1

Preliminaries

In this preliminary chapter, we give an account of some basic notions which will be used throughout the book. This chapter is not designed for a systematic reading but rather as a reference.

The first three sections contain notation and basic vocabulary. Each of the subsequent sections is an introduction to a topic which is not completely treated in this book. These sections are concerned mainly with the theory of automata. Kleene's theorem is given and we show how to construct a minimal automaton from a given automaton. Syntactic monoids are defined. These concepts and results will be discussed in another context in Chapter 9. We introduce formal power series and weighted automata. We give some basic properties and prove parts of Perron–Frobenius theorem.

1.1 Notation

As usual, $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ denote the sets of nonnegative integers, integers, and rational, real, and complex numbers, respectively. By convention, $0 \in \mathbb{N}$. We set

$$\mathbb{R}_+ = \{x \in \mathbb{R} \mid x \geq 0\}.$$

Next,

$$\binom{n}{p} = \frac{n!}{p!(n-p)!}$$

denotes the binomial coefficient of n and p.

For real numbers $x \leq y$, we denote by $[x, y)$ the set of real numbers z such that $x \leq z$ and $z < y$. In particular, if $x = y$ this set is empty.

Given two subsets X, Y of a set Z, we define

$$X \setminus Y = \{z \in Z \mid z \in X, z \notin Y\}.$$

Frequently, $\overline{X}$ will be used to denote the complement of a subset X of some set Z. An element x and the singleton set $\{x\}$ will usually not be distinguished. The set of all subsets of a set X is denoted by $\mathfrak{P}(X)$.

The function symbols are usually written on the left of their arguments but with some exceptions: When we consider the composition of actions on a set, the action is written on the right. In particular, permutations are written on the right.

A partition of a set X is a family $(X_i)_{i\in I}$ of *nonempty* subsets of X such that

(i) $X = \bigcup_{i\in I} X_i$,
(ii) $X_i \cap X_j = \emptyset, (i \neq j)$.

We usually define a partition as follows: "Let $X = \bigcup_{i\in I} X_i$ be a partition of X". We denote the cardinality of a set X by $\mathrm{Card}(X)$.

1.2 Monoids

A *semigroup* is a set equipped with an associative binary operation. The operation is usually written multiplicatively.

A *monoid* is a semigroup which, in addition, has a neutral element. The neutral element of a monoid M is unique and is denoted by 1_M or simply by 1.

For any monoid M, the set $\mathfrak{P}(M)$ is given a monoid structure by defining, for $X, Y \subset M$,

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

The neutral element is $\{1\}$.

A *submonoid* of M is a subset N which is stable under the operation and which contains the neutral element of M, that is $1_M \in N$ and

$$NN \subset N. \tag{1.1}$$

Note that a subset N of M satisfying (1.1) does not always satisfy $1_M = 1_N$ and therefore may be a monoid without being a submonoid of M.

A *morphism* from a monoid M into a monoid N is a function $\varphi : M \to N$ which satisfies, for all $m, m' \in M$,

$$\varphi(mm') = \varphi(m)\varphi(m'),$$

and furthermore

$$\varphi(1_M) = 1_N.$$

The notions of subsemigroup and semigroup morphism are then defined in the same way as the corresponding notions for monoids.

A *congruence* on a monoid M is an equivalence relation θ on M such that, for all $m, m' \in M, u, v \in M$

$$m \equiv m' \bmod \theta \;\Rightarrow\; umv \equiv um'v \bmod \theta.$$

Let φ be a morphism from M onto N. The equivalence θ defined by $m \equiv m' \bmod \theta$ if and only if $\varphi(m) = \varphi(m')$ is a congruence.It is called the *nuclear congruence*

induced by φ. Conversely, if θ is a congruence on the monoid M, the set M/θ of the equivalence classes of θ is equipped with a monoid structure, and the canonical function from M onto M/θ is a monoid morphism.

An *idempotent* of a monoid M is an element e of M such that

$$e = e^2.$$

For each idempotent e of a monoid M, the set eMe is a monoid contained in M. It is easily seen that it is the largest monoid contained in M having e as a neutral element. It is called the *monoid localized* at e.

An element 0 of a monoid M is a *zero* if $0 \neq 1$ and for all $m \in M$

$$0m = m0 = 0.$$

If M contains a zero it is unique.

Let M be a monoid. The set of (left and right) invertible elements of M is a group called the *group of units* of M.

A *cyclic monoid* is a monoid with just one generator, that is,

$$M = \{a^n \mid n \in \mathbb{N}\}$$

with $a^0 = 1$. If M is infinite, it is isomorphic to the additive monoid $\mathbb{N}$ of nonnegative integers. If M is finite, the *index* of M is the smallest integer $i \geq 0$ such that there exists an integer $r \geq 1$ with

$$a^{i+r} = a^i. \tag{1.2}$$

The smallest integer r such that (1.2) holds is called the *period* of M. The pair composed of index i and period p determines a monoid having $i + p$ elements,

$$M_{i,p} = \{1, a, a^2, \ldots, a^{i-1}, a^i, \ldots, a^{i+p-1}\}.$$

Its multiplication is conveniently represented in Figure 1.1.

The monoid $M_{i,p}$ contains two idempotents (provided $i \geq 1$). Indeed, assume that $a^j = a^{2j}$. Then either $j = 0$ or $j \geq i$ and j and $2j$ have the same residue mod p, hence $j \equiv 0 \bmod p$. Conversely, if $j \geq i$ and $j \equiv 0 \bmod p$, then $a^j = a^{2j}$.

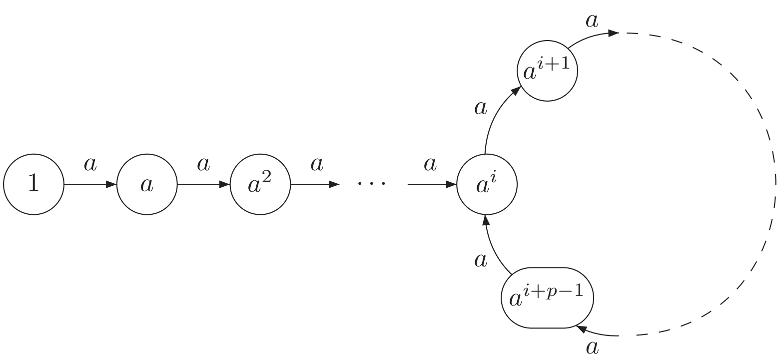

Figure 1.1 The monoid $M_{i,p}$.

Consequently, the unique idempotent $e \neq 1$ in $M_{i,p}$ is $e = a^j$, where j is the unique integer in $\{i, i+1, \ldots, i+p-1\}$ which is a multiple of p.

Let M be a monoid. For $x, y \in M$, we define

$$x^{-1}y = \{z \in M \mid xz = y\} \quad \text{and} \quad xy^{-1} = \{z \in M \mid x = zy\}.$$

For subsets X, Y of M, this notation is extended to

$$X^{-1}Y = \bigcup_{x \in X} \bigcup_{y \in Y} x^{-1}y \quad \text{and} \quad XY^{-1} = \bigcup_{x \in X} \bigcup_{y \in Y} xy^{-1}.$$

The set $X^{-1}Y$ is called a left *residual* of Y. The following identities hold for subsets X, Y, Z of M:

$$(XY)^{-1}Z = Y^{-1}(X^{-1}Z) \quad \text{and} \quad X^{-1}(YZ^{-1}) = (X^{-1}Y)Z^{-1}.$$

The notation $X^{-1}Y$ should not be confused with the product of the inverse of an element with another in some group. There is a case where the confusion could arise, in Chapter 14, where a due "caveat" will be found.

Given a subset X of a monoid M, we define

$$F(X) = M^{-1}XM^{-1}$$

to be the set of *factors* of elements in X. We have

$$F(X) = \{m \in M \mid \exists u, v \in M : umv \in X\}.$$

We sometimes use the notation $\overline{F}(X)$ to denote the complement of $F(X)$ in M,

$$\overline{F}(X) = M \setminus F(X).$$

A *relation* m over a set Q is a subset of $Q \times Q$. The *product* of two relations m and n over Q is the relation mn defined by

$$(p, r) \in mn \iff \exists q \in Q : (p, q) \in m \quad \text{and} \quad (q, r) \in n.$$

The set $\mathfrak{P}(Q \times Q)$ of relations over a set Q is a monoid for this product. Two remarkable relations are the *identity relation* id_Q and the *null relation*, which is the empty subset of $Q \times Q$. The identity relation id_Q is the neutral element of $\mathfrak{P}(Q \times Q)$. The null relation is a zero of this monoid.

A *monoid of relations* over some nonempty set Q is a submonoid of the monoid $\mathfrak{P}(Q \times Q)$. A monoid M of relations over Q is said to be *transitive* if for all $p, q \in Q$, there exists $m \in M$ such that $(p, q) \in m$.

1.3 Words

Let A be a set, which we call an *alphabet*. A *word* w on the alphabet A is a finite sequence of elements of A

$$w = (a_1, a_2, \ldots, a_n), \quad a_i \in A.$$

The set of all words on the alphabet A is denoted by A^* and is equipped with the associative operation defined by the concatenation of two sequences

$$(a_1, a_2, \ldots, a_n)(b_1, b_2, \ldots, b_m) = (a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_m).$$

This operation is associative. This allows us to write

$$w = a_1 a_2 \cdots a_n$$

instead of $w = (a_1, a_2, \ldots, a_n)$, by identifying each element $a \in A$ with the sequence (a). An element $a \in A$ is called a *letter*. The empty sequence is called the *empty word* and is denoted by 1 or ε. It is the neutral element for concatenation. Thus the set A^* of words is equipped with the structure of a monoid. The monoid A^* is called the *free monoid* on A. The set of nonempty words on A is denoted by A^+. We therefore have $A^+ = A^* \setminus 1$.

The *length* $|w|$ of the word $w = a_1 a_2 \ldots a_n$ with $a_i \in A$ is the number n of letters in w. Clearly, $|1| = 0$. The function $w \mapsto |w|$ is a morphism from A^* onto the additive monoid $\mathbb{N}$. For $n \geq 0$, we use the notation

$$A^{(n)} = \{w \in A^* \mid |w| \leq n - 1\}$$

and also

$$A^{[n]} = \{w \in A^* \mid |w| \leq n\}.$$

In particular, $A^{(0)} = \emptyset$ and $A^{[0]} = \{1\}$.

For a subset B of A, we denote by $|w|_B$ the number of letters of w which are in B. Thus

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

For a word $w \in A^*$, the set

$$\mathrm{alph}(w) = \{a \in A \mid |w|_a > 0\}$$

is the set of all letters occurring at least once in w. For a subset X of A^*, we set

$$\mathrm{alph}(X) = \bigcup_{x \in X} \mathrm{alph}(x).$$

A word $w \in A^*$ is a *factor* of a word $x \in A^*$ if there exist $u, v \in A^*$ such that $x = uwv$. The relation *is a factor of* is a partial order on A^*. A factor w of x is *proper* if $w \neq x$.

A word $w \in A^*$ is a *prefix* of a word $x \in A^*$ if there is a word $u \in A^*$ such that $x = wu$. The factor w is called *proper* if $w \neq x$. The relation *is a prefix of* is again a partial order on A^* called the *prefix order*. We write $w \leq x$ when w is a prefix of x and $w < x$ whenever $w \leq x$ and $w \neq x$. This orderhas the following fundamental

property. If, for some x,

$$w \leq x, \quad w' \leq x,$$

then w and w' are comparable, that is, $w \leq w'$ or $w' \leq w$. In other words, if $wu = w'u'$, then either there exists $s \in A^*$ such that $w = w's$ (and also $su = u'$) or there exists $t \in A^*$ such that $w' = wt$ (and then $u = tu'$).

In an entirely symmetric manner, we define a *suffix* w of a word x by $x = vw$ for some $v \in A^*$. A set $P \subset A^*$ is called *prefix-closed* if it contains the prefixes of its elements: $uv \in P \Rightarrow u \in P$. A suffix-closed set is defined symmetrically.

Consider a totally ordered alphabet A. The *lexicographic* or *alphabetic* order on A^* is defined by setting $u \prec v$ if u is a proper prefix of v, or if $u = ras$, $v = rbt$, $a < b$ for $a, b \in A$ and $r, s, t \in A^*$. The lexicographic order has the property

$$u \prec v \Leftrightarrow wu \prec wv$$

for any $u, v, w \in A^*$. Similarly, the *radix order* on A^* is defined by setting $u < v$ if $|u| < |v|$ or if $|u| = |v|$ and $u \prec v$ in the lexicographic order.

The *reversal* w of a word $w = a_1a_2 \cdots a_n$, with $a_i \in A$, is the word

$$\tilde{w} = a_n \cdots a_2a_1.$$

The notations $\tilde{w}$ and $w^{\sim}$ are equivalent. Note that for all $u, v \in A^*$,

$$(uv)^{\sim} = \tilde{v}\tilde{u}.$$

The *reversal* $\tilde{X}$ of a set $X \subset A^*$ is the set $\tilde{X} = \{\tilde{x} \mid x \in X\}$.

A *factorization* of a word $w \in A^*$ is a sequence $\{u_1, u_2, \ldots, u_n\}$ of $n \geq 0$ words in A^* such that

$$w = u_1u_2 \cdots u_n.$$

For a subset X of A^*, we denote by X^* the submonoid generated by X,

$$X^* = \{x_1x_2 \cdots x_n \mid n \geq 0, x_i \in X\}.$$

Similarly, we denote by X^+ the subsemigroup generated by X,

$$X^+ = \{x_1x_2 \cdots x_n \mid n \geq 1, x_i \in X\}.$$

We have

$$X^+ = \begin{cases} X^* \setminus 1 & \text{if } 1 \notin X, \\ X^* & \text{otherwise.} \end{cases}$$

By definition, each word w in X^* admits at least one factorization $(x_1, x_2, \ldots, x_n)$ whose elements are all in X. Such a factorization is called an *X-factorization*. We frequently use the pictorial representation of an X-factorization given in Figure 1.2.

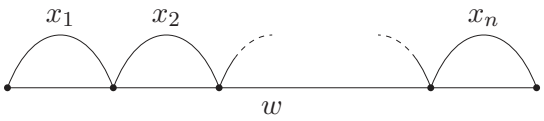

Figure 1.2 An X-factorization of w.

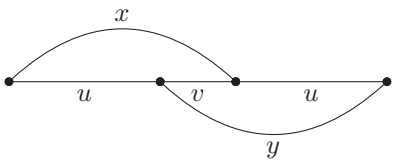

Figure 1.3 Two conjugate words x and y.

A word $x \in A^*$ is called *primitive* if it is not a power of another word. Thus x is primitive if and only if $x = y^n$ with $n \geq 0$ implies $x = y$. Observe that the empty word is not primitive.

Two words x, y are called *conjugate* if there exists words u, v such that $x = uv$, $y = vu$. (See Figure 1.3.) We frequently say that y is a conjugate of x. Two conjugate words are obtained from each other by a cyclic permutation. More precisely, let γ be the function from A^* into itself defined by

$$\gamma(1) = 1 \quad \text{and} \quad \gamma(av) = va \tag{1.3}$$

for $a \in A$, $v \in A^*$. It is clearly a bijection from A^* onto itself. Two words x and y are conjugate if and only if there exists an integer $n \geq 0$ such that

$$x = \gamma^n(y).$$

This easily implies that the conjugacy relation is an equivalence relation. A *conjugacy class* is a class of this equivalence relation. A conjugacy class is also called a *necklace*. The length of a necklace is the length of the words in the conjugacy class. A necklace is *primitive* if each word in the conjugacy class is primitive.

Proposition 1.3.1 *Each nonempty word is a power of a unique primitive word.*

Proof. Let $x \in A^+$ and δ be the restriction of the function γ defined by (1.3) to the conjugacy class of x. Then $\delta^k = 1$ if and only if x is a power of a word of length dividing k.

Let p be the order of δ, that is, the gcd of the integers k such that $\delta^k = 1$. Since $\delta^p = 1$, there exists a word r of length p such that $x = r^e$ with $e \geq 1$. The word r is primitive, otherwise there would be a word s of length q dividing p such that $r \in s^*$, which in turn implies that $x \in s^*$, contrary to the definition of p. This proves the existence of the primitive word. To show uniqueness, consider a word $t \in A^*$ such that $x \in t^*$ and let $k = |t|$. Since $\delta^k = 1$, the integer k is a multiple of p. Consequently $t \in r^*$. Thus, if t is primitive, we have $t = r$. □

Table 1.1 *The number $\ell_n(k)$ of primitive conjugacy classes over a k-letter alphabet.*

n	1	2	3	4	5	6	7	8	9	10	11	12
$\ell_n(2)$	2	1	2	3	6	9	18	30	56	99	186	335
$\ell_n(3)$	3	3	8	18	48	116	312	810				
$\ell_n(4)$	4	6	20	60	204	670						
$\ell_n(5)$	5	10	40	150	624							

Let $x \in A^+$. The unique primitive word r such that $x = r^n$ for some integer n is called the *root* of x. The integer n is the *exponent* of x.

Proposition 1.3.2 *Two nonempty conjugate words have the same exponent and their roots are conjugate.*

Proof. Let $x, y \in A^+$ be two conjugate words, and let i be an integer such that $y = \gamma^i(x)$. Set r and s be the roots of x and y respectively and let n be the exponent of x. Then

$$y = \gamma^i(r^n) = (\gamma^i(r))^n.$$

This shows that $\gamma^i(r) \in s^*$. Interchanging the roles of x and y, we have $\gamma^j(s) \in r^*$. It follows that $\gamma^i(r) = s$ and $\gamma^j(s) = r$. Thus r and s are conjugate and consequently x and y have the same exponent. □

Proposition 1.3.3 *All words in a conjugacy class have the same exponent. If C is a conjugacy class of words of length n with exponent e, then*

$$\mathrm{Card}(C) = n/e.$$

Proof. Let $x \in A^n$ and C be its conjugacy class. Let δ be the restriction of γ to C and p be the order of δ. The root of x is the word r of length p such that $x = r^e$. Thus $n = pe$. Now $C = \{x, \delta(x) \dots, \delta^{p-1}(x)\}$. These elements are distinct since p is the order of δ. Thus $\mathrm{Card}(C) = p$. □

We now compute the number of conjugacy classes of words of given length over a finite alphabet. Let A be an alphabet with k letters. For all $n \geq 1$, the number of conjugacy classes of primitive words in A^* of length n is denoted by $\ell_n(k)$. The notation is justified by the fact that this number depends only on k and not on A.

The first values of this function, for $k = 2, 3, 4$, are given in Table 1.1. Clearly $\ell_n(1) = 1$ if $n = 1$, and $\ell_n(1) = 0$ otherwise. Now for $n \geq 1$

$$k^n = \sum_{d|n} d\, \ell_d(k), \tag{1.4}$$

where d runs over the divisors of n. Indeed, every word of length n belongs to exactly one conjugacy class of wordsof length n. Each class has $d = n/e$ elements, where

e is the exponent of its words. Since there are as many classes whose words have exponent n/e as there are classes of primitive words of length $d = n/e$, the formula follows.

We can obtain an explicit expression for the numbers $\ell_n(k)$ by using the classical technique of Möbius inversion which we now recall.

The *Möbius function* is the function $\mu : \mathbb{N} \setminus 0 \to \mathbb{N}$ defined by $\mu(1) = 1$ and

$$\mu(n) = \begin{cases} (-1)^i & \text{if } n \text{ is the product of } i \text{ distinct prime numbers,} \\ 0 & \text{otherwise.} \end{cases}$$

Proposition 1.3.4 (Möbius inversion formula) *Let α, β be two functions from $\mathbb{N} \setminus 0$ into $\mathbb{N}$. Then*

$$\alpha(n) = \sum_{d|n} \beta(d) \quad (n \geq 1) \tag{1.5}$$

if and only if

$$\beta(n) = \sum_{d|n} \mu(d)\alpha(n/d) \quad (n \geq 1). \tag{1.6}$$

Proof. Let $\mathcal{S}$ be the set of functions from $\mathbb{N} \setminus 0$ into $\mathbb{N}$. Define a product on $\mathcal{S}$ by setting, for $f, g \in \mathcal{S}$

$$f * g(n) = \sum_{n=de} f(d)g(e).$$

It is easily verified that $\mathcal{S}$ is a commutative monoid for this product. Its neutral element is the function I taking the value 1 for $n = 1$ and 0 elsewhere.

Let $\iota \in \mathcal{S}$ be the constant function with value 1. Let us verify that

$$\iota * \mu = I. \tag{1.7}$$

Indeed $\iota * \mu(1) = 1$; for $n \geq 2$, let $n = p_1^{k_1} p_2^{k_2} \cdots p_m^{k_m}$ be the prime decomposition of n. If d divides n, then $\mu(d) \neq 0$ if and only if

$$d = p_1^{\ell_1} p_2^{\ell_2} \dots p_m^{\ell_m}$$

with all $\ell_i = 0$ or 1. Then $\mu(d) = (-1)^t$ with $t = \sum_{i=1}^m \ell_i$. It follows that

$$\iota * \mu(n) = \sum_{d|n} \mu(d) = \sum_{t=0}^{m} (-1)^t \binom{m}{t} = 0.$$

Now let $\alpha, \beta \in \mathcal{S}$. Then Formula (1.5) is equivalent to $\alpha = \iota * \beta$ and Formula (1.6) is equivalent to $\beta = \mu * \alpha$. By (1.7) these two formulas are equivalent. □

Proposition 1.3.5 *The number of conjugacy classes of primitive words of length n over an alphabet with k letters is*

$$\ell_n(k) = \frac{1}{n}\sum_{d|n} \mu(n/d)k^d.$$

Proof. This is immediate from Formula (1.4) by Möbius inversion. □

A word $w \in A^+$ is called *unbordered* if no proper nonempty prefix of w is a suffix of w. In other words, w is unbordered if and only if $w \in uA^+ \cap A^+u$ implies $u = 1$. If w is unbordered, then

$$wA^* \cap A^*w = wA^*w \cup w.$$

The following property holds.

Proposition 1.3.6 *Let A be an alphabet with at least two letters. For each word $u \in A^+$, there exists $v \in A^*$ such that uv is unbordered.*

Proof. Let a be the first letter of u, and let $b \in A \setminus a$. Let us verify that the word $w = uab^{|u|}$ is unbordered. A nonempty prefix t of w starts with the letter a. It cannot be a suffix of w unless $|t| > |u|$. But then we have $t = sab^{|u|}$ for some $s \in A^*$, and also $t = uab^{|s|}$. Thus $|s| = |u|$, hence $t = w$. □

Let A be an alphabet. The *free group* $A^{\odot}$ on A is defined as follows: Let $\bar{A}$ be an alphabet in bijection with A and disjoint from A. Denote by $a \mapsto \bar{a}$ the bijection from A onto $\bar{A}$. This notation is extended by setting, for all $a \in A \cup \bar{A}$, $\bar{\bar{a}} = a$. Let δ be the symmetric relation defined for $u, v \in (A \cup \bar{A})^*$ and $a \in A \cup \bar{A}$ by

$$ua\bar{a}v \equiv uv \bmod \delta.$$

Let ρ be the reflexive and transitive closure of δ. Then ρ is a congruence. The quotient monoid $A^{\odot} = (A \cup \bar{A})^*/\rho$ is a group. Indeed, for all $a \in A \cup \bar{A}$,

$$a\bar{a} \equiv 1 \bmod \rho.$$

Thus the images of the generators are invertible in $A^{\odot}$. This shows that all elements in $A^{\odot}$ are invertible.

Let A be an alphabet. The *free commutative monoid* $A^{\oplus}$ on A is the quotient of A^* by the congruence generated by the pairs (ab, ba) for $a, b \in A$, $a \neq b$. If $A = \{a_1, \ldots, a_k\}$, then the monoid $A^{\oplus}$ can be identified with the additive monoid $\mathbb{N}^k$ through the map $a_1^{n_1}a_2^{n_2}\cdots a_k^{n_k} \mapsto (n_1, n_2, \ldots, n_k)$.

We denote by $\alpha(w)$ the commutative image of a word $w \in A^*$. It is the element of $A^{\oplus}$ defined by

$$\alpha(w) = \prod_{a \in A} a^{|w|_a}.$$

Observe that α is a monoid morphism from A^* onto $A^{\oplus}$.