed of simple subsets of A^*.

A family $\mathcal{X}$ of n-subsets of A^* has finite *dimension* if it is the union of the n-sections of a finite number of simple families, i.e.

$$\mathcal{X} = \bigcup_{i=1}^{k} \sec_n(X_{i,1}, \ldots, X_{i,m(i)}) \tag{1.6.1}$$

where the $X_{i,j}$ are simple sets. The dimension of $\mathcal{X}$ is the minimum value of the maximum of the $m(i)$ for all representations of $\mathcal{X}$ in the form (1.6.1).

EXAMPLE 1.6.14. The family $\mathcal{F} = \{\{a\}, b^*\}$ of Example 1.6.13 is simple. Thus, the family of its 2-sections has dimension 2.

EXAMPLE 1.6.15. The family of one element sets has infinite dimension. Indeed otherwise there would exist a finite number of words such that every word is a factor of a power of one of these words.

Theorem 1.6.16. *For each integer $n \geq 0$, the family of unavoidable sets having n elements has dimension n.*

Proof. Let Y be a subset of A^*. Let $\mathcal{R}_n(Y)$ be the family of n-subsets X of A^* such that $X \cup Y$ is unavoidable. We prove by induction on n that, for every Y, the family $\mathcal{R}_n(Y)$ has dimension at most n.

The claim holds for $n = 0$ since the family $\mathcal{R}_0(Y)$ is formed of the empty set, and thus has dimension 0.

Let us suppose that the result holds for $\mathcal{R}_{n-1}(Y)$ for any set Y. Let now Y be any set, and consider $\mathcal{R}_n(Y)$. We distinguish two cases.

Case 1. There exist $n+1$ words $u_1, \ldots, u_{n+1}$ such that the sets u_i^ζ are disjoint and all their elements avoid Y. Let T be the set of words which are factor of at least two u_i^ζ. The set T is finite. Indeed, if u and v are any two primitive words which are not conjugate, then by Fine and Wilf's Theorem, any word in $F(u^\zeta) \cap F(v^\zeta)$ has length at most $|u| + |v| - 2$.

Any set $X \in \mathcal{R}_n(Y)$ contains an element of T. Indeed, each element of u_i^ζ has a factor in X and since there are $n+1$ words u_i, two of them have a common factor in X. If $t \in X$, we may write $X \cup Y$ as $Z \cup (Y \cup \{t\})$ with $Z = X - \{t\}$. This shows that

$$\mathcal{R}_n(Y) = \bigcup_{t \in T} \{Z + t \mid Z \in \mathcal{R}_{n-1}(Y + t)\}$$

By the induction hypothesis, each family $\mathcal{R}_{n-1}(Y + t)$ has dimension at most $n - 1$. Since T is finite, this implies that $\mathcal{R}_n(Y)$ has dimension at most n.

Case 2. The set of two-sided infinite periodic words avoiding Y is contained in a union $u_1^\zeta \cup \cdots \cup u_k^\zeta$ with $k \leq n$. Then $\mathcal{R}_n(Y)$ is the family of n-sections of the family $\{u_1^*, \ldots, u_k^*\}$.

This proves that the family of unavoidable set of cardinality n has dimension at most n. This results indeed from the case $Y = \emptyset$ in the above argument.

The proof that the dimension is exactly n relies on the existence of minimal unavoidable sets of arbitrary size (see Example 1.6.5).

We argue by contradiction and suppose that the family of unavoidable sets with n elements has dimension less than n. Let X be a minimal unavoidable set with n elements. By assumption, X is an n-section of some simple family $\mathcal{F}$ with less than n elements, and all n-sections of $\mathcal{F}$ are unavoidable. There is some $x \in X$ such that $Y = X - x$ in an $n-1$-section of $\mathcal{F}$. Then, for any $x' \in A^* - X$, the set $Y + x'$ is an n-section of $\mathcal{F}$, and thus it is unavoidable. It is clear that this implies that Y itself is unavoidable, a contradiction. ∎

Example 1.6.17. Let $\mathcal{U}_n$ denote the family of unavoidable sets with n elements on $A = \{a, b\}$. We give below a list of finite sets $\mathcal{F}_n$ of simple families for $n \leq 4$ such that $\mathcal{U}_n$ is the set of n-sections of the elements of $\mathcal{F}_i$ for $i \leq n$.

Let first $\mathcal{F}_1$ be reduced to $\{\varepsilon\}$ (we identify in this example $\{w\}$ with w). Next, $\mathcal{F}_2$ is composed of the two families

$$\{a, b^*\}, \quad \{a^*, b\}.$$

The family $\mathcal{F}_3$ is composed of the simple families

$$\{aa, bb, (ab)^*\}, \quad \{aa, b^*, bab\}, \quad \{a^*, b^*, ab\}$$

and the two additional ones obtained by exchanging a and b.

The family $\mathcal{F}_4$ is composed of the simple families

$$\begin{array}{lll}\{aa, bbb, babbab, (ab)^*\}, & \{aa, bbb, babab, (abb)^*\}, & \{aa, b^*, babab, bbabb\}, \\ \{aaa, b^*, bab, baab\}, & \{a^*, b^*, bab, baa\}, & \{a^*, b^*, bab, aab\}\end{array}$$

and the six additional ones obtained by exchanging a and b.

Problems

Section 1.1

1.1.1 Consider a binary operation on the set $\mathbb{N} \times \mathbb{N}$ defined by

$$(i,j)(k,\ell) = \begin{cases} (i+k-j, \ell) & \text{if } j \leq k \\ (i, j-k+\ell) & \text{otherwise}. \end{cases}$$

Show that this operation is associative, with neutral element $(0,0)$. The set $\mathbb{N} \times \mathbb{N}$ equipped with this operation is called the *bicyclic monoid*.

Section 1.2

1.2.1 Show that w is primitive if and only if its period is not a proper divisor of its length.

1.2.2 Let A be an alphabet with k elements. Let X be a subset of A^* such that $F(X) \neq A^*$. Show that $\rho_X > 1/k$. *Hint*: Take $w \notin F(X)$. If w is a letter, then $X \subset (A-w)^*$ and thus $\rho_X > 1/(k-1)$. In the general case, show that the result holds for each set $X_i = \{x \in X \mid |x| \equiv i \bmod |w|\}$.

1.2.3 Show that a set of infinite words is open for the topology if it is of the form XA^ω for some $X \subset A^*$

1.2.4 Let A be an alphabet, and let \$ be a letter not in A. Any word w over A can be viewed as the infinite word $w\$^\omega \in (A \cup \$)^{\mathbb{N}}$. Show that a sequence (u_n) of words over A converges to an infinite word x if and only if it is not ultimately constant and if the sequence $u_n\$^\omega$ converges to x in the topological space $(A \cup \$)^{\mathbb{N}}$.

1.2.5 Consider a closed curve in the plane which is *normal*, i.e. has only finitely many self-intersections and these are transverse double points. Label the intersections with distinct symbols from an alphabet A. The *Gauss code* of the curve is the word obtained as the successive intersection points met by proceeding along the curve and noting each crossing point label as it is traversed. The word obtained is really a conjugacy class.

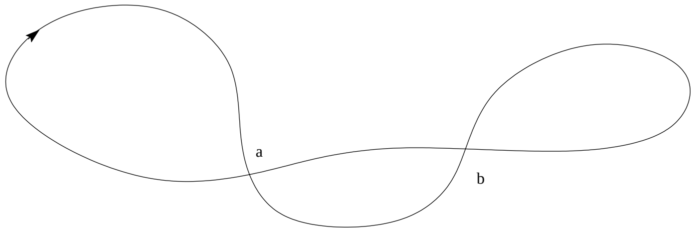

Figure 1.13. A closed curve

For example, the Gauss code of the curve of Figure 1.13 is *abba*.
Prove that every symbol appears exactly twice in a Gauss code. For each symbol $a \in A$, let $\Delta(a)$ be the set of symbols occurring exactly once between two occurrences of a. Prove that in a Gauss code, the cardinality of $\Delta(a)$ is even for each $a \in A$. Show that this condition is not sufficient for a word with two occurrences of each symbol to be a Gauss code. (Hint: consider the word *abcadcedbe*).

Section 1.3

1.3.1 Let $\mathcal{A} = (Q, E, i, T)$ be a finite trim deterministic automaton with a unique initial state recognizing a set X. Show that there is a function f from Q onto $Q(X)$ such that $f(i) = X$ and $f(q \cdot a) = f(q) \cdot a$ for all $q \in Q$ and $a \in A$. Derive from this that $\mathcal{A}(X)$ is the unique deterministic automaton recognizing X having a minimal number of states.

1.3.2 Show that the syntactic semigroup of X is the smallest semigroup recognizing X in the sense that, for every semigroup S recognizing X, there exists a morphism from S onto the syntactic semigroup of X. Show that X is recognizable if and only if its syntactic semigroup is finite.
Let $\mathcal{A}(X)$ be the minimal automaton of X. Define a semigroup morphism f from A^+ into $Q(X)^{Q(X)}$ by $f(w) : q \mapsto q \cdot w$. Show that the semigroup $f(A^+)$ is isomorphic to the syntactic semigroup of X.

1.3.3 Set $A = \{1, 2, 3\}$, and consider the following function $f : A^* \to A^*$. A word $x \in A^*$ can be written as

$$x = a_1^{i_1} a_2^{i_2} \cdots a_n^{i_n}$$

with $i_j \geq 1$ and $a_i \neq a_{i+1}$ for $1 \leq i < n$. Let X be the set of words x for which $i_j \leq 3$ for $j = 1, \ldots, n$. Define a function f on X by

$$f(x) = i_1 a_1 i_2 a_2 \cdots i_n a_n \,.$$

The iterates of f on the letter 1 are

$$\begin{array}{c} 1 \\ 1\,1 \\ 2\,1 \\ 1\,2\,1\,1 \\ 1\,1\,1\,2\,2\,1 \\ 3\,1\,2\,2\,1\,1 \end{array}$$

1. Show that $f(x) \in X$.
2. Show that f is both left subsequential and right subsequential.

1.3.4 Use the de Bruijn graph of order n over a q letter alphabet to prove the existence of an infinite word of period q^{n+1} having all words of length $n+1$ as its factors. (Hint: use the fact that in the de Bruijn graph, there exists a circuit using exactly once each edge (such a circuit is called Eulerian).)

Section 1.4

1.4.1 Let $f : A^* \to A^*$ be a morphism such that for any two letters $a, b \in A$ there is at least an occurrence of b in $f(a)$. Then any infinite word x such that $f(x) = x$ is uniformly recurrent (Martin 1971).

1.4.2 Show that a recurrent one-sided infinite word x over a finite alphabet is uniformly recurrent if and only if any infinite set of factors of x contains two elements which are factors one of the other.

1.4.3 Let (u_n) be a sequence of positive real numbers such that $u_{n+m} \leq u_n + u_m$. Show that $\lim_{n\to\infty} u_n/n$ exists and is equal to $\inf_{n\to\infty} u_n/n$.

Section 1.5

1.5.1 Let X be a finite unavoidable set over A. Let n be the maximal length of the elements of X and let $k = \mathrm{Card}(A)$. Show that any word of length $k^n + n - 1$ has a factor in X.

1.5.2 For a given set $X \subset A^+$, we define an order on A^* by requiring that for all $u, v \in A^*$ and $x \in X$, we have

$$uv < uxv$$

Prove the following generalization of Higman's theorem: the above defined order is a well quasi order if and only if the set X is unavoidable.

1.5.3 A *two-sided derivation* is obtained by adding in the definition of a derivation the following case.

(ii′) there exists a word $x = ay \in X$ with $a \in A$ such that, for each letter $b \in A$ there is a prefix z of y such that $bz \in X$, and

$$Y = X - x + y$$

Show that for any finite set X there is a two-sided derivation from X to the set Y of words avoided by S_X which are minimal for the factor ordering.

1.5.4 A set $X \subset A^+$ is called *separable* if for every $x \in X$ there is a periodic two-sided infinite word s such that $F(s) \cap X = \{x\}$. Show that a finite set X is a subset of a minimal unavoidable set if and only if it is separable.

To prove that the condition is sufficient, choose for each $x \in X$ a periodic infinite word s_x such that $F(s_x) \cap X = \{x\}$. Let $S(x)$ be the closure of s_x under the shift. Since

$$S = \cup_{x \in X} S(x)$$

is a subshift of finite type, there is a finite set I such that $S = S_I$. Then $X + I$ is unavoidable and there is a minimal unavoidable set Y such that $X \subset Y \subset X + I$.

1.5.5 An unavoidable set $X \subset A^+$ is called *irreducible* if it is minimal and if, for x in X and for every proper factor y of x the set $X - x + y$ is not minimal.

1. Show that there is, up to symmetries (exchange of a and b and reversal), only one irreducible unavoidable set with four elements, namely $\{aa, aba, abb, bbb\}$.
2. Show that for each integer n, there are only finitely many irreducible sets with n elements on a given alphabet.

Use Theorem 1.6.16 and Dickson's Lemma (see e.g. Problem 6.1.2 in Lothaire 1983)

1.5.6 Let $X = \{aaa, bbbb, abbbab, abbab, abab, bbaabb, baabaab\}$. Verify that X is unavoidable.

Show that for any word x in X and any letter a in A, the sets

$$Y = X - x + xa \quad \text{and} \quad Y' = X - x + ax$$

are avoidable.

Notes

There are numerous references concerning automata and formal languages, see e.g. Hopcroft and Ullman 1979, Eilenberg 1974. The original reference to de Bruijn graphs is de Bruijn 1946. See also Problem 1.3.4 and van Lint and Wilson 1992 for a reference to Eulerian circuits. Theorem 1.3.13 is due to Coven and Hedlund 1973.

A good reference on symbolic dynamics is the book of Lind and Marcus 1995. Our presentation here follows closely the survey by Béal and Perrin 1997. Theorems 1.5.9,1.5.12 are classical results due to Morse and Hedlund (Morse and Hedlund 1938).

Unavoidable sets seem to appear for the first time in Schützenberger 1964. In this paper, he proves an asymptotic estimate of the minimal cardinality $c(k, n)$

of an unavoidable set of words of fixed length n on a k-letter alphabet (see Lothaire 1983 p. 99). No exact formula is known for $c(k,n)$.

Further papers on unavoidable sets appeared later (see Choffrut and Culik 1984, Crochemore, Lerest, and Wender 1983)

Theorem 1.6.16 is due to Rosaz 1998. Laurent Rosaz has also obtained many other interesting results. Problem 1.5.5 appears in Rosaz 1998. The example of Problem 1.5.6 is due to Rosaz (Rosaz 1998). It answers by the negative a conjecture which had been formulated by Ehrenfeucht and asking whether for any unavoidable set X there is a word x in X and a letter a in A such that $X + xa - x$ is still unavoidable. Problem 1.5.4 is a result again due to Rosaz 1995.

The Gauss codes of Problem 1.2.5 have been introduced by Gauss in Gauss 1900. Several characterizations of Gauss codes has been given (see Treybig 1968, Marx 1969, Lovasz and Marx 1976, Rosenstiehl 1976). For a history of the subject, see Grünbaum 1972 or Rosenstiehl 1999.

Problem 1.3.3 describes a transformation studied by J. H. Conway under the name of "audioactive decay" (Conway 1987). A number of amazing results were obtained by Conway (see also Vardi (1991), Ekhad and Zeilberger (1997))

Problem 1.5.2 is due to Ehrenfeucht, Hausler, and Rozenberg 1983a. It is a generalization of the famous Higman Theorem (see Lothaire 1983).