1

Preliminaries

Valérie Berthé and Michel Rigo

1.1 Conventions

Let us briefly start with some basic notation used throughout this book. The set of non-negative integers (respectively integers, rational numbers, real numbers, complex numbers) is $\mathbb{N}$ (respectively $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$). In particular, the set $\mathbb{N}$ is $\{0,1,2,\ldots\}$. We use the notation $[\![i,j]\!]$ for the set of integers $\{i,i+1,\ldots,j\}$.The floor of a real number x is $\lfloor x \rfloor = \sup\{z \in \mathbb{Z} \mid z \le x\}$, whereas $\{x\} = x - \lfloor x \rfloor$ stands for the fractional part of x.

1.2 Words

This section is only intended to give basic definitions about words. For material not covered in this book, classical textbooks on finite or infinite words and their properties are (Lothaire, 1983, 2002, 2005), (Allouche and Shallit, 2003), and (Queffélec, 1987). See also the chapter (Choffrut and Karhumäki, 1997) or the tutorial (Berstel and Karhumäki, 2003). The book (Rigo, 2014) can also serve as introductory lecture notes on the subject.

1.2.1 Finite words

An *alphabet* is a finite, non-empty set. Its elements are referred to as *symbols* or *letters*. In this book, depending on the specific context or conventions of a given chapter, alphabets will be denoted by capital letters like Σ or A.

Definition 1.2.1 A (finite) *word* over Σ is a finite sequence of letters from Σ. The empty sequence is called the *empty word* and it is denoted by ε. The sets of all finite words, finite non-empty words and infinite words over Σ are denoted by Σ^*, Σ^+ and Σ^ω, respectively. A word $w = w_0 w_2 \cdots w_n$ where $w_i \in \Sigma$, $0 \le i \le n$, can be seen as a function $w : \{0,1,\ldots,n\} \to \Sigma$ in which $w(i) = w_i$ for all i.

Combinatorics, Words and Symbolic Dynamics, ed. Valérie Berthé and Michel Rigo. Published by Cambridge University Press. ©Cambridge University Press 2016.

Definition 1.2.2 Let $\mathbb{S}$ be a set equipped with a single binary operation

$$\star : \mathbb{S} \times \mathbb{S} \to \mathbb{S}.$$

It is convenient to call this operation a *multiplication* over $\mathbb{S}$, and the product of $x, y \in \mathbb{S}$ is usually denoted by xy. If this multiplication is *associative*, i.e., for all $x, y, z \in \mathbb{S}$, $(xy)z = x(yz)$, then the algebraic structure given by the pair $(\mathbb{S}, \star)$ is a *semigroup*. If, moreover, multiplication has an identity element, i.e., there exists some element $1 \in \mathbb{S}$ such that, for all $x \in \mathbb{S}$, $x1 = x = 1x$, then $(\mathbb{S}, \star)$ is a *monoid*. In addition if every element $x \in \mathbb{S}$ has an *inverse*, i.e., there exists $y \in \mathbb{S}$ such that $xy = 1 = yx$, then $(\mathbb{S}, \star)$ is a *group*.

Let $u = u_0 \cdots u_{m-1}$ and $v = v_0 \cdots v_{n-1}$ be two words over Σ. The *concatenation* of u and v is the word $w = w_0 \cdots w_{m+n-1}$ defined by $w_i = u_i$ if $0 \leq i < m$, and $w_i = v_{i-m}$ otherwise. We write $u \cdot v$ or simply uv to express the concatenation of u and v. The concatenation (or catenation) of words is an associative operation, i.e., given three words u, v and w, $(uv)w = u(vw)$. Hence, parenthesis can be omitted. In particular, the set Σ^* (respectively, Σ^+) equipped with the concatenation product is a monoid (respectively, a semigroup).

Concatenating a word w with itself k times is abbreviated by w^k. In particular, $w^0 = \varepsilon$. Furthermore, for an integer m and a word $w = w_1 w_2 \cdots w_n$, where $w_i \in \Sigma$ for $1 \leq i \leq n$, the *rational power*

$$w^{m/n}$$

is $w^q w_1 w_2 \cdots w_r$, where $m = qn + r$ for $0 \leq r < n$. For instance, we have

$$(\texttt{abbab})^{9/5} = \texttt{abbababba}. \tag{1.1}$$

The *length* of a word w, denoted by $|w|$, is the number of occurrences of the letters in w. In other words, if $w = w_0 w_2 \cdots w_{n-1}$ with $w_i \in \Sigma$, $0 \leq i < n$, then $|w| = n$. In particular, the length of the empty word is zero. For $a \in \Sigma$ and $w \in \Sigma^*$, we write $|w|_a$ for the number of occurrences of a in w. Therefore, we have

$$|w| = \sum_{a \in \Sigma} |w|_a.$$

A word u is a *factor* of a word v (respectively, a *prefix*, or a *suffix*), if there exist words x and y such that $v = xuy$ (respectively, $v = uy$, or $v = xu$). A factor (respectively, prefix, suffix) u of a word v is called *proper* if $u \neq v$ and $u \neq \varepsilon$. Thus, for example, if $w =$ `concatenation`, then `con` is a prefix, `ate` is a factor, and `nation` is a suffix.

The *mirror* (sometimes called *reversal*) of a word $u = u_0 \cdots u_{m-1}$ is the word $\tilde{u} = u_{m-1} \cdots u_0$. It can be defined inductively on the length of the word by $\tilde{\varepsilon} = \varepsilon$ and $\widetilde{au} = \tilde{u}a$ for $a \in \Sigma$ and $u \in \Sigma^*$. Notice that for $u, v \in \Sigma^*$, $\widetilde{uv} = \tilde{v}\tilde{u}$. A *palindrome* is a word u such that $\tilde{u} = u$. For instance, the palindromes of length at most 3 in $\{0,1\}^*$ are $\varepsilon, 0, 1, 00, 11, 000, 010, 101, 111$.

1.2.2 Infinite words

Definition 1.2.3 An (one-sided right) *infinite word* is a map from $\mathbb{N}$ to Σ. If $\mathbf{w}$ is an infinite word, we often write

$$\mathbf{w} = a_0 a_1 a_2 \cdots ,$$

where each $a_i \in \Sigma$. The set of all infinite words of Σ is denoted Σ^ω (one can also find the notation $\Sigma^{\mathbb{N}}$).

The notions of *factor*, *prefix* or *suffix* introduced for finite words can be extended to infinite words. Factors and prefixes are finite words, but a suffix of an infinite word is also infinite.

Definition 1.2.4 A two-sided or *bi-infinite word* is a map from $\mathbb{Z}$ to Σ. The set of all bi-infinite words is denoted ${}^\omega\Sigma^\omega$ (one can also find the notation $\Sigma^{\mathbb{Z}}$).

Example 1.2.5 Consider the infinite word $\mathbf{x} = x_0 x_1 x_2 \cdots$ where the letters $x_i \in \{0,\dots,9\}$ are given by the digits appearing in the usual decimal expansion of $\pi - 3$,

$$\pi - 3 = \sum_{i=0}^{+\infty} x_i \, 10^{-i-1},$$

i.e., $\mathbf{x} = 14159265358979323846264338327950288419\cdots$ is an infinite word.

Definition 1.2.6 An infinite word $\mathbf{x} = x_0 x_1 \cdots$ is *(purely) periodic* if there exists a finite word $u = u_0 \cdots u_{k-1} \neq \varepsilon$ such that $x = u^\omega$, i.e., for all $n \geq 0$, we have $x_n = u_r$ where $n = dk + r$ with $r \in \{0,\dots,k-1\}$. An infinite word x is *eventually periodic* (or, *ultimately periodic*) if there exist two finite words $u, v \in \Sigma^*$, with $v \neq \varepsilon$ such that $x = uvvv\cdots = uv^\omega$. Notice that purely periodic words are special cases of eventually periodic words. For any eventually periodic word x, there exist words u, v of shortest length such that $x = uv^\omega$, then the integer $|u|$ (respectively $|v|$) is referred to as the *preperiod* (respectively *period*) of x. An infinite word is said to be *non-periodic* if it is not ultimately periodic.

Definition 1.2.7 The *language* of the infinite word $\mathbf{x}$ is the set of all its factors. It is denoted by $L(\mathbf{x})$. The set of factors of length n occurring in $\mathbf{x}$ is denoted by $L_n(\mathbf{x})$.

Definition 1.2.8 An infinite word $\mathbf{x}$ is *recurrent* if all its factors occur infinitely often in $\mathbf{x}$. It is *uniformly recurrent* (also called *minimal*), if it is recurrent and for every factor u of $\mathbf{x}$, if $T_{\mathbf{x}}(u) = \left\{ i_1^{(u)} < i_2^{(u)} < i_3^{(u)} < \cdots \right\}$ is the infinite set of positions where u occurs in $\mathbf{x}$, then there exists a constant C_u such that, for all $j \geq 1$,

$$i_{j+1}^{(u)} - i_j^{(u)} \leq C_u.$$

Definition 1.2.9 One can endow Σ^ω with a *distance* d defined as follows. Let $\mathbf{x}, \mathbf{y}$ be two infinite words over Σ. Let $\mathbf{x} \wedge \mathbf{y}$ denote the longest common prefix of $\mathbf{x}$ and $\mathbf{y}$. Then the distance d is given by

$$d(\mathbf{x},\mathbf{y}) := \begin{cases} 0, & \text{if } \mathbf{x} = \mathbf{y}, \\ 2^{-|\mathbf{x}\wedge\mathbf{y}|}, & \text{otherwise.} \end{cases}$$

It is obvious to see that, for all $\mathbf{x},\mathbf{y},\mathbf{z} \in \Sigma^\omega$, $d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x})$, $d(\mathbf{x},\mathbf{z}) \le d(\mathbf{x},\mathbf{y}) + d(\mathbf{y},\mathbf{z})$ and $d(\mathbf{x},\mathbf{y}) \le \max(d(\mathbf{x},\mathbf{z}), d(\mathbf{y},\mathbf{z}))$. This last property is not required to have a distance, but when it holds, the distance is said to be *ultrametric*. Note that we obtain an equivalent distance if we replace 2 with any real number $r > 1$.

This notion of distance extends to $\Sigma^{\mathbb{Z}}$. Notice that the topology on Σ^ω is the product topology (of the discrete topology on Σ). The space Σ^ω is a compact *Cantor set*, that is, a totally disconnected compact space without isolated points. Since Σ^ω is a (complete) metric space, it is therefore relevant to speak of convergent sequences of infinite words. The sequence $(\mathbf{z}_n)_{n\ge 0}$ of infinite words over Σ *converges* to $\mathbf{x} \in \Sigma^\omega$, if for all $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that, for all $n \ge N$, $d(\mathbf{z}_n, \mathbf{x}) < \varepsilon$. To express the fact that a sequence of finite words $(w_n)_{n\ge 0}$ over Σ converges to an infinite word $\mathbf{y}$, it is assumed that Σ is extended with an extra letter $c \notin \Sigma$. Any finite word w_n is replaced with the infinite word $w_n ccc\cdots$ and if the sequence of infinite words $(w_n ccc\cdots)_{n\ge 0}$ converges to $\mathbf{y}$, then the sequence $(w_n)_{n\ge 0}$ is said to converge to $\mathbf{y}$.

Let $(u_n)_{n\ge 0}$ be a sequence of non-empty finite words. If we define, for all $\ell \ge 0$, the finite word v_ℓ as the concatenation $u_0 u_1 \cdots u_\ell$, then the sequence $(v_\ell)_{\ell\ge 0}$ of finite words converges to an infinite word. This latter word is said to be the concatenation of the elements in the infinite sequence of finite words $(u_n)_{n\ge 0}$. In particular, for a constant sequence $u_n = u$ for all $n \ge 0$, $v_\ell = u^{\ell+1}$ and the concatenation of an infinite number of copies of the finite word u is denoted by u^ω.

1.3 Morphisms

Particular infinite words of interest can be obtained by iterating morphisms (or homomorphisms of free monoids). Morphisms are also called *substitutions*. A map $h : \Sigma^* \to \Delta^*$, where Σ and Δ are alphabets, is called a *morphism* if h satisfies $h(xy) = h(x)h(y)$ for all $x, y \in \Sigma^*$. A morphism may be specified by providing the values $h(a)$ for all $a \in \Sigma$. For example, we may define a morphism $h : \{0,1,2\}^* \to \{0,1,2\}^*$ by

$$\begin{aligned} 0 &\mapsto 01201 \\ 1 &\mapsto 020121 \\ 2 &\mapsto 0212021. \end{aligned} \tag{1.2}$$

This domain of a morphism is easily extended to (one-sided) infinite words.

A morphism $h : \Sigma^* \to \Sigma^*$ such that $h(a) = ax$ for some $a \in \Sigma$ and $x \in \Sigma^*$ with $h^i(x) \ne \varepsilon$ for all i is said to be *prolongable on* a; we may then repeatedly iterate h to obtain the infinite *fixed point*

$$h^\omega(a) = a\,x\,h(x)\,h^2(x)\,h^3(x)\cdots.$$

This infinite word is said to be *purely morphic*. The morphism h given by (1.2) above is prolongable on 0, so we have the fixed point

$$h^\omega(0) = 012010201210212021012010201 21\cdots.$$

A morphism h is *non-erasing* if $h(a) \neq \varepsilon$ for all $a \in \Sigma$. Otherwise it is *erasing*. A morphism is k-*uniform* if $|h(a)| = k$ for all $a \in \Sigma$; it is *uniform* if it is k-uniform for some k.

Example 1.3.1 (Thue–Morse word) For example, if the morphism $\mu : \{0,1\}^* \to \{0,1\}^*$ is defined by

$$\begin{aligned} 0 &\mapsto 01 \\ 1 &\mapsto 10, \end{aligned}$$

then μ is 2-uniform. This morphism is often referred to as the *Thue–Morse morphism*. The fixed point

$$\mathbf{t} = \mu^{\omega}(0) = 0110100110010110\cdots$$

is known as the *Thue–Morse word*.

Example 1.3.2 (Fibonacci word) Another significant example of a purely morphic word is the *Fibonacci word*. It is obtained from the non-uniform morphism defined over the alphabet $\{0,1\}$ by $\sigma : 0 \mapsto 01, 1 \mapsto 0$,

$$\sigma^{\omega}(0) = (x_n)_{n\geq 0} = 0100101001001010010100100101001001010010100\cdots.$$

It is a Sturmian word and can be obtained as follows. Let $\phi = (1+\sqrt{5})/2$ be the Golden Ratio. For all $n \geq 1$, if $\lfloor (n+1)\phi \rfloor - \lfloor n\phi \rfloor = 2$, then $x_{n-1} = 0$, otherwise $x_{n-1} = 1$.

1.4 Languages and machines

Formal languages theory is mostly concerned with the study of the mathematical properties of sets of words. For an exhaustive exposition on regular languages and automata theory, see (Sakarovitch, 2003) and (Perrin and Pin, 2004) for the connections with infinite words. Also see the chapter (Yu, 1997), or (Sudkamp, 1997), (Hopcroft and Ullman, 1979) and the updated revision (Hopcroft *et al.*, 2006) for general introductory books on formal languages theory.

1.4.1 Languages of finite words

Let Σ be an alphabet. A subset L of Σ^* is said to be a *language*. Note for instance that this terminology is consistent with the one of Definition 1.2.7. Since a language is a *set* of words, we can apply all the usual set operations like union, intersection or set difference: $\cup$, $\cap$ or $\setminus$. The concatenation of words can be extended to define an operation on languages. If L, M are languages, LM is the language of the words obtained by concatenation of a word in L and a word in M, i.e.,

$$LM = \{uv \mid u \in L, v \in M\}.$$

We can of course define the concatenation of a language with itself, so it permits us to introduce the power of a language. Let $n \in \mathbb{N}$, Σ be an alphabet and $L \subseteq \Sigma^*$ be a language. The language L^n is the set of words obtained by concatenating n words in L. We set $L^0 := \{\varepsilon\}$. In particular, we recall that Σ^n denotes the set of words of length n over Σ, i.e., concatenations of n letters in Σ. The *(Kleene) star* of the language L is defined as

$$L^* = \bigcup_{i \geq 0} L^i.$$

Otherwise stated, L^* contains the words that are obtained as the concatenation of an arbitrary number of words in L. Notice that the definition of Kleene star is compatible with the notation Σ^* introduced to denote the set of finite words over Σ. We also write $L^{\leq n}$ as a shorthand for

$$L^{\leq n} = \bigcup_{i=0}^{n} L^i.$$

Note that if the empty word belongs to L, then $L^{\leq n} = L^n$. We recall that $\Sigma^{\leq n}$ is the set of words over Σ of length at most n. More can be found in Section 6.3.1 where the notion of code is introduced.

Example 1.4.1 Let $L = \{a, ab, aab\}$ and $M = \{a, ab, ba\}$ be two finite languages. We have

$$L^2 = \{aa, aab, aaab, aba, abab, abaab, aaba, aabab, aabaab\}$$

and

$$M^2 = \{aa, aab, aba, abab, abba, baa, baab, baba\}.$$

One can notice that $\operatorname{Card}(L^2) = (\operatorname{Card} L)^2$ but $\operatorname{Card}(M^2) < (\operatorname{Card} M)^2$. This is due to the fact that all words in L^2 have a unique factorisation as concatenation of two elements in L but this is not the case for M, where $(ab)a = a(ba)$. We can notice that

$$L^* = \{a\}^* \cup \{a^{i_1} b a^{i_2} b \cdots a^{i_n} b a^{i_{n+1}} \mid \forall n \geq 1, i_1, \ldots, i_n \geq 1,\ i_{n+1} \geq 0\}.$$

Since languages are sets of (finite) words, a language can be either *finite* or *infinite*. For instance, a language L differs from $\emptyset$ or $\{\varepsilon\}$ if, and only if, the language L^* is infinite. Let L be a language, we set $L^+ = LL^*$. The mirror operation can also be extended from words to languages: $\tilde{L} = \{\tilde{u} \mid u \in L\}$.

Definition 1.4.2 A language is *prefix-closed* (respectively *suffix-closed*) if it contains all prefixes (respectively suffixes) of any of its elements. A language is *factorial* if it contains all factors of any of its elements.

Obviously, any factorial language is prefix-closed and suffix-closed. The converse does not hold. For instance, the language $\{a^n b \mid n > 0\}$ is suffix-closed but not factorial.

Example 1.4.3 The set of words over $\{0,1\}$ containing an even number of 1s is the language

$$\begin{aligned} E &= \{w \in \{0,1\}^* \mid |w|_1 \equiv 0 \pmod 2\} \\ &= \{\varepsilon, 0, 00, 11, 000, 011, 101, 110, 0000, 0011, \ldots\}. \end{aligned}$$

This language is closed under mirror, i.e., $\tilde{L} = L$. Notice that the concatenation $E\{1\}E$ is the language of words containing an odd number of 1s and $E \cup E\{1\}E = E(\{\varepsilon\} \cup \{1\}E) = \{0,1\}^*$. Notice that E is neither prefix-closed, since $1001 \in E$ but $100 \notin E$, nor suffix-closed. See also Example 8.1.3 and Example 9.2.9.

If a language L over Σ can be obtained by applying to some finite languages a finite number of operations of union, concatenation and Kleene star, then this language is said to be a *regular language*. This generation process leads to *regular expressions* which are well-formed expressions used to describe how a regular language is built in terms of these operations. From the definition of a regular language, the following result is immediate.

Theorem 1.4.4 *The class of regular languages over Σ is the smallest subset of 2^{Σ^*} (for inclusion) containing the languages $\emptyset$, $\{a\}$ for all $a \in \Sigma$ and closed under union, concatenation and Kleene star.*

Example 1.4.5 For instance, the language L over $\{0,1\}$ whose words do not contain the factor 11 is regular. It is called the *Golden mean shift*, see also Example 9.2.1. This language can be described by the regular expression $L = \{0\}^*\{1\}\{0,01\}^* \cup \{0\}^*$. Otherwise stated, it is generated from the finite languages $\{0\}$, $\{0,01\}$ and $\{1\}$ by applying union, concatenation and star operations. Its complement in Σ^* is also regular and is described by the regular expression $\Sigma^*\{11\}\Sigma^*$. The language E from Example 1.4.3 is also regular, we have the following regular expression $\{0\}^*(\{1\}\{0\}^*\{1\}\{0\}^*)^*$ describing E.

1.4.2 Automata

As we shall briefly explain in this section, the regular languages are exactly the languages recognised by finite automata.

Definition 1.4.6 A *finite automaton* is a labelled graph given by a 5-tuple $\mathscr{A} = (Q, \Sigma, E, I, T)$ where Q is the (finite) *set of states*, $E \subseteq Q \times \Sigma^* \times Q$ is the finite set of *edges* defining the *transition relation*, $I \subseteq Q$ is the set of *initial states* and T is the *set of terminal (or final) states*. A *path* in the automaton is a sequence

$$(q_0, u_0, q_1, u_1, \ldots, q_{k-1}, u_{k-1}, q_k)$$

such that, for all $i \in \{0, \ldots, k-1\}$, $(q_i, u_i, q_{i+1}) \in E$, $u_0 \cdots u_{k-1}$ is the *label* of the path. Such a path is *successful* if $q_0 \in I$ and $q_k \in T$. The language $L(\mathscr{A})$ *recognised* (or *accepted*) by $\mathscr{A}$ is the set of labels of all successful paths in $\mathscr{A}$.

Any finite automaton $\mathscr{A}$ gives a partition of Σ^* into $L(\mathscr{A})$ and $\Sigma^* \setminus L(\mathscr{A})$. When depicting an automaton, initial states are marked with an incoming arrow and terminal states are marked with an outgoing arrow. A transition like (q,u,r) is represented by a directed edge from q to r with label u, $q \xrightarrow{u} r$.

Example 1.4.7 In Figure 1.1 the automaton has two initial states p and r, three terminal states q, r and s. For instance, the word ba is recognised by the automaton. There are two successful paths corresponding to the label ba: (p,b,q,a,s) and (p,b,p,a,s). For this latter path, we can write $p \xrightarrow{b} p \xrightarrow{a} s$. On the other hand, the word $baab$ is not recognised by the automaton.

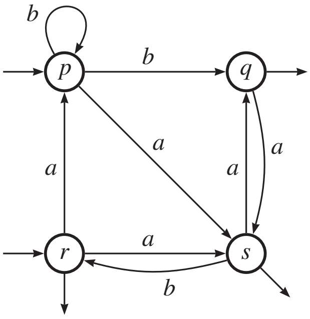

Figure 1.1 A finite automaton.

Example 1.4.8 The automaton in Figure 1.2 recognises exactly the language E of the words having an even number of 1 from Example 1.4.3.

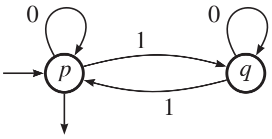

Figure 1.2 An automaton recognising words with an even number of 1.

Definition 1.4.9 Let $\mathscr{A} = (Q,\Sigma,E,I,T)$ be a finite automaton. A state $q \in Q$ is *accessible* (respectively *co-accessible*) if there exists a path from an initial state to q (respectively from q to some terminal state). If all states of $\mathscr{A}$ are both accessible and co-accessible, then $\mathscr{A}$ is said to be *trim*.

Definition 1.4.10 A finite automaton $\mathscr{A} = (Q,\Sigma,E,I,T)$ is said to be *deterministic* (*DFA*) if it has only one initial state q_0, if E is a subset of $Q \times \Sigma \times Q$ and for each $(q,a) \in Q \times \Sigma$ there is at most one state $r \in Q$ such that $(q,a,r) \in E$. In that case,

E defines a partial function $\delta_{\mathscr{A}} : Q \times \Sigma \to Q$ that is called the *transition function* of $\mathscr{A}$. The adjective *partial* means that the domain of $\delta_{\mathscr{A}}$ can be a strict subset of $Q \times \Sigma$. To express that the partial transition function is total, the DFA can be said to be *complete*. To get a total function, one can add to Q a new 'sink state' s and, for all $(q,a) \in Q \times \Sigma$ such that $\delta_{\mathscr{A}}$ is not defined, set $\delta_{\mathscr{A}}(q,a) := s$. This operation does not alter the language recognised by $\mathscr{A}$. We can extend $\delta_{\mathscr{A}}$ to be defined on $Q \times \Sigma^*$ by $\delta_{\mathscr{A}}(q,\varepsilon) = q$ and, for all $q \in Q$, $a \in \Sigma$ and $u \in \Sigma^*$, $\delta_{\mathscr{A}}(q,au) = \delta_{\mathscr{A}}(\delta_{\mathscr{A}}(q,a),u)$. Otherwise stated, the language recognised by $\mathscr{A}$ is $L(\mathscr{A}) = \{u \in \Sigma^* \mid \delta_{\mathscr{A}}(q_0,u) \in F\}$ where q_0 is the initial state of $\mathscr{A}$. If the automaton is deterministic, it is sometimes convenient to refer to the 5-tuple $\mathscr{A} = (Q,\Sigma,\delta_{\mathscr{A}},I,T)$.

As explained by the following result, for languages of finite words, finite automata and deterministic finite automata recognise exactly the same languages.

Theorem 1.4.11 (Rabin and Scott (1959)) *If L is recognised by a finite automaton $\mathscr{A}$, there exists a DFA which can be effectively computed from $\mathscr{A}$ and recognising the same language L.*

A proof and more details about classical results in automata theory can be found in textbooks like (Hopcroft *et al.*, 2006), (Sakarovitch, 2003) or (Shallit, 2008). For standard material in automata theory we shall not refer again to these references below.

One important result is that the set of regular languages coincides with the set of languages recognised by finite automata.

Theorem 1.4.12 (Kleene (1956)) *A language is regular if, and only if, it is recognised by a (deterministic) finite automaton.*

Observe that if L, M are two regular languages over Σ, then $L \cap M$, $L \cup M$, LM and $L \setminus M$ are also regular languages. In particular, a language over Σ is regular if, and only if, its complement in Σ^* is regular.

Example 1.4.13 The regular language $L = \{0\}^*\{1\}\{0,01\}^* \cup \{0\}^*$ from Example 1.4.5 is recognised by the DFA depicted in Figure 1.3. Notice that the state s is a *sink*: non-terminal state and all transitions remain in s.

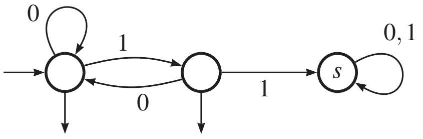

Figure 1.3 A DFA accepting words without factor 11.

1.5 Symbolic dynamics

Let us introduce some basic notions in symbolic dynamics. For expository books on the subject, see (Cornfeld *et al.*, 1982), (Kitchens, 1998), (Lind and Marcus, 1995), (Perrin, 1995) and (Queffélec, 1987). For references on ergodic theory, also see, e.g., (Walters, 1982).

1.5.1 Codings of dynamical systems

A (discrete) dynamical system is a pair (X,T) where $T : X \to X$ is a map acting on a convenient space X (e.g., X is a topological space or a metric space, in the usual setting, X is generally compact and T is continuous). We are interested in iterating the map T and we look at orbits $(T^n(x))_{n\geq 0}$ of points in X under the action T. The *trajectory* of $x \in X$ is the sequence $(T^n(x))_{n\geq 0}$. Roughly speaking, infinite words appear naturally as a convenient coding (with a priori some loss of information) of these trajectories $(T^n(x))_{n\geq 0}$. So one can gain insight about the dynamical system by studying these words, with an interplay between combinatorics on words and dynamics. In that setting, the space X is discretised, i.e., it is partitioned into finitely many sets $X_1,\dots,X_k$ and the trajectory of x is thus coded by the corresponding sequence of visited subsets, such as illustrated in Figure 1.4. Precisely, the coding of $(T^n(x))_{n\geq 0}$ is the word $\mathbf{w}_x = w_0w_1w_2\cdots$ over the alphabet $\{1,\dots,k\}$ where $w_i = j$ if and only if $T^i(x) \in X_j$. Even though the infinite word $\mathbf{w}_x$ contains less information than the original trajectory $(T^n(x))_{x\geq 0}$, this discretised and simplified version of the original system can help us to understand the dynamics of the original system.

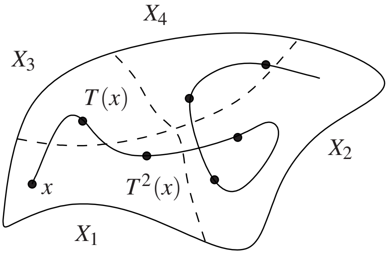

Figure 1.4 Trajectory of x in a space $X = X_1 \cup X_2 \cup X_2 \cup X_4$.

Example 1.5.1 (Rotation words) One of the simplest dynamical systems can be obtained from the coding of a rotation on a circle identified with the interval $[0,2\pi)$. Instead of working modulo 2π, it is convenient to normalise the interval $[0,2\pi)$ and