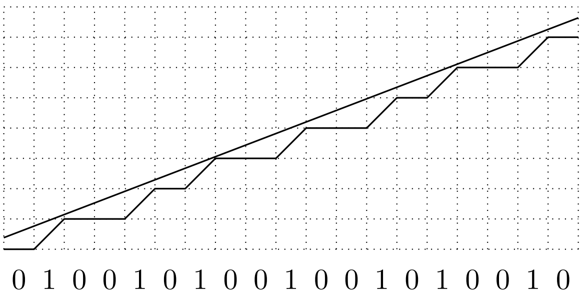
Combinatorics on Words

Aleksi Saarela

2020



Contents

Preface

Combinatorics on words is an area of discrete mathematics that studies combinatorial properties of finite and infinite sequences of symbols. It has applications in many fields of mathematics and computer science. The goal of these lecture notes is to provide an introduction to this area.

Not many prerequisites are needed. The reader is assumed to be familiar with things like vectors, matrices, and modular arithmetic, and know the definitions of a group and a homomorphism. Of course, because this is an advanced level course, some mathematical maturity is expected, and the emphasis is on theorems and proofs. Some theorems, examples and exercises require a little bit of knowledge on other topics such as graph theory or programming. These can be skipped if necessary.

The lecture notes were written for a half-semester advanced course in the University of Turku. They are partially based on old full semester lecture notes by Juhani Karhumäki. Several books mentioned in the bibliography have also been used as references. There are connections to many other courses such as *Automata and Formal Languages*, *Semigroup Theory*, and *Symbolic Dynamics*.

Chapter 1

Words and concatenation

1.1 Basic definitions

An *alphabet* is a nonempty finite set. The elements of an alphabet are called *letters* (or *symbols*). Alphabets of size k are called *k-ary*. In the cases $k = 1$, $k = 2$, $k = 3$, we can use the terms *unary*, *binary*, *ternary*, respectively. Most often we use Σ to denote an alphabet.

Example 1.1.1. Some typical examples of alphabets are the alphabet of binary digits $\{0, 1\}$, the alphabet of decimal digits $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$, and the alphabet of the 26 English letters $\{\mathsf{a}, \ldots, \mathsf{z}\}$.

A *word* over an alphabet Σ is a finite sequence of elements of Σ. A word $(a_1, \ldots, a_n)$ is usually written without the commas and parentheses as $a_1 \cdots a_n$. We allow the case $n = 0$, which gives the *empty word*, denoted by ε. The set of all words over Σ is denoted by Σ^*.

The *length* of a word $w = a_1 \cdots a_n$ is n and it is denoted by $|w|$. The set of words of length n over Σ is denoted by Σ^n. If Σ is k-ary, then there are k^n words in Σ^n. We identify words of length one with letters, that is, $\Sigma^1 = \Sigma$.

A word is called *k-ary* if it contains at most k different letters. Thus a word over a k-ary alphabet is k-ary, but sometimes it might be also j-ary for some $j < k$. Every k-ary word is also l-ary for all $l \geq k$.

Example 1.1.2. Let $\Sigma = \{a, b\}$. Then $\Sigma^0 = \{\varepsilon\}$, $\Sigma^1 = \{a, b\}$, $\Sigma^2 = \{aa, ab, ba, bb\}$, and $\Sigma^3 = \{aaa, aab, aba, abb, baa, bab, bba, bbb\}$.

The *concatenation* or *product* of words u and v, denoted by $u \cdot v$ or uv, is the word consisting of the letters of u followed by the letters of v. In other words, if $u = a_1 \cdots a_m$ and $v = b_1 \cdots b_n$, then $uv = a_1 \cdots a_m b_1 \cdots b_n$. It is clear that $|uv| = |u| + |v|$.

Concatenation is associative, that is, $(uv)w = u(vw)$ for all words u, v, w. Therefore, we can write the concatenation of several words without parentheses. The empty word acts as a neutral element, that is, $\varepsilon w = w \varepsilon = w$ for all words w. If the alphabet is not unary, then there exist words u, v such that $uv \neq vu$, so concatenation is not commutative. For example, if a and b are distinct letters, then $ab \neq ba$. A characterization of pairs of words (x, y) such that $xy = yx$ is proved later in Theorem 1.2.2.

Example 1.1.3. Consider the words $u = \mathsf{side}$ and $v = \mathsf{road}$ over the alphabet $\{\mathsf{a}, \ldots, \mathsf{z}\}$. Then $uv = \mathsf{sideroad}$, $vu = \mathsf{roadside}$ and $uuvu = \mathsf{sidesideroadside}$.

Let w be a word. A word u is a

- *factor* of w if $w = xuy$ for some words x, y,
- *prefix* of w if $w = uy$ for some word y,
- *suffix* of w if $w = xu$ for some word x,
- *subword* of w if $w = x_0 u_1 x_1 \cdots u_k x_k$ and $u = u_1 \cdots u_k$ for some words $x_0, \ldots, x_k$ and $u_1, \ldots, u_k$.

A factor (prefix, suffix) u of w is a *proper factor* (*proper prefix*, *proper suffix*, respectively) if $u \neq w$.

If w is a word of length n and $k \in \{0, \ldots, n\}$, then w has exactly one prefix (suffix) of length k, denoted by $\mathrm{pref}_k(w)$ ($\mathrm{suff}_k(w)$, respectively). On the other hand, the number of factors of a word depends not only on the length but also the structure of the word. For example, if a and b are distinct letters, then aa has three factors, but ab has four. If u and v are prefixes (suffixes) of w, then one of u and v is a prefix (suffix, respectively) of the other.

Example 1.1.4. The word $0122 \in \{0, 1, 2\}^*$ has

- ten factors ε, 0, 1, 2, 01, 12, 22, 012, 122, 0122,
- five prefixes ε, 0, 01, 012, 0122,
- five suffixes ε, 2, 22, 122, 0122,
- twelve subwords ε, 0, 1, 2, 01, 02, 12, 22, 012, 022, 122, 0122.

Remark 1.1.5. Sometimes in the literature, factors are called subwords, and subwords are called *scattered subwords* or *sparse subwords*.

Powers of a word w are defined in the usual way: $w^0 = \varepsilon$ and $w^{n+1} = w^n \cdot w$ for all $n \in \mathbb{Z}_{\geq 0}$. Then w^n is called the *nth power* or *n-power* of w. The words w^2 and w^3 can be called the *square* and *cube* of w, respectively. Clearly $w^m w^n = w^{m+n}$ and $(w^m)^n = w^{mn}$ for all $m, n \in \mathbb{Z}_{\geq 0}$.

Example 1.1.6. Let us consider words over the alphabet $\{\mathsf{a}, \ldots, \mathsf{z}\}$. The English word $\mathsf{hotshots} = (\mathsf{hots})^2$ is a square. The Finnish word $\mathsf{kokoko} = (\mathsf{ko})^3$ is a cube.

Let Σ and Γ be alphabets. A mapping $h : \Sigma^* \to \Gamma^*$ is a *morphism* (or *homomorphism*) if $h(uv) = h(u)h(v)$ for all $u, v \in \Sigma^*$. It follows that if $h : \Sigma^* \to \Gamma^*$ is a morphism, then $h(u_1 \cdots u_n) = h(u_1) \cdots h(u_n)$ for all $u_1, \ldots, u_n \in \Sigma^*$.

Every mapping $h_1 : \Sigma \to \Gamma^*$ can be extended to a morphism $h : \Sigma^* \to \Gamma^*$ in a unique way by the formula $h(a_1 \cdots a_n) = h_1(a_1) \cdots h_1(a_n)$ for all $a_1, \ldots, a_n \in \Sigma$. For this reason, when defining a morphism, it is enough to specify the images of the letters.

Example 1.1.7. Consider the morphism

$$h : \{a, b, c\}^* \to \{a, b, c\}^*, \ h(a) = abca, \ h(b) = a, \ h(c) = \varepsilon.$$

Then

$$h(abbc) = h(a)h(b)h(b)h(c) = abca \cdot a \cdot a \cdot \varepsilon = abcaaa.$$

Example 1.1.8. The mapping

$$f : \{a, b\}^* \to \{a, b\}^*,\ f(w) = (ab)^{|w|}$$

is a morphism, because

$$f(uv) = (ab)^{|uv|} = (ab)^{|u|+|v|} = (ab)^{|u|}(ab)^{|v|} = f(u)f(v)$$

for all $u, v \in \Sigma^*$. The mapping

$$g : \{a, b\}^* \to \{a, b\}^*,\ g(w) = w^2$$

is not a morphism, because

$$g(ab) = abab \neq aabb = g(a)g(b).$$

We conclude this section with some examples about how words can be used in different areas. The notation defined in Example 1.1.9 is used later in these lecture notes. Otherwise, the purpose of these examples is mostly just to give perspective, and they are not necessary for the material that follows.

Example 1.1.9 (Automata and formal languages)**.** A *language* (or *formal language*) is a set of words. Theory of formal languages is closely related to combinatorics on words. If A and B are languages and $n \in \mathbb{Z}_{\geq 0}$, we can use the following notation:

$$\begin{aligned} AB &= \{uv \mid u \in A,\ v \in B\}, \\ A^n &= \{u_1 \cdots u_n \mid u_1, \ldots, u_n \in A\}, \\ A^* &= \bigcup_{n=0}^{\infty} A^n, \\ A^+ &= \bigcup_{n=1}^{\infty} A^n. \end{aligned}$$

This is consistent with how we defined Σ^n and Σ^* earlier. If w is a word, we can use the notation

$$\begin{aligned} wA &= \{w\}A = \{wu \mid u \in A\}, \\ Aw &= A\{w\} = \{uw \mid u \in A\}, \\ w^* &= \{w\}^* = \{w^n \mid n \in \mathbb{Z}_{\geq 0}\}, \\ w^+ &= \{w\}^+ = \{w^n \mid n \in \mathbb{Z}_+\}. \end{aligned}$$

The language AB can be called the *concatenation* of A and B, and the language A^* can be called the *Kleene star* of A. A language is *regular* if it can be constructed from finite languages by repeatedly applying the operations union, concatenation and Kleene star. Regular languages have many equivalent definitions. For example, they can be defined as the languages recognized by *deterministic finite automata.*

Example 1.1.10 (Algebra)**.** Associativity and existence of a neutral element mean that, using algebraic terminology, $(\Sigma^*, \cdot)$ is a *monoid.* It is not a group, because nonempty words do not have inverses. Theory of monoids can be useful when studying words, and words can be useful when studying monoids or groups. We formally define and consider monoids in Section 2.3.

Example 1.1.11 (Number theory)**.** Let $B \geq 2$ be an integer. Every nonnegative integer n can be represented in the form

$$n = a_k B^k + \cdots + a_0 B^0,$$

where $k \geq 0$ and $a_0, \ldots, a_k \in \{0, \ldots, B-1\}$. Then the word $a_k \cdots a_0$ over the alphabet $\Sigma = \{0, \ldots, B-1\}$ can be called a *B-ary representation* of n. Often we require that $a_k \neq 0$ to make the representation unique. Conversely, we can define a function

$$N_B : \Sigma^* \to \mathbb{Z}_{\geq 0}, \; N_B(a_k \cdots a_0) = a_k B^k + \cdots + a_0 B^0$$

that maps a word to the number it represents. Then

$$N_B(uv) = N_B(u) \cdot B^{|v|} + N_B(v)$$

for all $u, v \in \Sigma^*$. Representations of real numbers can be viewed as infinite words, which are defined and studied in Chapter 3.

Example 1.1.12 (Programming)**.** In programming languages, words are usually called *strings* and letters are called *characters*. Concatenation is often denoted by the symbol +. In some languages, characters and strings are different types, so for example, the character a is not the same as the string a. In some other languages, characters are simply strings of length one, so the character a is the same as the string a. As stated above, we use the latter convention.

Example 1.1.13 (Algorithms)**.** The study of string algorithms in computer science is closely related to combinatorics on words. Finding a longest common factor of two words is an example of a nontrivial algorithmic problem. It can be solved by dynamic programming, or more efficiently by using a data structure called *suffix tree*. For example, an entire book can be viewed as a single word. Then the longest common factor of the books *Alice's Adventures in Wonderland* by Lewis Carroll and *Metamorphosis* by Franz Kafka is " he could think of nothing ". If *Metamorphosis* is replaced by the book *Through the Looking-Glass* by Lewis Carroll, then the longest common factor is " in one hand and a piece of bread-and-butter in the other. ".

Example 1.1.14 (Encodings)**.** Let $\Sigma = \{\mathsf{a}, \ldots, \mathsf{z}\} \cup \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. We could try to represent the *Morse code* as a morphism $f : \Sigma^* \to \{\cdot, -\}^*$, where $\cdot$ represents a short and $-$ a long signal. Then, for example, $f(\mathsf{e}) = \cdot$, $f(\mathsf{t}) = -$, and $f(\mathsf{a}) = \cdot-$. However, $f(\mathsf{et}) = f(\mathsf{a})$, so this morphism is not injective. This means that if we receive an encoded message $\cdot-$, we do not know whether it is an encoding of a or et. In practise, this is solved by adding a pause after the encoding of each letter. This idea can be represented by the morphism $g : \Sigma^* \to \{\cdot, -, \square\}^*$ defined by $g(a) = f(a)\square$ for all $a \in \Sigma$, where $\square$ represents a pause. The morphism g is injective, so from an encoded message $g(w)$ we can always deduce the original message w. This is related to the definition of a *code* in Section 2.3.

Example 1.1.15 (Bioinformatics)**.** A DNA strand consists of a sequence of the four bases *adenine*, *cytosine*, *guanine* and *thymine*. These strands are often modelled as words over the alphabet $\{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$.

Exercises

1.1.1. How many words are there in $\{a, b, c\}^n$ that contain all three letters at least once?

1.1.2. How many words in $\{a, b\}^5$ contain aab as a factor? How many words in $\{a, b\}^5$ contain aba as a factor? Can you describe the "reason" why these numbers are different?

*1.1.3.** (Graph theory.) Let $k, n \geq 1$ and let Σ be a k-ary alphabet. The n-dimensional k-ary *De Bruijn graph* is the directed graph (with loops) whose set of vertices is Σ^n and there is an edge from u to v if and only if $\mathrm{suff}_{n-1}(u) = \mathrm{pref}_{n-1}(v)$. Show that this graph is Eulerian. Conclude that there exists a word, called a *De Bruijn word*, of length $k^{n+1} + n$ having every word in Σ^{n+1} as a factor.

*1.1.4.** What is the maximum number of factors a binary word of length 10 can have?

1.1.5. Are the following functions morphisms:

$$f : \{a, b\}^* \to \{a\}^*,\ f(w) = a^{|w|+1},$$
$$g : \{a, b\}^* \to \{a, b\}^*,\ g(a_1 \cdots a_n) = a_1^2 \cdots a_n^2,$$

where $a_1, \dots, a_n \in \{a, b\}$.

1.1.6. Consider the following properties of a word x or a pair of words (x, y): $x = \varepsilon$, $x \neq \varepsilon$, x is a square, $|x| = |y|$, x is a factor of y, x is a prefix of y. Which of these properties are preserved by morphisms in the sense that if x or (x, y) has the property, then also $h(x)$ or $(h(x), h(y))$ has the property for all morphisms h?

1.1.7. Let $B \geq 2$ be an integer and $\Sigma = \{0, \dots, B - 1\}$ an alphabet. We can define a function

$$\overline{N}_B : \Sigma^* \to \mathbb{Z}_{\geq 0},\ \overline{N}_B(a_0 \cdots a_k) = a_0 B^0 + \cdots + a_k B^k$$

that maps a word to the number it represents in *reverse B-ary notation.* Give a formula for $\overline{N}_B(uv)$, where $u, v \in \Sigma^*$ (similar to the one in Example 1.1.11).

1.1.8. Many text editors have a "search and replace" feature. If w is a word representing the text and we want to replace the first occurrence of a word u in w by a word v, then this can be formally described by the function

$$\mathrm{replace}(w, u, v) = \begin{cases} xvy & \text{if } w = xuy \text{ and } u \text{ occurs in } xu \text{ only as a suffix,} \\ w & \text{if } u \text{ does not occur in } w. \end{cases}$$

Let w_0, u, v be fixed words and let $w_{i+1} = \mathrm{replace}(w_i, u, v)$ for all $i \geq 0$. What does the sequence $w_0, w_1, w_2, \dots$ look like if (w_0, u, v) is (a, a, aa) or (a^4b, ab, ba) or $(ababbaba, abba, baab)$?

*1.1.9.** With the notation of the previous exercise, show that if $|u| = |v|$, then $w_{i+1} = w_i$ for all large enough i.

*1.1.10.** (Programming.) Write a program that takes as input a word and a morphism (choose a suitable way to represent the morphism) and returns the image of the word under the morphism.

1.2 Commutation, primitivity and conjugacy

The following result is known as *Levi's lemma*. It is frequently used without specifically referring to it.

Lemma 1.2.1. *Let u, v, x, y be words and $uv = xy$. Then exactly one of the following is true:*

1. *$|u| < |x|$ and $x = ut$ and $v = ty$ for some nonempty word t.*
2. *$|u| = |x|$ and $u = x$ and $v = y$.*
3. *$|u| > |x|$ and $u = xt$ and $y = tv$ for some nonempty word t.*

Proof. Let $uv = xy = a_1 \cdots a_n$, where $a_1, \dots, a_n$ are letters. There exist numbers k, m such that $u = a_1 \cdots a_k$ and $x = a_1 \cdots a_m$. If $k < m$, then the first condition is satisfied with $t = a_{k+1} \cdots a_m$. If $k = m$, then the second condition is satisfied. If $k > m$, then the third condition is satisfied with $t = a_{m+1} \cdots a_k$. □

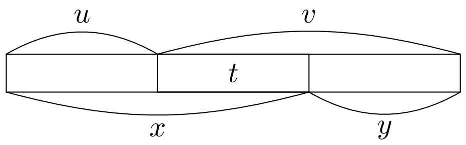


Figure 1.1: Illustration of the first condition of Levi's lemma. These kinds of diagrams are frequently used in combinatorics on words.

The second condition of Levi's lemma is often combined with the first or the last one. For example, the combination of the first two conditions could be stated as follows: $|u| \leq |x|$ and $x = ut$ and $v = ty$ for some word t.

Letting $u = x$ in Levi's lemma gives the following cancellation property: If $uv = uy$, then $v = y$. Similarly, if $uv = xv$, then $u = x$.

We saw earlier that concatenation is not a commutative operation. However, there are some words that commute with each other. Trivially, if w is any word and m, n any nonnegative integers, then $w^m \cdot w^n = w^n \cdot w^m$. Next we prove that there are no other examples of commuting words, that is, words commute if and only if they are powers of a common word.

Theorem 1.2.2. *Let x and y be words. Then $xy = yx$ if and only if there exists a word w such that $x, y \in w^*$.*

Proof. The "if" direction is clear. The "only if" direction is proved by induction on $|xy|$. If $|xy| = 0$, then $x = y = \varepsilon$ and the claim is clear. Let $|xy| > 0$. By symmetry, we can assume that $|x| \leq |y|$. By Levi's lemma, $y = xt$ for some word t. From $xy = yx$ it follows that $xxt = xtx$, and cancelling x from the left gives $xt = tx$. If $|x| = 0$, then $x, y \in y^*$. If $|x| > 0$, then $|xy| = |xxt| > |xt|$, so it follows from $xt = tx$ and the induction hypothesis that $x, t \in w^*$ for some word w, and then also $y = xt \in w^*$. This completes the induction. □

A word is *primitive* if it is not an nth power of any word for any integer $n \geq 2$. A primitive word p is a *primitive root* of a word w if $w \in p^+$. We prove in Theorem 1.2.4 that every nonempty word has a unique primitive root. The primitive root of w is denoted by $\rho(w)$.

Example 1.2.3. Consider words over the alphabet $\{a, b\}$. Of the words of length at most four, only ε, aa, bb, aaa, bbb, $aaaa$, $bbbb$, $abab$, $baba$ are not primitive. The empty word is not primitive because $\varepsilon = \varepsilon^2$. The primitive root of $abab$ is ab.

Theorem 1.2.4. *Every nonempty word has a unique primitive root.*

Proof. First we prove that a nonempty word w has at least one primitive root. Let p be a shortest word such that $w \in p^*$ (p exists because $w \in w^*$). If p is not primitive, then $p = q^n$ for some word q and $n \geq 2$, and then $w \in q^*$, which is a contradiction because $|q| = |p|/n < |p|$. Thus p is primitive and therefore a primitive root of w.

Then we prove that any two primitive roots of w must be equal. Let $w = p^m = q^n$, where p and q are primitive and $m, n \geq 1$. By symmetry, we can assume that $|p| \leq |q|$. By Levi's lemma, $q = pt$ for some word t. Then $p^m = (pt)^n = p(tp)^{n-1}t$. Cancelling p from the left gives $p^{m-1} = (tp)^{n-1}t$, and then multiplying by p from the right gives $p^m = (tp)^n$. Thus $(pt)^n = p^m = (tp)^n$, so $pt = \mathrm{pref}_{|pt|}(p^m) = tp$. By Theorem 1.2.2, $p, t \in r^*$ for some word r. Then also $q = pt \in r^*$. Because p and q are primitive, it must be $p = r = q$. This completes the proof. □

The following result is often useful.

Lemma 1.2.5. *Let p, x, y be words. If p is primitive and $pp = xpy$, then $x = \varepsilon$ or $y = \varepsilon$.*

Proof. Clearly $|x| \leq |xy| = |p|$, so by Levi's lemma, $p = xt$ for some word t. Then $xtxt = xxty$ and thus $txt = xty$, and then $tx = xt$. By Theorem 1.2.2, $x, t \in w^*$ for some word w. Then also $p = xt \in w^*$. Because p is primitive, it must be $p = w$, and thus either $x = w$ and $t = \varepsilon$, or $x = \varepsilon$ and $t = w$. If $x = w = p$, then $y = \varepsilon$. This completes the proof. □

Words u and v are *conjugates* if there exist words p, q such that $u = pq$ and $v = qp$. In other words, if $a_1, \ldots, a_n \in \Sigma$, then the conjugates of the word $a_1 \cdots a_n$ are the words $a_i \cdots a_n \cdot a_1 \cdots a_{i-1}$, where $i \in \{1, \ldots, n\}$.

By the next theorem, words x and y are conjugates if and only if $xz = zy$ for some word z.

Theorem 1.2.6. *Let $x, y, z \in \Sigma^*$. Then $xz = zy$ if and only if $x = y = \varepsilon$ or there exist words $p, q \in \Sigma^*$ and $k \in \mathbb{Z}_{\geq 0}$ such that $x = pq$, $y = qp$, and $z = (pq)^k p$.*

Proof. The "if" direction is clear. To prove the "only if" direction, let $xz = zy$. If one of x, y is empty, then both of them are, and the claim is true, so let $x \neq \varepsilon \neq y$. We prove the claim by induction on $|z|$. The case $|z| = 0$ is clear. Let $|z| > 0$. If $|z| \leq |x|$, then $x = zt$ for some word t, and then $tz = y$, so we can let $p = z$, $q = t$, $k = 0$. If $|x| < |z|$, then $z = xt$ for some word t, and then $xt = ty$. Because $|t| = |z| - |x| < |z|$, it follows from the induction hypothesis that $x = pq$, $y = qp$, and $t = (pq)^k p$ for some p, q, k, and then $z = (pq)^{k+1} p$. This completes the induction. □

Conjugacy is an equivalence relation. The next result gives the size of the equivalence classes, which can also be called conjugacy classes.

Theorem 1.2.7. *A nonempty word w has exactly $|\rho(w)|$ conjugates.*

Proof. Let $\rho(w) = r$ and $w = r^n$. Let $q_i = \mathrm{suff}_i(r)$ and $p_i = \mathrm{pref}_{|r|-i}(r)$ for $i \in \{0, \dots, |r| - 1\}$. Then the conjugates of r are the words $q_i p_i$. If $q_i p_i = q_j p_j$ for some $i \leq j$, then $(q_i p_i)^2 q_i = (q_j p_j)^2 q_i$ and thus $q_j = q_i t$ for some word t. By cancelling q_i and using the equalities $p_i q_i = p_j q_j = r$, we get $r^2 = trp_j q_i$. By Lemma 1.2.5, $t = \varepsilon$ and thus $i = j$. We have proved that the $|r|$ words $q_i p_i$ are all different. The conjugates of w are the $|r|$ words $(q_i p_i)^n$, which are also all different. This completes the proof. □

In the next theorem, we study connections between primitivity and conjugacy.

Theorem 1.2.8. *Let u and v be nonempty words.*

1. *If u and v are conjugates, then u is primitive if and only if v is primitive.*
2. *If u and v are conjugates, then $\rho(u)$ and $\rho(v)$ are conjugates.*
3. *If $\rho(u)$ and $\rho(v)$ are conjugates and $|u| = |v|$, then u and v are conjugates.*

Proof. Let $u = \rho(u)^n$.

1. We assume that $u = pq$ is not primitive, that is, $n \geq 2$, and show that its conjugate $v = qp$ is not primitive. It must be $p = \rho(u)^k p'$, $q = q' \rho(u)^{n-k-1}$ for some integer $k \geq 0$ and words p', q' such that $\rho(u) = p'q'$. Then $qp = (q'p')^n$ is not primitive. Similarly, if v is not primitive, then u is not primitive.
2. If u and v are conjugates, then $u = pq$ and $v = qp$ for some words p, q, and then $p = \rho(u)^k p'$, $q = q' \rho(u)^{n-k-1}$ for some integer $k \geq 0$ and words p', q' such that $\rho(u) = p'q'$. Then $v = (q'p')^n$ and $q'p'$ is primitive by Claim 1, so $\rho(v) = q'p'$.
3. If $\rho(u)$ and $\rho(v)$ are conjugates, then $\rho(u) = pq$ and $\rho(v) = qp$ for some words p, q, and then $u = (pq)^n$ and $v \in (qp)^*$. If $|u| = |v|$, it must be $v = (qp)^n$. □

Exercises

1.2.1. Let x, y, z be words and $xyx = yxz$. Show that x, y, z are powers of a common word.

1.2.2. Let x and y be words and $xxyy = yxyx$. Show that x and y are powers of a common word.

1.2.3. Let x and y be nonempty words. Show that there exist $m, n \geq 1$ such that $x^m = y^n$ if and only if $\rho(x) = \rho(y)$.

1.2.4. Let p, x, y be words and m, n positive integers. Let p be primitive, $p^m x = yp^n$, and $|y| \leq (m-1)|p|$. Show that $x, y \in p^*$.

1.2.5. Let $n \geq 1$. Show that words u and v are conjugates if and only if there exists a word w such that $u^n w = wv^n$.

*1.2.6.** (Number theory.) Let $k, n \geq 1$ and let Σ be a k-ary alphabet. Show that the number of primitive words in Σ^n is

$$\sum_{d|n} \mu(n/d) k^d,$$

where μ is the Möbius function. Conclude that if $k \geq 2$, then the proportion of words in Σ^n that are primitive approaches 1 as $n \to \infty$.

1.3 Palindromes, anagrams and alphabetical order

Many people are familiar with palindromes, anagrams and alphabetical order from a non-mathematical context. All of these concepts are important in combinatorics on words. In this section, we define them formally and prove some results related them.

The *reverse* of a word $w = a_1 \cdots a_n$, where $a_1, \dots, a_n \in \Sigma$, is $w^R = a_n \dots a_1$. Clearly $(w^R)^R = w$, and if u and v are words, then $(uv)^R = v^R u^R$. A word w is a *palindrome* if $w^R = w$.

Example 1.3.1. Let us consider words over the alphabet $\{\mathsf{a}, \dots, \mathsf{z}\}$. The English word racecar and the Finnish word saippuakauppias are palindromes.

In the next theorem, we study connections between primitivity, conjugacy, and reversal.

Theorem 1.3.2. *Let u and v be nonempty words.*

1. *The word u is primitive if and only if u^R is primitive.*
2. *If $v = u^R$, then $\rho(v) = \rho(u)^R$.*
3. *If $\rho(v) = \rho(u)^R$ and $|u| = |v|$, then $v = u^R$.*
4. *The word u is a palindrome if and only if $\rho(u)$ is a palindrome.*
5. *If u and v are conjugates, then u^R and v^R are conjugates.*

Proof. Let $u = \rho(u)^n$.

1. If u is not primitive, that is, $n \geq 2$, then $u^R = (\rho(u)^n)^R = (\rho(u)^R)^n$ is not primitive. Similarly, if u^R is not primitive, then u is not primitive.
2. If $v = u^R$, then $v = (\rho(u)^n)^R = (\rho(u)^R)^n$, and $\rho(u)^R$ is primitive by Claim 1, so $\rho(v) = \rho(u)^R$.
3. If $\rho(v) = \rho(u)^R$, then $v = (\rho(u)^R)^m = (\rho(u)^m)^R$ for some number m. If $|u| = |v|$, it must be $m = n$ and thus $v = u^R$.
4. If u is a palindrome, then $\rho(u) = \rho(u^R)$, and $\rho(u^R) = \rho(u)^R$ by Claim 2, so also $\rho(u)$ is a palindrome. If $\rho(u)$ is a palindrome, then $u^R = (\rho(u)^n)^R = (\rho(u)^R)^n = \rho(u)^n = u$, so also u is a palindrome.
5. If $u = pq$ and $v = qp$ for some words p, q, then $u^R = q^R p^R$ and $v^R = p^R q^R$. □

The number of occurrences of a factor x in a word w is denoted by $|w|_x$. Words $u, v \in \Sigma^*$ are *abelian equivalent* if $|u|_a = |v|_a$ for all $a \in \Sigma$. Abelian equivalent words can also be called *anagrams*. In other words, if $a_1, \dots, a_n \in \Sigma$, then the anagrams of the word $a_1 \cdots a_n$ are the words $a_{\sigma(1)} \cdots a_{\sigma(n)}$, where $\sigma : \{1, \dots, n\} \to \{1, \dots, n\}$ is a permutation.

Example 1.3.3. If we ignore whitespace, then "combinatorics on words" and "win or rot in abcd cosmos" are anagrams.

Abelian equivalence is an equivalence relation. The next result gives the size of the equivalence classes.

Theorem 1.3.4. *Let $\Sigma = \{a_1, \dots, a_k\}$ and $w \in \Sigma^*$. Let $|w| = n$ and $|w|_{a_i} = n_i$ for all i. Then there are*

$$\frac{n!}{n_1! \cdots n_k!}$$

words that are abelian equivalent to w.

Proof. If we are contructing a word of length n that is abelian equivalent to w, we can first choose which n_1 of the n positions contain a_1, then we can choose which n_2 of the remaining $n - n_1$ positions contain a_2, and so on. There are a total of

$$\binom{n}{n_1}\binom{n-n_1}{n_2}\cdots\binom{n-n_1-\cdots-n_{k-1}}{n_k} = \frac{n!}{n_1!\cdots n_k!}$$

ways to make the choices. □

Example 1.3.5. The word 0012 has $4!/(2! \cdot 1! \cdot 1!) = 12$ anagrams. They are

$$0012, 0021, 0102, 0120, 0201, 0210, 1002, 1020, 1200, 2001, 2010, 2100.$$

A relation $\leq$ on a set S is a *total order* if the following conditions are satisfied for all $x, y, z \in S$:

1. $x \leq y$ or $y \leq x$.
2. If $x \leq y$ and $y \leq x$, then $x = y$.
3. If $x \leq y$ and $y \leq z$, then $x \leq z$.

Naturally, we can use the notation $x < y$ if $x \leq y$ and $x \neq y$.

Assume that we have fixed a total order $\leq$ on Σ. The *lexicographic order* (or *alphabetical order*) $\leq_{\text{lex}}$ is defined as follows. If $u, v \in \Sigma^*$, then $u \leq_{\text{lex}} v$ if one of the following holds:

1. u is a prefix of v.
2. u has a prefix xa and v has a prefix xb for some word x and letters $a < b$.

The *radix order* $\leq_{\text{rad}}$ is defined as follows. If $u, v \in \Sigma^*$, then $u \leq_{\text{rad}} v$ if one of the following holds:

1. $|u| < |v|$.
2. $|u| = |v|$ and $u \leq_{\text{lex}} v$.

Theorem 1.3.6. *Both $\leq_{\text{lex}}$ and $\leq_{\text{rad}}$ are total orders on Σ^*.*

Proof. For any words u and v, exactly one of the following holds:

1. $u = v$.
2. u is a proper prefix of v.
3. v is a proper prefix of u.
4. There exist words x, y, z and distinct letters a, b such that $u = xay$ and $v = xbz$.

Based on this, the three conditions in the definition of a total order can be verified for both $\leq_{\text{lex}}$ and $\leq_{\text{rad}}$. Details are left as an exercise. □

A *Lyndon word* is a primitive word that is lexicographically smaller than all of its other conjugates. Lyndon words are not needed in the remaining part of these lecture notes, but they are used quite often in combinatorics on words. For example, the following theorem is well-known.

Theorem 1.3.7. *Every word w has a unique representation as a product $w = u_1 \cdots u_n$, where $n \geq 0$, $u_1, \dots, u_n$ are Lyndon words, and $u_n \leq_{\text{lex}} \cdots \leq_{\text{lex}} u_1$.*

Exercises

1.3.1. Find a meaningful sentence in some language that is a palindrome if we ignore capitalization, whitespace, and punctuation. An example would be "I prefer pi". Find also a meaningful anagram of your name or some other phrase (again, ignoring capitalization, whitespace, and punctuation).

1.3.2. Let Σ be a k-ary alphabet and $n \geq 0$. How many palindromes are there in Σ^n?

*1.3.3.** What is the maximum number of palindromic factors a word of length n can have? (Words with this maximum number of palindromic factors are called *rich*.)

*1.3.4.** Let u and v be words. Show that if u is a palindrome and uv is a palindrome, then v is a product of two palindromes.

1.3.5. Show that every nonunary word has a primitive anagram. Show that a word w has a nonprimitive anagram if and only if there exists $n \geq 2$ such that n divides $|w|_a$ for all letters a.

*1.3.6.** Let $k \geq 1$. Words $u, v \in \Sigma^*$ are *k-abelian equivalent* if $|u|_x = |v|_x$ for all $x \in \Sigma^*$ such that $|x| \leq k$. Let $|u|, |v| \geq k-1$. Show that u and v are k-abelian equivalent if and only if $\text{pref}_{k-1}(u) = \text{pref}_{k-1}(v)$ and $|u|_x = |v|_x$ for all $x \in \Sigma^k$.

1.3.7. Prove Theorem 1.3.6.

1.3.8. Let us say that u is the lexicographically previous word before v if $u <_{\text{lex}} v$ and there does not exist a word x such that $u <_{\text{lex}} x <_{\text{lex}} v$. The lexicographically next word after v is defined in a similar way. Does every nonempty word have a lexicographically previous and next word? What about if we use the radix order?

1.3.9. Let u, v, x, y be words. Show that $u \leq_{\text{lex}} v$ if and only if $xu \leq_{\text{lex}} xv$. Show that if $u \leq_{\text{lex}} v$ and u is not a prefix of v, then $ux \leq_{\text{lex}} vy$. Give an example of words u, v, x such that $u \leq_{\text{lex}} v$ but not $ux \leq_{\text{lex}} vx$.

1.3.10. Show that if u and v are words and $u \neq \varepsilon$, then uvu is not a Lyndon word.

*1.3.11.** Show that a nonempty word is a Lyndon word if and only if it is the lexicographically smallest of its nonempty suffixes.

1.3.12. Find the representation mentioned in Theorem 1.3.7 for the word 1010010100, where $0 \leq_{\text{lex}} 1$.

*1.3.13.** (Programming.) Write a program that, given a word w and a number n, checks whether w can be written as a product of n palindromes.

1.4 Periodicity

Let $w = a_1 \cdots a_n$, where $a_1, \ldots, a_n \in \Sigma$. A positive integer k is a *period* of w if $a_{i+k} = a_i$ for all $i \in \{1, \ldots, n-k\}$.

Example 1.4.1. Let $\Sigma = \{a, b\}$. The word $w = ababaababababaababa$ has periods 7, 12, 14, 16. Trivially, it also has period k for all $k \geq |w| = 17$.

If w is a word and $\alpha \in \mathbb{Q}_{\geq 0}$ is such that $\alpha|w| \in \mathbb{Z}$, then we can define a *fractional power* $w^\alpha = w^n u$, where $n = \lfloor \alpha \rfloor$ and u is the prefix of w of length $(\alpha - n)|w|$. Then w^α is called the *αth power* or *α-power* of w. Note that if $\alpha, \beta \in \mathbb{Q}_{\geq 0}$, then often $w^{\alpha+\beta} \neq w^\alpha w^\beta$ and $(w^\alpha)^\beta \neq w^{\alpha\beta}$ even if all the fractional powers here are defined. For example, if a and b are distinct letters, then $(ab)^{1/2}(ab)^{1/2} = aa \neq ab = (ab)^{1/2+1/2}$ and $((ab)^{1/2})^2 = aa \neq ab = (ab)^{(1/2)\cdot 2}$.

Example 1.4.2. Let us consider words over the alphabet $\{\mathsf{a}, \ldots, \mathsf{z}\}$. The French word $\mathsf{entente} = (\mathsf{ent})^{7/3}$ is a $(7/3)$-power. The Finnish word $\mathsf{taltalta} = (\mathsf{tal})^{8/3}$ is an $(8/3)$-power.

A nonempty word u is a *border* of a word w if u is a proper prefix and a proper suffix of w.

Example 1.4.3. The word w of Example 1.4.1 has borders a, aba, $ababa$, $ababaababa$.

In the next theorem we see that periods, borders and fractional powers are closely related.

Theorem 1.4.4. *Let w be a word and $k \geq 1$. The following are equivalent:*

1. *w has period k.*
2. *w is a fractional power of a word of length k.*
3. *$w = u^{|w|/k}$ for some word u of length k, and if $k \leq |w|$, then $u = \mathrm{pref}_k(w)$.*
4. *$k \geq |w|$ or w has a border of length $|w| - k$.*

Proof. Left as an exercise. □

Given two positive integers k and l, we are interested in words that have periods k and l. Every word that has period $\gcd(k, l)$ certainly has periods k and l, but if we exclude these words, then there are only finitely many words with periods k and l. We want to find out how long such words can be. We start with an example before giving the general answer.

Example 1.4.5. Let us try to find a word of length 10 with periods 4 and 7. Let $u = a_1 \cdots a_{10}$, where $a_1, \ldots, a_{10} \in \Sigma$. If u has periods 4 and 7, then

$$a_4 = a_8 = a_1 = a_5 = a_9 = a_2 = a_6 = a_{10} = a_3 = a_7,$$

so $u = a_1^{10}$, that is, u is unary.

On the other hand, we can find a nonunary word of length 9 with periods 4 and 7. Let $v = b_1 \cdots b_9$, where $b_1, \ldots, b_9 \in \Sigma$. If v has periods 4 and 7, then

$$b_4 = b_8 = b_1 = b_5 = b_9 = b_2 = b_6 \qquad \text{and} \qquad b_3 = b_7,$$

so $v = aabaaabaa$ for some $a, b \in \Sigma$. If $a \neq b$, then v does not have period 1, but it has periods 4 and 7.

The following theorem, or alternatively the reformulation in Corollary 1.4.7, is known as the *periodicity theorem of Fine and Wilf.* We give two different proofs.

Theorem 1.4.6. *Let u and v be nonempty words. If a power of u and a power of v have a common prefix of length $|uv| - \gcd(|u|, |v|)$, then u and v are powers of a common word of length $\gcd(|u|, |v|)$.*

Proof. The first proof is related to example 1.4.5. Let w be the common prefix, $d = \gcd(|u|, |v|)$, $k = |u|/d$, $l = |v|/d$, and $m = |w|/d = k + l - 1$. Then we can write $u = u_1 \cdots u_k$, $v = v_1 \cdots v_l$, and $w = w_1 \cdots w_m$, where all u_i, v_i, w_i are words of length d.

Let

$$f : \{0, \dots, m\} \to \{0, \dots, m\}, \ f(n) = \begin{cases} n + k & \text{if } n < l \\ n - l & \text{if } n \geq l. \end{cases}$$

If n is such that $n, f(n) > 0$, then $w_n = w_{f(n)}$. Let $n_0 = 0$ and $n_{i+1} = f(n_i)$ for all $i \geq 0$. If we can prove that $\{n_1, \dots, n_m\} = \{1, \dots, m\}$, then it follows that $w_1 = \cdots = w_m$ and thus $w \in w_1^*$, and then $u, v \in w_1^*$, which proves the theorem.

For all i, j, $i < j$, there are integers $a, b \geq 0$ such that $n_j = n_i + ak - bl$ and $a + b = j - i$. If $n_i = n_j$, then $ak - bl = 0$, and then $a \geq l$ and $b \geq k$, so $j - i \geq k + l = m + 1$. It follows that $n_0, \dots, n_m$ are pairwise distinct, and thus $\{n_1, \dots, n_m\} = \{1, \dots, m\}$. This completes the first proof. □

Proof. The second proof requires some polynomial algebra. Let us assume that the alphabet is a subset of $\mathbb{Z}$ (we can always rename the letters, so this is not a restriction). For any word $t = a_0 \cdots a_n$, where $a_0, \dots, a_n$ are letters, we define a polynomial

$$P_t = \sum_{i=0}^{n} a_i X^i \in \mathbb{Z}[X].$$

Then it is easy to check that

$$P_{t^m} = \frac{1 - X^{m|t|}}{1 - X^{|t|}} \cdot P_t$$

for all m. Let $d = \gcd(|u|, |v|)$. Then $1 - X^{|u|} = (1 - X^d)Q$ and $1 - X^{|v|} = (1 - X^d)R$ for some polynomials Q, R. We have

$$P_{u^{|v|}} - P_{v^{|u|}} = \frac{1 - X^{|u||v|}}{1 - X^{|u|}} \cdot P_u - \frac{1 - X^{|u||v|}}{1 - X^{|v|}} \cdot P_v = \frac{1 - X^{|u||v|}}{(1 - X^d)QR} \cdot (RP_u - QP_v).$$

By the assumption about the common prefix, $P_{u^{|v|}} - P_{v^{|u|}}$ is divisible by $X^{|uv|-d}$, so $RP_u - QP_v$ must be divisible by $X^{|uv|-d}$. But the degree of $RP_u - QP_v$ is at most $|uv| - d - 1$, so $RP_u - QP_v = 0$. It follows that $P_{u^{|v|}} - P_{v^{|u|}} = 0$ and thus $u^{|v|} = v^{|u|}$. The claim follows quite easily. □

Corollary 1.4.7. *Let w be a word with periods k and l. If $|w| \geq k + l - \gcd(k, l)$, then w has period $\gcd(k, l)$.*

Proof. The word w is a prefix of a power of $\mathrm{pref}_k(w)$ and of a power of $\mathrm{pref}_l(w)$. By Theorem 1.4.6, w is a prefix of a power of a word of length $\gcd(k, l)$, and therefore it has period $\gcd(k, l)$. □

The bound $|w| \geq k + l - \gcd(k, l)$ is optimal in the sense that if neither of k, l divides the other, then there exists a word of length $k + l - \gcd(k, l) - 1$ that has periods k and l but not period $\gcd(k, l)$. This can be proved by using similar ideas, and it is often stated as part of the theorem of Fine and Wilf.

The next two results are sometimes useful.

Lemma 1.4.8. *Let x and y be words and $y \neq \varepsilon$. If x is a prefix of yx, then x is a fractional power of y.*

Proof. The case $x = \varepsilon$ is trivial. If $x \neq \varepsilon$, then yx has a border x, so by Theorem 1.4.4, yx is a fractional power of y. Then also x is a fractional power of y. □

Lemma 1.4.9. *Let x and y be words and $m, n \geq 1$. If one of $x^m y$ and $y^n x$ is a prefix of the other, then x and y are powers of a common word.*

Proof. The word x is a prefix of $y^n x$, so x is a fractional power of y^n by Lemma 1.4.8. It follows that $y^n x$ is a fractional power of y. Similarly, $x^m y$ is a fractional power of x. This means that a power of x and a power of y have a common prefix of length

$$\min\{|x^m y|, |y^n x|\} \geq |xy|,$$

so x and y are powers of a common word by Theorem 1.4.6. □

Exercises

1.4.1. Find some words that have a meaning in some natural language and are α-powers for some rational number $\alpha \geq 2$.

1.4.2. Show that a word and its reverse have the same periods. Show that conjugates do not necessarily have the same periods.

1.4.3. Prove Theorem 1.4.4.

1.4.4. Show that if a word w has a border, then its shortest border is of length at most $|w|/2$.

1.4.5. What is the maximum length of a nonunary word that has periods 5 and 8? Give an example of such a word. What if the word has to be nonbinary?

1.4.6. Let u, v, w be words. Show that if $w^2 = uvuvu$, then $u, v, w \in r^*$ for some word r. Give an example of nonempty words u, v, w such that $w^2 = uvu$ but $\rho(u) \neq \rho(v)$.

*1.4.7.** Prove the optimality result that was mentioned after Corollary 1.4.7.

*1.4.8.** Let w be a word and k its smallest period. Show that w is nonprimitive if and only if k divides $|w|$ and $k < |w|$.

1.4.9. Let x, y be words. Show that if xy and yx have a common prefix of length $|x| + |y| - \gcd(|x|, |y|)$, then $xy = yx$.

Chapter 2

Equations and codes

2.1 Word equations

We are interested in finding words that satisfy a given equality. For example, which words x satisfy $xaxbab = abaxbx$, where a and b are letters, or which triples of words x, y, z satisfy $xyz = zyx$? To help us study these kinds of questions, we give some definitions. Most importantly, we give a formal definition for a *word equation* and its *solutions*.

In this section, let Σ be an alphabet of constants and Ξ an alphabet of variables. A morphism $h : (\Xi \cup \Sigma)^* \to \Sigma^*$ is *constant-preserving* if $h(a) = a$ for all $a \in \Sigma$, and it is *periodic* if there exists a word w such that $h(X) \in w^*$ for all $X \in \Xi$. If $\Xi = \{X_1, \dots, X_n\}$, then by the morphism

$$(X_1, \dots, X_n) \mapsto (w_1, \dots, w_n)$$

we mean the unique constant-preserving morphism $h : (\Xi \cup \Sigma)^* \to \Sigma^*$ such that $h(X_i) = w_i$ for all i.

A *word equation* is a pair of words $(U, V) \in (\Xi \cup \Sigma)^* \times (\Xi \cup \Sigma)^*$. A *solution* of this equation is a constant-preserving morphism h such that $h(U) = h(V)$. An equation (U, V) is *constant-free* if $U, V \in \Xi^*$, and it is *trivial* if $U = V$. If $n = |\Xi|$, then (U, V) can be called an *n-variable equation*, and it can be called an *equation over* Ξ. The words U and V can be called the *left-hand side* and the *right-hand side* of (U, V), respectively.

Example 2.1.1. Let $\Xi = \{X\}$ and $\Sigma = \{a, b\}$. Consider the word equation $(XaXbab, abaXbX)$. Let f be the morphism $(X) \mapsto (\varepsilon)$ and let g be the morphism $(X) \mapsto (ab)$. These two morphisms are solutions of the equation:

$$\begin{aligned} f(XaXbab) &= f(X)af(X)bab = abab = abaf(X)bf(X) = f(abaXbX), \\ g(XaXbab) &= g(X)ag(X)bab = abaabbab = abag(X)bg(X) = g(abaXbX). \end{aligned}$$

It can be shown that the equation has no other solutions (see Example 2.2.7).

Example 2.1.2. Let $\Xi = \{X, Y, Z\}$ and $\Sigma = \{a, b\}$. The constant-free word equation (XYZ, ZYX) has infinitely many solutions (see Example 2.2.2), for example the following two:

$$(X, Y, Z) \mapsto (a, b, aba), \qquad (X, Y, Z) \mapsto (ab, (ab)^2, (ab)^3).$$

The first one is nonperiodic and the second one is periodic.

A *system of equations* is a set of equations. A morphism is a solution of a system if it is a solution of every equation in the system. A system of size two can be called a *pair of equations.* Two equations or systems are *equivalent* if they have the same set of solutions.

Example 2.1.3. Let $\Xi = \{X, Y, Z\}$ and $\Sigma = \{a, b\}$. The pair of equations $\{(XYZ, ZYX), (XYYZ, ZYYX)\}$ has the solution

$$(X, Y, Z) \mapsto (a, b, a).$$

The pair of equations has infinitely many other solutions as well.

Let $U, V, S, T \in (\Xi \cup \Sigma)^*$. The equations (U, V) and (V, U) are equivalent. This means that we can always swap the left-hand side and the right-hand side of an equation. The equations (SUT, SVT) and (U, V) are also equivalent. This means that we can always cancel common prefixes and common suffixes of the left-hand side and the right-hand side of an equation.

We can turn some earlier results into results about word equations. In particular, we can solve two simple constant-free equations: The *commutation equation* (XY, YX) and the *conjugacy equation* (XZ, ZY).

Theorem 2.1.4. *Let $\Xi = \{X, Y\}$. The solutions of the equation (XY, YX) are the morphisms*

$$(X, Y) \mapsto (p^i, p^j),$$

where $p \in \Sigma^*$ and $i, j \geq 0$.

Proof. Follows from Theorem 1.2.2. To be more specific, if h is a solution of (XY, YX), then $h(X)h(Y) = h(Y)h(X)$, so $h(X) = p^i$ and $h(Y) = p^j$ for some $p \in \Sigma^*$ and $i, j \geq 0$ by Theorem 1.2.2. On the other hand, if g is a morphism, $g(X) = p^i$, and $g(Y) = p^j$, then $g(XY) = p^{i+j} = g(YX)$, so g is a solution (XY, YX). □

Theorem 2.1.5. *Let $\Xi = \{X, Y, Z\}$. The solutions of the equation (XZ, ZY) are the morphisms*

$$(X, Y, Z) \mapsto (pq, qp, (pq)^k p), \qquad (X, Y, Z) \mapsto (\varepsilon, \varepsilon, p),$$

where $p, q \in \Sigma^*$ and $k \geq 0$.

Proof. Follows from Theorem 1.2.6. □

Theorem 2.1.6. *Let $X, Y \in \Xi$, $U, V \in \Xi^*$ and $m, n \geq 1$. If h is a solution of the equation $(X^m YU, Y^n XV)$, then $h(X)$ and $h(Y)$ are powers of a common word.*

Proof. Follows from Lemma 1.4.9. □

The next theorem is a generalization of Theorem 2.1.4. It is sometimes stated in the following form: If two words satisfy a nontrivial relation, then they are powers of a common word.

Theorem 2.1.7. *A nontrivial constant-free two-variable word equation has only periodic solutions.*

Proof. Given a nontrivial constant-free equation over $\{X, Y\}$, we can cancel common prefixes and maybe swap the sides to either get an equation of the form (W, ε), where $W \in \{X, Y\}^+$, or of the form (XU, YV), where $U, V \in \{X, Y\}^*$. The former clearly has only periodic solutions, and the latter is equivalent to the equation $(XUXY, YVXY)$, which has only periodic solutions by Theorem 2.1.6. □

Example 2.1.8. Consider the equation (X^3, Y^2) over $\{X, Y\}$. If h is a solution of this equation, then $h(X) = p^m$ and $h(Y) = p^n$ for some word p and numbers $m, n \geq 0$. But not all such morphisms h are solutions: It must be

$$p^{3m} = h(X^3) = h(Y^2) = p^{2n},$$

which is equivalent to $3m = 2n$ (if $p \neq \varepsilon$), so the solutions are exactly the morphisms

$$(X, Y) \mapsto (p^m, p^n), \qquad 3m = 2n,$$

or equivalently, the morphisms

$$(X, Y) \mapsto (p^{2k}, p^{3k}), \qquad k \geq 0.$$

We conclude this section by stating a result sometimes known as the theorem of Lyndon and Schützenberger.

Theorem 2.1.9. *Let $\Xi = \{X, Y, Z\}$ and let $k, m, n \geq 2$ be integers. The word equation $(X^k, Y^m Z^n)$ has only periodic solutions.*

Exercises

2.1.1. Let $\Xi = \{X, Y\}$. Find all solutions of the word equation (XY^{12}, Y^2X^5).

2.1.2. Let $\Xi = \{X, Y\}$. Find all solutions of the word equation (X^2, YZ).

2.1.3. Let $\Xi = \{X, Y, Z\}$. Find all solutions of the word equation $(XYZY, Y^2XZ)$.

2.1.4. Let $\Xi = \{S, T, X, Y\}$. Find a solution h for the equation (ST^3S, XY^3X) such that $h(SX) \neq h(XS)$.

2.1.5. Let $\Xi = \{S, T, X, Y\}$. Find a solution h for the equation (ST^4S, XY^4X) such that $h(SX) \neq h(XS)$.

2.1.6. Give an example of a word equation or a system of word equations that has at least 2020 solutions but not infinitely many. If you are feeling competitive, you can try to give an example that is as short as possible. (The length of an equation (U, V) is defined to be $|UV|$, and the length of a system is defined to be the sum of the lengths of the equations in the system.)

2.1.7. Let $\Xi = \{X, Y, Z\}$ and let $k, m \geq 2$ be integers. Find a nonperiodic solution for the word equation $(X^k, Y^m Z)$.

*2.1.8.** Prove Theorem 2.1.9 in the case $k = m = n$.

2.2 Solving word equations

In this section, we give many examples of solving a word equation, that is, finding all of its solutions. These examples illustrate different techniques that can be used in some particular cases. We do not give any general method that would work for all equations. In fact, even checking whether a given word equation has at least one solution is very complicated (see Remark 2.2.9).

First, we solve two constant-free three-variable equations.

Example 2.2.1. Let $\Xi = \{X, Y, Z\}$. The solutions of the equation (XYZ, ZXY) are the morphisms

$$(X, Y, Z) \mapsto ((pq)^i p, q(pq)^j, (pq)^k), \qquad (X, Y, Z) \mapsto (\varepsilon, \varepsilon, p),$$

where $p, q \in \Sigma^*$ and $i, j, k \geq 0$. It is easy to see that all these morphisms are solutions. To see that there are no other solutions, let x, y, z be words such that $xyz = zxy$. Then xy and z commute, so $xy = r^m$ and $z = r^k$ for some word r and integers $m, k \geq 0$. If $m \geq 1$, then there exist words p, q and integers i, j such that $r = pq$, $m = i + j + 1$, $x = r^i p$, $y = qr^j$. If $m = 0$, then $x = y = \varepsilon$.

Example 2.2.2. Let $\Xi = \{X, Y, Z\}$. The solutions of the equation (XYZ, ZYX) are the morphisms

$$(X, Y, Z) \mapsto ((pq)^i p, q(pq)^j, (pq)^k p), \quad (X, Y, Z) \mapsto (p, \varepsilon, \varepsilon), \quad (X, Y, Z) \mapsto (\varepsilon, \varepsilon, p),$$

where $p, q \in \Sigma^*$ and $i, j, k \geq 0$. It is easy to see that all these morphisms are solutions. To see that there are no other solutions, let x, y, z be words such that $xyz = zyx$. Then $xyzy = zyxy$, so $xy = r^m$ and $zy = r^n$ for some word r and integers $m, n \geq 0$. If $m, n \geq 1$, then there exist words p, q and integers i, j, k such that $r = pq$, $m = i + j + 1$, $n = k + j + 1$, $x = r^i p$, $y = qr^j$, $z = r^k p$. If $m = 0$ or $n = 0$, then $x = y = \varepsilon$ or $z = y = \varepsilon$.

Next, we give examples of various *length arguments.*

Example 2.2.3. Let $\Xi = \{X, Y, Z\}$ and consider the equation $(XYZY, YXYZ)$. If $x, y, z \in \Sigma^*$ and $xyzy = yxyz$, then $xy = \mathrm{pref}_{|xy|}(xyzy) = \mathrm{pref}_{|xy|}(yxyz) = yx$. Thus x and y are powers of a common word. Similarly, $zy = yz$ and y and z are powers of a common word. So if $y \neq \varepsilon$, then x, y, z are powers of a common word. It follows that the only nonperiodic solutions of the equation are the morphisms $(X, Y, Z) \mapsto (p, \varepsilon, q)$, where $p, q \in \Sigma^*$ and $pq \neq qp$.

Example 2.2.4. Let $\Xi = \{X, Y, Z\}$ and consider the equation $(XZX, Y^4 Z)$. If $x, y, z \in \Sigma^*$ and $xzx = y^4 z$, then $|xzx| = |y^4 z|$ and thus $2|x| = 4|y|$. Then $x = y^2$ and $y^2 z y^2 = y^4 z$, and thus y and z are powers of a common word, and also x is a power of the same word. It follows that the equation has only periodic solutions.

Example 2.2.5. Let $\Xi = \{X, Y, Z\}$ and consider the equation (X^2, YZ^2Y). If $x, y, z \in \Sigma^*$ and $x^2 = yz^2y$, then $|x| = |x^2|/2 = |yz^2y|/2 = |yz|$. Then $x = yz = zy$ and thus x, y, z are powers of a common word. It follows that the equation has only periodic solutions.

Next, we consider some one-variable equations.

Example 2.2.6. Let $\Xi = \{X\}$, $\Sigma = \{a, b\}$, $u \in \Sigma^*$, $v \in \Sigma^+$, and uv primitive. The solutions of the word equation (Xvu, uvX) are the morphisms

$$(X) \mapsto ((uv)^i u),$$

where $i \geq 0$. It is easy to see that these are solutions. To see that there are no other solutions, let x be a word such that $xvu = uvx$. By Example 2.2.2,

$$(x, v, u) = ((pq)^i p, q(pq)^j, (pq)^k p)$$

for some $p, q \in \Sigma^*$ and $i, j, k \geq 0$. By the primitivity of uv, $j = k = 0$.

It is known that every nontrivial one-variable equation with infinitely many solutions is equivalent to an equation of the form (Xvu, uvX).

Example 2.2.7. Let $\Xi = \{X\}$ and $\Sigma = \{a, b\}$. The solutions of the word equation $(XaXbab, abaXbX)$ are the morphisms

$$(X) \mapsto (\varepsilon), \qquad (X) \mapsto (ab).$$

It is easy to see that these are solutions. To see that there are no other solutions, let x be a nonempty word such that $xaxbab = abaxbx$. By Lemma 1.4.8, x must be a fractional power of aba, and x must end in b, so $x = (aba)^n ab$ for some integer $n \geq 0$. If $n \geq 1$, then x ends in aab, which is a contradiction, because $xaxbab = abaxbx$ ends in bab.

Using similar ideas, it can be shown that the equation

$$(XaXbXaabbabaXbabaabbab, abaabbabaXbabaabbXaXbX)$$

has exactly three solutions

$$(X) \mapsto (\varepsilon), \qquad (X) \mapsto (ab), \qquad (X) \mapsto (abaabbab).$$

More specifically, if $(X) \mapsto (x)$ is a solution, then x must be a fractional power of $abaabbaba$, and if $|x| \geq 9$, then x must end in $babaabbab$. No fractional power of $abaabbaba$ ends in $babaabbab$, so $|x| \leq 8$. Checking the nine proper prefixes of $abaabbaba$ gives the three solutions mentioned above.

It is known that every one-variable equation with only finitely many solutions has at most three solutions.

If the alphabet of constants is unary, then word equations are essentially linear Diophantine equations.

Example 2.2.8. Let $\Xi = \{X, Y, Z\}$ and $\Sigma = \{a\}$. The solutions of the word equation (Xa^2Y^3, Z^2) are the morphisms

$$(X, Y, Z) \mapsto (a^{2i}, a^{2j}, a^{i+3j+1}), \qquad (X, Y, Z) \mapsto (a^{2i+1}, a^{2j+1}, a^{i+3j+3}),$$

where $i, j \geq 0$. Note that for $x, y, z \in a^*$, $xa^2y^3 = z^2$ is equivalent to

$$|x| + 3|y| + 2 = 2|z|.$$

We see that $|x|$ and $|y|$ must have the same parity. This leads to the above solutions.

Remark 2.2.9. Mostly, algorithmic questions are not considered in these lecture notes, but let us make a remark about the following three algorithmic decision problems:

1. Given a morphism $h : \Xi^* \to \Sigma^*$, does there exist two distinct words $U, V \in \Xi^*$ such that $h(U) = h(V)$? In other words, is h noninjective?
2. Given words $U, V \in (\Xi \cup \Sigma)^*$, does there exist a constant-preserving morphism $h : (\Xi \cup \Sigma)^* \to \Sigma^*$ such that $h(U) = h(V)$? In other words, does the word equation (U, V) have a solution?
3. Given two morphisms $f, g : \Xi^* \to \Sigma^*$, does there exist a word $W \in \Xi^+$ such that $f(W) = g(W)$?

These problems seem similar, but the first one is easy, the second one is difficult, and the third one is impossible:

1. The first problem is closely related to *codes*, which are studied in Section 2.3. It can be solved efficiently by the *Sardinas–Patterson algorithm.*
2. Using terminology of complexity theory, it is known that the second problem is NP-hard and in PSPACE, and it has been conjectured that it is in NP (and therefore NP-complete).
3. The third problem is called the *Post correspondence problem* (or *PCP*) and it is known to be undecidable.

Exercises

2.2.1. Let $\Xi = \{X, Y\}$ and $\Sigma = \{a, b\}$. Solve the equation (XaY, YaX).

2.2.2. Let $\Xi = \{S, T, X, Y\}$. Solve the equation (S^2XYT, T^2YXS).

2.2.3. Let $\Xi = \{X, Y, Z\}$. Solve the equation (XZX, Y^2).

2.2.4. Let $\Xi = \{X, Y, Z, T\}$ and $\Sigma = \{a, b\}$. Solve the equation

$$(XYZTZYXTa, ZYXTaXYZT).$$

2.2.5. Let $\Xi = \{X\}$ and $\Sigma = \{a, b\}$. Solve the equation $(XXbaaba, aabaXbX)$.

2.2.6. Let $\Xi = \{X, Y, Z\}$ and $\Sigma = \{a\}$. Solve the pair of equations

$$\{(XZY, Z^2), (XaYaX, ZaZ)\}.$$

Because Σ is unary, this is essentially linear algebra.

*2.2.7.** Let $\Xi = \{X, Y, Z\}$. Solve the equation $(X^2Y^2Z^2, (XYZ)^2)$.

*2.2.8.** Show that if a one-variable equation has the solutions $(X) \mapsto (a^m)$ and $(X) \mapsto (a^n)$, where $a \in \Sigma$ and $m \neq n$, then it has the solution $(X) \mapsto (a^k)$ for all $k \geq 0$.

*2.2.9.** Let $a, b \in \Sigma$, $a \neq b$, and $U, V, U', V' \in (\Xi \cup \Sigma)^*$. Show that the pair of equations $\{(U, V), (U', V')\}$ is equivalent to the equation $(UaU'UbU', VaV'VbV')$. Conclude that every finite system of equations is equivalent to a single equation (assuming that Σ is not unary). Give an example of a pair of constant-free equations that is not equivalent to any single constant-free equation.

2.3 Free monoids and codes

Let M be a nonempty set and $*$ a binary operation that is defined on all pairs of elements of M. Then $(M,*)$ is a *monoid* if the following conditions are satisfied:

1. $x*y \in M$ for all $x,y \in M$.
2. $(x*y)*z = x*(y*z)$ for all $x,y,z \in M$.
3. There exists $e \in M$ such that $e*x = x*e = x$ for all $x \in M$.

The element e is called the *neutral element*. It is unique, because if e and e' are neutral elements, then $e = e*e' = e'$. If $L \subseteq M$ and $(L,*)$ and $(M,*)$ are monoids with the same neutral element, then $(L,*)$ is a *submonoid* of $(M,*)$.

Remark 2.3.1. A technical remark about the definitions: A *binary operation* on a set S is a function $S \times S \to S$. The image of a pair $(x,y) \in S \times S$ under $*$ is denoted by $x*y$. In the above definition of a monoid, $*$ can be a binary operation on M, but it can also be a binary operation on a superset of M. This allows us to use the same operation for many different sets M.

Example 2.3.2. Here are some examples of monoids:

1. Every group is a monoid.
2. If $(R,+,\cdot)$ is a ring, then $(R,\cdot)$ is a monoid.
3. $(\mathbb{Z}_{\geq 0},+)$ is a monoid.
4. Let $\mathcal{P}(S)$ the set of subsets of a set S. Then $(\mathcal{P}(S),\cup)$ and $(\mathcal{P}(S),\cap)$ are monoids. The neutral elements are $\varnothing$ and S, respectively.

If Σ is an alphabet, then $(\Sigma^*,\cdot)$ is a monoid and ε is the neutral element. From now on, we concentrate on the monoid $(\Sigma^*,\cdot)$ and its submonoids. Because the binary operation is always concatenation, we can talk about a monoid M instead of a monoid $(M,\cdot)$.

Theorem 2.3.3. *Let $L \subseteq \Sigma^*$. The following are equivalent:*

1. *L is a submonoid of Σ^*.*
2. *$L^* = L$.*
3. *$L^2 \cup \{\varepsilon\} \subseteq L$.*

Proof. $1 \implies 2$: Clearly $L \subseteq L^*$. Every element of L^* is ε, an element of L, or a product of two or more elements of L. If L is a submonoid, then all of these must be in L, so $L^* \subseteq L$.

$2 \implies 3$: Clearly $L^2 \cup \{\varepsilon\} \subseteq L^*$.

$3 \implies 1$: The first condition of the definition of a monoid is satisfied because $L^2 \subseteq L$, the second one because $L \subseteq \Sigma^*$, and the third one because $\varepsilon \in L$. Thus L is a monoid. It is a submonoid of Σ^* because $L \subseteq \Sigma^*$ and $\varepsilon \in L$. □

Example 2.3.4. Let $L \subseteq \Sigma^*$. The set L^* is a monoid because $(L^*)^* = L^*$.

Let $\Sigma = \{a,b\}$. The set $M = \{w \in \Sigma^* \mid |w|_a \equiv 0 \pmod 2\}$ is a monoid because $\varepsilon \in M$ and $uv \in M$ for all $u,v \in M$.

A subset G of a monoid $M \subseteq \Sigma^*$ is a *generating set* of M if $G^* = M$. A generating set of M is *minimal* if it does not have a proper subset that is a generating set of M. By the next theorem, every monoid $M \subseteq \Sigma^*$ has a unique minimal generating set, and it consists of those nonempty words in M that cannot be written as a product of two nonempty words in M.

Theorem 2.3.5. *A submonoid M of Σ^* has the unique minimal generating set*

$$G = (M \smallsetminus \{\varepsilon\}) \smallsetminus (M \smallsetminus \{\varepsilon\})^2.$$

Proof. First, we prove that G is a generating set of M. Clearly $G^* \subseteq M$, so it is sufficient to prove that $M \cap \Sigma^n \subseteq G^*$ for all $n \geq 0$, which can be done by induction as follows. The case $n = 0$ is clear. Let $n \geq 1$. Let $x \in M \cap \Sigma^n$. If $x \in G$, then $x \in G^*$. If $x \notin G$, then $x \in (M \smallsetminus \{\varepsilon\})^2$, so $x = yz$ for some $y, z \in M \smallsetminus \{\varepsilon\}$. Because $|y|, |z| < |x| = n$, it follows from the induction hypothesis that $y, z \in G^*$ and thus $x \in G^*$. We have proved that $M \cap \Sigma^n \subseteq G^*$, which completes the induction.

Then, we prove that every generating set H of M contains G as a subset and therefore G is the unique minimal generating set. If $x \in M \smallsetminus H \smallsetminus \{\varepsilon\}$, then $x = y_1 \cdots y_n$ for some $n \geq 2$ and $y_1, \ldots, y_n \in H \smallsetminus \{\varepsilon\}$. But then $y_1, y_2 \cdots y_n \in M \smallsetminus \{\varepsilon\}$ and thus $x \in (M \smallsetminus \{\varepsilon\})^2$, so $x \notin G$. This shows that $G \subseteq H$. □

Example 2.3.6. The minimal generating set of the monoid M of Example 2.3.4 is $ab^*a \cup \{b\}$.

If $M \subseteq \Sigma^*$ is a monoid and $S \subseteq M$, then an *S-factorization* of $x \in M$ is a sequence $(x_1, \ldots, x_n)$ of elements of S such that $x = x_1 \cdots x_n$. Then S^* is the set of elements that have at least one S-factorization.

A subset C of a monoid $M \subseteq \Sigma^*$ is a *code* if every element of C^* has a unique C-factorization. A monoid $M \subseteq \Sigma^*$ is *free* if it has a generating set that is a code.

Example 2.3.7. An alphabet Σ is a code and therefore Σ^* is a free monoid. Also Σ^n is a code for all $n \geq 1$. The free monoid $(\Sigma^n)^*$ consists of the words whose length is divisible by n. The empty set is a code and $\varnothing^* = \{\varepsilon\}$ is a free monoid. The empty word can never be in a code, because (ε) and $(\varepsilon, \varepsilon)$ are distinct $\{\varepsilon\}$-factorizations of the same element.

Example 2.3.8. The set $C = \{a, bba, bbaab\}$ is a code. To see this, let $(x_1, \ldots, x_m)$ and $(y_1, \ldots, y_n)$ be C-factorizations of the same element and $|x_1| < |y_1|$. Then it must be $x_1 = bba$ and $y_1 = bbaab$, and then $x_2 = a$ and $x_3 \in \{bba, bbaab\}$. But then there is no possible value for y_2, which is a contradiction.

It follows from the next theorem that a submonoid of Σ^* is free if and only if its minimal generating set is a code.

Theorem 2.3.9. *If C is a code, then C is the minimal generating set of C^*.*

Proof. Clearly C is a generating set of C^*. Let $B \subseteq C$ be a generating set of C^*. If $x \in C$, then $x = y_1 \cdots y_n$ for some $n \geq 1$ and $y_1, \ldots, y_n \in B$. If $n \geq 2$, then x has two distinct C-factorizations (x) and $(y_1, \ldots, y_n)$, which is a contradiction. Thus $n = 1$ and $x \in B$. We have shown that $C \subseteq B$. This proves that C is a minimal generating set. □

A set $P \subseteq \Sigma^+$ is a *prefix code* (*suffix code*) if there does not exist words $u, v \in P$ such that u is a proper prefix (proper suffix, respectively) of v.

Theorem 2.3.10. *Prefix codes and suffix codes are codes.*

Proof. Let $P \subseteq \Sigma^*$ be a prefix code. Let $(x_1, \dots, x_m)$ and $(y_1, \dots, y_n)$ be P-factorizations of the same word w. If $m = 0$ or $n = 0$, then $w = \varepsilon$ and $m = n = 0$, so the factorizations are equal. Consider the case $m, n > 0$. If $|x_1| \neq |y_1|$, then one of x_1, y_1 is a proper prefix of the other, which is a contradiction. Therefore $|x_1| = |y_1|$ and thus $x_1 = y_1$. It follows that $(x_2, \dots, x_m)$ and $(y_2, \dots, y_n)$ are P-factorizations of the same word. Continuing the same way, we see that $m = n$ and $x_i = y_i$ for all i. This shows that no word has two distinct P-factorizations, so P is a code. The claim about suffix codes can be proved in a similar way. □

Example 2.3.11. Let $\Sigma = \{a, b\}$. The set $\{ab, aab\}$ is a prefix code, so $\{ab, aab\}^*$ is a free monoid. The set $\{ab, aab, aabab\}$ is not a code, because $aab \cdot ab = aabab$, but the monoid $\{ab, aab, aabab\}^*$ is free, because $\{ab, aab, aabab\}^* = \{ab, aab\}^*$.

Example 2.3.12. The monoid M of Example 2.3.4 is free, because its minimal generating set $ab^*a \cup \{b\}$ is a prefix code.

A submonoid M of Σ^* is *stable* if $x, z, xy, yz \in M$ implies $y \in M$.

Theorem 2.3.13. *A submonoid of Σ^* is free if and only if it is stable.*

Proof. First, let M be a free submonoid of Σ^*. Then $M = C^*$ for some code C. If $x, z, xy, yz \in M$, then $x = x_1 \cdots x_k$, $z = z_1 \cdots z_l$, $xy = u_1 \cdots u_m$, $yz = v_1 \cdots v_n$, where $k, l, m, n \geq 0$ and $x_i, y_i, u_i, v_i \in C$ for all i. It follows that xyz has C-factorizations $(x_1, \cdots, x_k, v_1, \dots, v_n)$ and $(u_1, \dots, u_m, z_1, \dots, z_l)$, which must be equal, because C is a code. Then $k \leq m$ and $x_i = u_i$ for all $i \in \{1, \dots, k\}$, and then $y = u_{k+1} \cdots u_m \in M$. This shows that M is stable.

Then, let M be a stable submonoid of Σ^* and let G be the minimal generating set of M. We assume that some element has two distinct G-factorizations $(x_1, \dots, x_m)$ and $(z_1, \dots, z_n)$ and derive a contradiction. We can assume that $x_1 \neq z_1$ (otherwise, consider the factorizations $(x_2, \dots, x_m)$ and $(z_2, \dots, z_n)$), and by symmetry, we can assume that $|x_1| < |z_1|$. Then $z_1 = x_1 y$ for some word $y \neq \varepsilon$. If we let $x = x_1$ and $z = z_2 \cdots z_n$, then $x, z, xy, yz \in M$, and thus $y \in M$ because M is stable. But then $z_1 \in (M \smallsetminus \{\varepsilon\})^2$, which is a contradiction by Theorem 2.3.5. Therefore, G is a code and M is free. □

A submonoid M of Σ^* is *right unitary* if $x, xy \in M$ implies $y \in M$, and *left unitary* if $z, yz \in M$ implies $y \in M$. If a monoid is right unitary or left unitary, then it is clearly stable and therefore free.

Example 2.3.14. Another way to see that the monoid M of Example 2.3.4 is free is to notice that it is right unitary.

There is a close connection between codes and injective morphisms.

Theorem 2.3.15. *Let Σ and Γ be alphabets and let $h : \Sigma^* \to \Gamma^*$ be a morphism that is injective on Σ. Then h is injective if and only if $h(\Sigma)$ is a code.*

Proof. The language $h(\Sigma)$ is not a code if and only if some word in Γ^* has two distinct $h(\Sigma)$-factorizations $(h(a_1), \dots, h(a_m))$ and $(h(b_1), \dots, h(b_n))$, where all $a_i, b_i \in \Sigma$. Because h is injective on Σ, these factorizations being distinct is equivalent to $a_1 \cdots a_m \neq b_1 \cdots b_n$. Thus $h(\Sigma)$ is not a code if and only if there exists $u, v \in \Sigma^*$ such that $h(u) = h(v)$ and $u \neq v$. This last condition is equivalent to h being noninjective. □

The next example collects together many things we have proved, and gives a characterization of two-element codes.

Example 2.3.16. Let $x, y \in \Sigma^+$ and $x \neq y$. The following are equivalent:

1. $\rho(x) = \rho(y)$.
2. $x, y \in w^*$ for some word w.
3. $xy = yx$.
4. $x^{|y|} = y^{|x|}$.
5. $x^m = y^n$ for some integers $m, n \geq 1$.
6. $(X, Y) \mapsto (x, y)$ is a solution of a nontrivial constant-free equation on $\{X, Y\}$.
7. The morphism $h : \{0,1\}^* \to \Sigma^*,\ h(0) = x,\ h(1) = y$, is not injective.
8. $\{x, y\}$ is not a code.

Most of the equivalences follow directly from results we have proved. Checking the details is left as an exercise. What about the case $x = y$ or $x = \varepsilon$ or $y = \varepsilon$?

Exercises

2.3.1. Give an example of two monoids M, N such that $M \cup N = \{a, b\}^*$ and $M \cap N = \{\varepsilon\}$.

2.3.2. Check that the following sets are monoids and find their minimal generating sets: $\{a, ab\}^* \cap \{a, ba\}^*$, $\{w \in \{a, b\}^* \mid |w| \geq 3\} \cup \{\varepsilon\}$.

2.3.3. Is $\{a, ab, aba, abb\}^*$ a free monoid? Is $\{ab, aba, baab\}^*$ a free monoid? Is $\{a, ab, bba\}^*$ a free monoid?

2.3.4. Show that the set $\{w \in \{a, b\}^* \mid |w|_{ab} \geq 1\} \cup \{\varepsilon\}$ is a monoid. Find its minimal generating set. Is this monoid free?

2.3.5. Show that if C is a code, then $\{x^R \mid x \in C\}$ is a code.

2.3.6. If C and D are codes, then is CD necessarily a code? If C and D are prefix codes, then is CD necessarily a prefix code?

2.3.7. Let $\Sigma = \{a, b\}$, $\alpha, \beta \in \mathbb{R}$, and $f : \Sigma^* \to \mathbb{R},\ f(w) = \alpha|w|_a + \beta|w|_b$. Show that $\{w \in \Sigma^* \mid f(w) = 0\}$ is a free monoid.

2.3.8. Prove that a submonoid of Σ^* is right (left) unitary if and only if its minimal generating set is a prefix code (suffix code, respectively).

2.3.9. Answer the question at the end of Example 2.3.16.

*2.3.10.** Show that every submonoid of a^* has a finite generating set.

2.4 Rank and defect

In linear algebra, there exist the powerful notions of *dimension* and *rank*. If the rank of a set S of vectors is denoted by $r(S)$, then we can prove, for example, the following:

1. If S is linearly independent, then $r(S) = |S|$.
2. If S is not linearly independent, then $r(S) < |S|$.
3. If the vectors in S satisfy two independent linear relations, then $r(S) \leq |S| - 2$.
4. The vectors in S can satisfy at most $|S|$ independent linear relations.

In combinatorics on words, we can define the rank of a language, and then prove analogous versions of the first two statements above, but not of the other two. So words satisfy some "dimension properties", but they are much weaker than those of vectors. We need the following lemma before we can give the definition.

Lemma 2.4.1. *Let F be a family of submonoids of Σ^* and let M be the intersection of the monoids in F. Then M is a monoid. Moreover, if every monoid in F is free, then M is free.*

Proof. Follows from Theorems 2.3.3 and 2.3.13. □

The *free rank* of a language L, denoted by $r_{\mathrm{f}}(L)$, is the size of the minimal generating set of the smallest free monoid containing L (the smallest free monoid containing L is the intersection of all free monoids containing L, which is a free monoid by Theorem 2.4.1). If L is a code, then clearly $r_{\mathrm{f}}(L) = |L|$.

We concentrate mostly on the free rank, but there are also other rank functions. For example, the *combinatorial rank* of a language L, denoted by $r_{\mathrm{c}}(L)$, is the size of a smallest language K such that $L \in K^*$. Clearly, $r_{\mathrm{c}}(L) \leq r_{\mathrm{f}}(L)$ and $r_{\mathrm{c}}(L) \leq |\Sigma|$, where Σ is the alphabet.

Example 2.4.2. If $L = \varnothing$ or $L = \{\varepsilon\}$, then $r_{\mathrm{c}}(L) = r_{\mathrm{f}}(L) = 0$. Otherwise, if $L \subseteq w^*$ for some word w, then $r_{\mathrm{c}}(L) = r_{\mathrm{f}}(L) = 1$. In all other cases, $r_{\mathrm{f}}(L) \geq r_{\mathrm{c}}(L) \geq 2$.

Example 2.4.3. Consider the language $L = \{a, bba, abbaab, bbaabbba\}$. Clearly $r_{\mathrm{c}}(L) = 2$. Note that L is not a code, because $a \cdot bbaabbba = abbaab \cdot bba$. If $L \subseteq C^*$, where C is a code, then it must be $bbaab \in C^*$ by Theorem 2.3.13. Thus $\{a, bba, bbaab\} \subseteq C^*$. Because $\{a, bba, bbaab\}$ is a code by Example 2.3.8, $\{a, bba, bbaab\}^*$ is the smallest free monoid containing L. Thus $r_{\mathrm{f}}(L) = 3$.

The main goal in this section is to prove the *defect theorem* (Theorem 2.4.5). It claims that if a language is not a code, then its free rank (and therefore also combinatorial rank) is less than its size. This is analogous to the fact that $r(S) < |S|$ if S is a linearly dependent set of vectors. We start with a lemma.

Lemma 2.4.4. *Let L be a language and $\varepsilon \notin L$. Let C be the minimal generating set of the smallest free monoid containing L. We can define a function $f : L \to C$ by letting $f(w)$ be the first element in the unique C-factorization of w. Then f is surjective.*

Proof. We assume that f is not surjective and derive a contradiction. Let $v \in C \smallsetminus f(L)$ and $B = (C \smallsetminus \{v\})v^*$. Clearly $v \notin (C \smallsetminus \{v\})^*$, and every word in $(C \smallsetminus \{v\})v^+$ is longer than v, so $v \notin B^*$. It follows that $B^* \subsetneq C^*$, and from $f(L) \subseteq C \smallsetminus \{v\}$ it follows that $L \subseteq B^*$. If some word u has B-factorizations

$$(x_1 v^{i_1}, \dots, x_m v^{i_m},), (y_1 v^{j_1}, \dots, y_n v^{j_n})$$

where $x_1, \dots, x_m, y_1, \dots, y_n \in C \smallsetminus \{v\}$, then u has the C-factorizations

$$(x_1, \underbrace{v, \dots, v}_{i_1}, \dots, x_m, \underbrace{v, \dots, v}_{i_m}),\ (y_1, \underbrace{v, \dots, v}_{j_1}, \dots, y_n, \underbrace{v, \dots, v}_{j_n}),$$

which must be equal. It follows that the B-factorizations of u are also equal. This shows that B is a code, which contradicts the minimality of C^*. □

Theorem 2.4.5. *If a finite language L is not a code, then $r_{\mathrm{f}}(L) \leq |L| - 1$.*

Proof. First, let $\varepsilon \notin L$ and let C and f be as in Lemma 2.4.4. Then $r_{\mathrm{f}}(L) = |C|$. Because L is not a code, some word $u \in L^*$ has two L-factorizations $(x_1, \dots x_m)$ and $(y_1, \dots, y_n)$ such that $x_1 \neq y_1$. But then we get the unique C-factorization of u by concatenating the C-factorizations of the words x_i, and also by concatenating the C-factorizations of the words y_i. If the first element of the C-factorization of u is z, then $f(x_1) = z = f(y_1)$. This shows that f is not injective, but it is surjective by Lemma 2.4.4, so $|C| < |L|$.

Then, let $\varepsilon \in L$ and let $K = L \smallsetminus \{\varepsilon\}$. Clearly, $r_{\mathrm{f}}(L) = r_{\mathrm{f}}(K)$. If K is a code, then $r_{\mathrm{f}}(K) = |K|$, and otherwise $r_{\mathrm{f}}(K) < |K|$ by the above. In any case, $r_{\mathrm{f}}(L) = r_{\mathrm{f}}(K) \leq |K| = |L| - 1$. □

The defect theorem can also be formulated for word equations.

Corollary 2.4.6. *Let h be a solution of a nontrivial constant-free word equation over Ξ. Then $r_{\mathrm{f}}(h(\Xi)) \leq |\Xi| - 1$.*

Proof. Because h is a solution of a nontrivial constant-free equation, h is not injective. By Theorem 2.3.15, h is not injective on Ξ or $h(\Xi)$ is not a code. In the former case, $r_{\mathrm{f}}(h(\Xi)) \leq |h(\Xi)| < |\Xi|$, and in the latter case, $r_{\mathrm{f}}(h(\Xi)) < |h(\Xi)| \leq |\Xi|$ by Theorem 2.4.5. □

A system of word equations is *independent* if it is not equivalent to any of its proper subsets. If h is a solution of an independent pair of equations, then does it follow that $r_{\mathrm{f}}(h(\Xi)) \leq |\Xi| - 2$? The answer is no (exercise). However, we can prove the following two theorems. The first one is a special case of the second one.

Theorem 2.4.7. *Let $\Xi = \{X, Y, Z\}$ and $S, T, U, V \in \Xi^*$. Let h be a solution of the pair of word equations $\{(XS, YT), (XU, ZV)\}$. Then h is periodic or $\varepsilon \in h(\Xi)$.*

Proof. Let $L = h(\Xi)$. If $\varepsilon \notin L$, then we can let C and f be as in Lemma 2.4.4. Like in the proof of Theorem 2.4.5, we see that $f(h(X)) = f(h(Y))$ and $f(h(X))) = f(h(Z))$. It follows that $f(h(\Xi)) = C$ is a singleton, and thus $r_{\mathrm{f}}(h(\Xi)) = 1$, meaning that h is periodic. □

The statement of the next theorem requires some terminology from graph theory.

Theorem 2.4.8. *Let h be a solution of a system of word equations*

$$\{(X_iU_i, Y_iV_i) \mid 1 \le i \le n\},$$

where $X_i, Y_i \in \Xi$, $X_i \ne Y_i$, $U_i, V_i \in \Xi^*$ for all i. Let c be the number of connected components of the graph with the set of vertices Ξ and the set of edges

$$\{\{X_i, Y_i\} \mid 1 \le i \le n\}.$$

Then $r_{\mathrm{f}}(h(\Xi)) \le c$ or $\varepsilon \in h(\Xi)$.

Proof. Left as an exercise. □

Example 2.4.9. Let $\Xi = \{X, Y, Z\}$ and $\Sigma = \{a, b\}$. Let us solve the word equation $(XYZZYXZ, ZXZYZYX)$. By a length argument, it is equivalent to the pair of equations $\{(XYZZ, ZXZY), (YXZ, ZYX)\}$. By Theorem 2.4.7, every solution h is periodic or maps some variable to ε. The periodic solutions are

$$(X, Y, Z) \mapsto (p^i, p^j, p^k),$$

where $p \in \Sigma^*$ and $i, j, k \ge 0$. If $h(X) = \varepsilon$, then $h(Y)$ and $h(Z)$ commute, and if $h(Y) = \varepsilon$, then $h(X)$ and $h(Z)$ commute, so this does not give any nonperiodic solutions. However, if $h(Z) = \varepsilon$, then $h(X)$ and $h(Y)$ can be arbitrary, so we get the solutions

$$(X, Y, Z) \mapsto (p, q, \varepsilon),$$

where $p, q \in \Sigma^*$.

How large can an independent system of constant-free word equations be? For one variable the answer is 1, for two variables it is 2, and for three variables the answer is known to be between 3 and 18. In the general case of n variables, there are independent systems larger than Cn^4 for certain constant $C \in \mathbb{R}_+$, and the only known upper bound is the following result, known as *Ehrenfeucht's conjecture* or *Ehrenfeucht's compactness property.*

Theorem 2.4.10. *Every system of word equations is equivalent to one of its finite subsets. Consequently, every independent system of word equations is finite.*

Exercises

2.4.1. Let $L = \{a, ba, abb, bbba\}$. Find $r_{\mathrm{c}}(L)$ and $r_{\mathrm{f}}(L)$.

2.4.2. Give an example of an independent system of three constant-free three-variable word equations.

2.4.3. Let $\Xi = \{X, Y, Z\}$ and $\Sigma = \{a, b\}$. Show that the pair of equations $S = \{(XYZ, ZYX), (XYYZ, ZYYX)\}$ is independent. (It has a nonperiodic solution by Example 2.1.3.)

2.4.4. Let $\Xi = \{X, Y, Z\}$ and $\Sigma = \{a, b\}$. Solve the pair of word equations $\{(XYXZY, YZXYX), (XZY, ZYX)\}$.

*2.4.5.** Prove Theorem 2.4.8.

Chapter 3

Infinite words

3.1 Constructing infinite words

An *infinite word* over an alphabet Σ is an infinite sequence of elements of Σ. An infinite word $(a_1, a_2, a_3, \dots)$ is usually written without the commas and parentheses as $a_1 a_2 a_3 \cdots$. The set of all infinite words over Σ is denoted by Σ^ω.

Remark 3.1.1. The infinite words defined above could be more specifically called *right-infinite*. It would be possible to define also *two-way infinite* words

$$\cdots a_{-3} a_{-2} a_{-1} a_0 a_1 a_2 a_3 \cdots ,$$

but we consider only right-infinite words here.

Many of the definitions we gave for finite words can be adapted for infinite words:

- An infinite word is called *k-ary* if it contains at most k different letters.
- The *length* of an infinite word w is $|w| = \infty$.
- The *concatenation* or *product* of a finite word u and an infinite word w, denoted by $u \cdot w$ or uw, is the word consisting of the letters of u followed by the letters of v. In other words, if $u = a_1 \cdots a_m$ and $w = b_1 b_2 b_3 \cdots$, then $uw = a_1 \cdots a_m b_1 b_2 b_3 \cdots$.
- The concatenation wu of an infinite word w and a finite word u is not defined, and neither is the concatenation of two infinite words.
- A finite word $u \in \Sigma^*$ is a *factor* of $w \in \Sigma^\omega$ if $w = xuy$ for some $x \in \Sigma^*$ and $y \in \Sigma^\omega$.
- A finite word $u \in \Sigma^*$ is a *prefix* of $w \in \Sigma^\omega$ if $w = uy$ for some $y \in \Sigma^\omega$.
- An infinite word $u \in \Sigma^\omega$ is a *suffix* of $w \in \Sigma^\omega$ if $w = xu$ for some $x \in \Sigma^*$.

If $u, v \in \Sigma^* \cup \Sigma^\omega$, then we denote their longest common prefix by $u \wedge v$. In the trivial case $u = v \in \Sigma^\omega$, there are arbitrarily long common prefixes and we define $u \wedge u = u$.

Let $(w_n)_{n=0}^\infty$ be a sequence of finite or infinite words. A finite or infinite word w is the *limit* of this sequence if $\lim_{n\to\infty} |w_n \wedge w| = \infty$ or $w_n = w$ for all sufficiently large n. If the sequence $(w_n)_{n=0}^\infty$ has a limit, then the limit is unique and it is denoted by $\lim_{n\to\infty} w_n$. We can also say that the sequence *converges* to the limit. We are mostly interested in the case where the words w_n are finite but the limit is an infinite word.

Remark 3.1.2. A remark for those familiar with topology: Let

$$d : (\Sigma^* \cup \Sigma^\omega) \times (\Sigma^* \cup \Sigma^\omega),\ d(u, v) = \begin{cases} 0 & \text{if } u = v \\ 2^{-|u \wedge v|} & \text{if } u \neq v. \end{cases}$$

Then $(\Sigma^* \cup \Sigma^\omega, d)$ is a compact metric space, and a sequence has a limit w in this space if and only if it has a limit w in the sense of the definition given above. Proving this is left as an exercise.

If w_n is a prefix of w_{n+1} for all n, then clearly the sequence $(w_n)_{n=0}^\infty$ has a limit. If u and $u_1, u_2, u_3 \dots$ are finite words, then we can define the infinite power u^ω and the infinite product $\prod_{i=1}^\infty u_n = u_1 u_2 u_3 \cdots$ by

$$u^\omega = \lim_{n \to \infty} u^n, \qquad \prod_{i=1}^\infty u_n = u_1 u_2 u_3 \cdots = \lim_{n \to \infty} u_1 \cdots u_n.$$

Example 3.1.3. Let $\Sigma = \{a, b\}$. Then

$$\lim_{n \to \infty} (ab)^n a = (ab)^\omega, \qquad \lim_{n \to \infty} a^n b^n = a^\omega.$$

An infinite word $a_1 a_2 a_3 \cdots$, where $a_1, a_2, a_3 \ldots \in \Sigma$, is *ultimately periodic* if there exist numbers $k, n \geq 1$ such that $a_{i+k} = a_i$ for all $i \geq n$, and it is *periodic* if we can choose $n = 1$. Clearly, an infinite word w is ultimately periodic if and only if there exist words $u \in \Sigma^*$ and $v \in \Sigma^+$ such that $w = uv^\omega$, and w is periodic if and only if there exists a word $v \in \Sigma^+$ such that $w = v^\omega$. An infinite word is *aperiodic* if it is not ultimately periodic.

Example 3.1.4. Let $\Sigma = \{a, b\}$. The word $a(ba)^\omega = (ab)^\omega$ is periodic. It is quite easy to see that the word $a(ab)^\omega$ is ultimately periodic but not periodic, and the word

$$\prod_{i=1}^\infty a^i b = abaabaaabaaaab \cdots$$

is aperiodic. Details are left as an exercise.

In the next two examples, we define two famous infinite words that are used later several times. Like in many other examples in this chapter, we use the alphabet $\{0, 1\}$.

Example 3.1.5. Let $F_0 = 0$, $F_1 = 01$, and $F_{n+2} = F_{n+1} F_n$ for all $n \geq 0$. Clearly F_n is a prefix of F_{n+1} for all $n \geq 0$, so the limit

$$F = \lim_{n \to \infty} F_n = 0100101001001010010100100101001001 \cdots$$

exists, and it is infinite. It is called the *Fibonacci word.*

For a word $w \in \{0, 1\}^*$, let $\overline{w}$ be the image of w under the morphism defined by $0 \mapsto 1$, $1 \mapsto 0$, that is, $\overline{w}$ is the word we get from w by swapping 0's and 1's.

Example 3.1.6. Let $T_0 = 0$ and $T_{n+1} = T_n \overline{T_n}$ for all $n \geq 0$. Clearly T_n is a prefix of T_{n+1} for all n, so the limit

$$T = \lim_{n \to \infty} T_n = 01101001100101101001011001101001 \cdots$$

exists, and it is infinite. It is called the *Thue–Morse word.*

For the rest of the chapter, let F, F_n, T, T_n be as in the previous two examples. The next theorem gives an alternative way to define the Thue–Morse word.

Theorem 3.1.7. *For $n \geq 0$, let $a_n \in \{0, 1\}$ be the sum of the digits in the binary representation of n modulo 2. The infinite word $a_0 a_1 a_2 \cdots$ is the Thue–Morse word.*

Proof. We prove by induction that $T_n = a_0 \cdots a_{2^n - 1}$ for all $n \geq 0$. The case $n = 0$ is clear. Let $n \geq 1$. If $i \in \{0, \ldots, 2^{n-1} - 1\}$, then we obtain the binary representation of $i + 2^{n-1}$ from the binary representation of i by adding 10^k for some $k \geq 0$ to the beginning. Consequently, $a_{i+2^{n-1}} = a_i + 1 \bmod 2$. It follows that $a_{2^{n-1}} \cdots a_{2^n - 1} = \overline{a_0 \cdots a_{2^{n-1}-1}}$. From the induction hypothesis it follows that $a_0 \cdots a_{2^n - 1} = T_{n-1}\overline{T_{n-1}} = T_n$. This completes the proof. □

We continue with some other examples of infinite words.

Example 3.1.8. Let $P_0 = \varepsilon$ and $P_{n+1} = P_n 0 P_n 1 P_n 0 P_n$ for all $n \geq 0$. Clearly P_n is a prefix of P_{n+1} for all n, so the limit

$$\lim_{n \to \infty} P_n = 010001010100010 \cdots$$

exists, and it is infinite. It is called the *period-doubling word.* It is closely related to the Thue–Morse word, as we see in the exercises.

Example 3.1.9. Let $S_0 = 0$ and $S_{n+1} = S_n 1^{3^n} S_n$ for all $n \geq 0$. Clearly S_n is a prefix of S_{n+1} for all n, so the limit

$$\lim_{n \to \infty} S_n = 010111010111111111010111010 \cdots$$

exists, and it is infinite. It is called the *Sierpinski word* or the *Cantor word.*

Example 3.1.10. There exists a unique infinite word $a_0 a_1 a_2 \cdots \in \{1, 2\}^\omega$ such that

$$a_0 a_1 a_2 \cdots = \prod_{n=0}^{\infty} 1^{a_{2n}} 2^{a_{2n+1}} = 1^{a_0} 2^{a_1} 1^{a_2} 2^{a_3} \cdots = 12211212212211211221211 \cdots .$$

It is called the *Oldenburger–Kolakoski word.*

All the named words we have introduced are aperiodic.

Theorem 3.1.11. *The Fibonacci word, the Thue–Morse word, the period-doubling word, the Sierpinski word and the Oldenburger–Kolakoski word are aperiodic.*

Proof. We prove the claims about the Fibonacci word and the Thue–Morse word; the others are left as an exercise.

First, we consider the Fibonacci word F, assume that it is ultimately periodic, and derive a contradiction. We can write $F = uv^\omega$ where v is primitive and u is of minimal length. Let N be the smallest integer such that $|F_N| \geq |uv|$. Clearly $N \geq 1$. For all $n \geq N$, we can write $F_n = uv^{k_n} v_n$, where k_n is a positive integer and v_n is a proper prefix of v. Then

$$uv^{k_{n+2}} v_{n+2} = F_{n+2} = F_{n+1} F_n = uv^{k_{n+1}} v_{n+1} uv^{k_n} v_n$$

and thus $v_{n+1} uv$ is a prefix of a power of v. It follows from Lemma 1.2.5 that $v_{n+1} u$ must be a power of v. If $u \neq \varepsilon$, then this contradicts the minimality of u, so $u = \varepsilon$, and then also $v_{n+1} = \varepsilon$. We have shown that $F_{n+1} \in v^*$ for all $n \geq N$. But if $F_{N+2}, F_{N+1} \in v^*$, then also $F_N \in v^*$, and then $F_{N-1} \in v^*$, which is a contradiction by the definition of N. This contradiction shows that F is aperiodic.

Then, we consider the Thue–Morse word $T = a_0 a_1 a_2 \cdots$, assume that it is ultimately periodic, and derive a contradiction. We could use a somewhat similar strategy as for the Fibonacci word, but we use Theorem 3.1.7 instead. Let k, n be such that $a_{i+k} = a_i$ for all $i \geq n$. Let m be such that $mk - 1 \geq n$. Then

$$a_{2(mk-1)+1} = a_{mk-1+mk} = a_{mk-1}.$$

On the other hand, we obtain the binary representation of $2(mk-1)+1$ from the binary representation of $mk-1$ by adding the digit 1 to the end, so by Theorem 3.1.7,

$$a_{2(mk-1)+1} = a_{mk-1} + 1 \bmod 2.$$

This contradiction shows that T is aperiodic. □

Exercises

3.1.1. Let u and $w_0, w_1, w_2, \ldots$ be finite words and let $w = \lim_{n\to\infty} w_n$ be an infinite word. If u is a factor of w_i for all i, then is u necessarily a factor of w? If u is a factor of w, then is u necessarily a factor of w_i for some i?

3.1.2. Let $u, v \in \Sigma^+$. Show that if $|u^\omega \wedge v^\omega| \geq |uv|$, then $u^\omega = v^\omega$.

3.1.3. Justify in detail the claims of Example 3.1.4 about $a(ab)^\omega$ being not periodic and $\prod_{i=1}^{\infty} a^i b$ being aperiodic.

3.1.4. For $n \geq 1$, let k_n be the largest integer k such that $2^k | n$, and let $b_n = k_n \bmod 2$. Show that $b_1 b_2 b_3 \cdots$ is the period-doubling word.

3.1.5. Let $a_0 a_1 a_2 \cdots$ be the Thue–Morse word and $b_1 b_2 b_3 \cdots$ the period-doubling word. Show that $b_n \equiv a_n + a_{n-1} + 1 \pmod 2$ for all $n \geq 1$.

3.1.6. Find out what the Sierpinski triangle and the Cantor set are (if you do not know already), and think how they are similar to the Sierpinski word.

3.1.7. Let $F_{2n-1} = G_{2n-1}01$ and $F_{2n} = G_{2n}10$ for all $n \geq 1$. Show that G_n is a palindrome for all n.

3.1.8. Show that if u is a factor of the Fibonacci word, then so is u^R.

*3.1.9.** Show that the word equation $(X01Y, Y10X)$ over $\{X, Y\}$ has infinitely many solutions h such that $h(X)$ and $h(Y)$ are prefixes of the Fibonacci word.

*3.1.10.** Complete the proof of Theorem 3.1.11.

*3.1.11.** (Topology.) Prove the claims in Remark 3.1.2.

3.2 Morphic words

If Σ and Γ are alphabets and $h : \Sigma^* \to \Gamma^*$ is a morphism, then we can define the image of an infinite word $w = a_1 a_2 a_3 \cdots \in \Sigma^\omega$ under h by

$$h(w) = h(a_1)h(a_2)h(a_3) \cdots .$$

A finite or infinite word w is a *fixed point* of a morphism h if $h(w) = w$.

Infinite words are often constructed by iterating a morphism $h : \Sigma^* \to \Sigma^*$ on a letter a, that is, as the limit of the sequence $h(a), h^2(a), h^3(a), \ldots$. The limit can be denoted by $h^\omega(a)$. In some cases, the limit is finite or does not exist, but under certain simple conditions, the sequence is guaranteed to converge to an infinite word.

A morphism $h : \Sigma^* \to \Sigma^*$ is *prolongable on* $a \in \Sigma$ if $h(a) = au$ for some $u \in \Sigma^+$ and $h^n(u) \neq \varepsilon$ for all $n \geq 0$ (note that if h is *nonerasing*, that is, $h(b) \neq \varepsilon$ for all $b \in \Sigma$, as is the case in most of our examples, then the second condition $h^n(u) \neq \varepsilon$ is automatically satisfied). Next we prove that if h is prolongable on a, then $h^\omega(a)$ is an infinite word and a fixed point of h.

Theorem 3.2.1. *Let $h : \Sigma^* \to \Sigma^*$ be prolongable on $a \in \Sigma$. Then*

$$h^\omega(a) = auh(u)h^2(u)h^3(u) \cdots ,$$

and $h^\omega(a)$ is the unique fixed point of h beginning with a.

Proof. We can prove by induction that $h^n(a) = auh(u) \cdots h^{n-1}(u)$ for all $n \geq 1$: Clearly this is true for $n = 1$, and if it is true for n, then

$$h^{n+1}(a) = h^n(au) = h^n(a)h^n(u) = auh(u) \cdots h^{n-1}(u)h^n(u).$$

This proves that $h^\omega(a) = auh(u)h^2(u)h^3(u) \cdots$.

It is clear that $h^\omega(a)$ begins with a and $h(h^\omega(a)) = h^\omega(a)$. On the other hand, if $w = h(w)$ and $\mathrm{pref}_1(w) = a$, then we can prove by induction that w has a prefix $h^n(a)$ for all $n \geq 0$: Clearly this is true for $n = 0$, and if it is true for n, then $w = h(w)$ has a prefix $h(h^n(a)) = h^{n+1}(a)$. This proves that $w = h^\omega(a)$. □

An infinite word $w \in \Sigma^\omega$ is *pure morphic* if there exists a letter $a \in \Sigma$ and a morphism $h : \Sigma^* \to \Sigma^*$ prolongable on a such that $w = h^\omega(a)$. An infinite word $w \in \Sigma^\omega$ is *morphic* if there exists an alphabet Γ, a pure morphic word $u \in \Gamma^\omega$, and a morphism $g : \Gamma^* \to \Sigma^*$ such that $w = g(u)$.

Next we are going to see that the Fibonacci word and the Thue–Morse word are pure morphic.

Theorem 3.2.2. *The Fibonacci word is the fixed point $\phi^\omega(0)$ of the morphism*

$$\phi : \{0,1\}^* \to \{0,1\}^*, \ \phi(0) = 01, \ \phi(1) = 0.$$

Proof. We can prove by induction that $\phi^n(0) = F_n$ and $\phi^n(1) = F_{n-1}$ for all $n \geq 1$: Clearly this is true for $n = 1$, and if it is true for n, then

$$\phi^{n+1}(0) = \phi^n(01) = \phi^n(0)\phi^n(1) = F_n F_{n-1} = F_{n+1}, \qquad \phi^{n+1}(1) = \phi^n(0) = F_n.$$

This proves the theorem. □

Theorem 3.2.3. *The Thue–Morse word is the fixed point $\tau^\omega(0)$ of the morphism*

$$\tau : \{0,1\}^* \to \{0,1\}^*, \ \tau(0) = 01, \ \tau(1) = 10.$$

Proof. We can prove by induction that $\tau^n(0) = T_n$ and $\tau^n(1) = \overline{T_n}$ for all $n \geq 0$: Clearly this is true for $n = 0$, and if it is true for n, then

$$\begin{aligned}\tau^{n+1}(0) &= \tau^n(01) = \tau^n(0)\tau^n(1) = T_n\overline{T_n} = T_{n+1},\\ \tau^{n+1}(1) &= \tau^n(10) = \tau^n(1)\tau^n(0) = \overline{T_n}T_n = \overline{T_{n+1}}.\end{aligned}$$

This proves the theorem. □

For the rest of the chapter, let ϕ and τ be the morphisms defined in the previous two theorems.

We continue with some other examples of morphic words.

Example 3.2.4. It can be proved that the period-doubling word is the fixed point $h^\omega(0)$ of the morphism

$$h : \{0,1\}^* \to \{0,1\}^*, \ h(0) = 01, \ h(1) = 00.$$

Example 3.2.5. It can be proved that the Sierpinski word is the fixed point $h^\omega(0)$ of the morphism

$$h : \{0,1\}^* \to \{0,1\}^*, \ h(0) = 010, \ h(1) = 111.$$

Example 3.2.6. Let

$$h : \{0,1\}^* \to \{0,1\}^*, \ h(0) = 001, \ h(1) = 110,$$

be a morphism. Its fixed point

$$h^\omega(0) = 001001110001001110110110001\cdots$$

is called the *Mephisto waltz word.*

Example 3.2.7. Let

$$h : \{0,1,2\}^* \to \{0,1,2\}^*, \ h(0) = 01, \ h(1) = 02, \ h(2) = 0,$$

be a morphism. Its fixed point

$$h^\omega(0) = 010201001020101020100102\cdots$$

is called the *Tribonacci word.*

Example 3.2.8. Let

$$\begin{aligned}&h : \{0,1,2,3\}^* \to \{0,1,2,3\}^*, \ h(0) = 0, \ h(1) = 1, \ h(2) = 203, \ h(3) = 213,\\ &g : \{0,1,2,3\}^* \to \{0,1\}^*, \ g(0) = 0, \ g(1) = 1, \ h(2) = h(3) = \varepsilon,\end{aligned}$$

be morphisms. The morphic word

$$g(h^\omega(2)) = 001001100011011\cdots$$

is called the *paperfolding word.* It can also be defined with the help of the morphisms

$$\begin{aligned}&h_1 : \{0,1,2,3\}^* \to \{0,1,2,3\}^*, \ h(0) = 02, \ h(1) = 03, \ h(2) = 12, \ h(3) = 13,\\ &g_1 : \{0,1,2,3\}^* \to \{0,1\}^*, \ g(0) = g(2) = 0, \ g(1) = g(3) = 1,\end{aligned}$$

as the word $g_1(h_1^\omega(0))$.

Next, we study the relations between ultimately periodic and morphic words.

Theorem 3.2.9. *Every periodic word is pure morphic, and every ultimately periodic word is morphic. There are ultimately periodic words that are not pure morphic.*

Proof. First, consider an arbitrary periodic word $(au)^\omega \in \Sigma^\omega$, where $a \in \Sigma$ and $u \in \Sigma^*$. Let $h : \Sigma^* \to \Sigma^*$ be the morphism defined by $h(b) = (au)^2$ for all $b \in \Sigma$. Then $h^\omega(a) = (au)^\omega$, so $(au)^\omega$ is pure morphic.

Then, consider an arbitrary ultimately periodic word $uv^\omega \in \Sigma^\omega$, where $u \in \Sigma^*$ and $v \in \Sigma^+$. Let

$$h : \{0,1\}^* \to \{0,1\}^*,\ h(0) = 01,\ h(1) = 1,$$
$$g : \{0,1\}^* \to \Sigma^*,\ g(0) = u,\ g(1) = v,$$

be morphisms. Then $g(h^\omega(0)) = uv^\omega$, so uv^ω is morphic.

Finally, consider the ultimately periodic word $w = 010^\omega \in \{0,1\}^\omega$. If $h : \{0,1\}^* \to \{0,1\}^*$ is a morphism prolongable on 0, then $h(w)$ begins with 00 or contains infinitely many 1's, so $h(w) \neq w$. This shows that w cannot be pure morphic. □

Exercises

3.2.1. Let $h : \{0,1\}^* \to \{0,1\}^*,\ h(0) = 010,\ h(1) = 1$, be a morphism. Show that $h^\omega(0)$ is periodic.

3.2.2. Prove the claims in Examples 3.2.4 and 3.2.5.

3.2.3. Find a way to define the Mephisto waltz word using recurrence relations similar to those in Example 3.1.6.

3.2.4. Let $w_0 = 0$, $w_1 = 010$, and $w_{n+2} = w_{n+1}w_nw_{n+1}$ for all $n \geq 0$. Find a morphism h such that $h^\omega(0) = \lim_{n\to\infty} w_n$.

3.2.5. For which words $u \in \{0,1\}^*$ is the word $0u0^\omega$ pure morphic? What about the word $0u1^\omega$?

3.2.6. Show that the aperiodic word in Example 3.1.4 is morphic but not pure morphic.

*3.2.7.** Let $a_0a_1a_2\cdots$ be the Thue–Morse word. Let N be a positive integer. Show that

$$\sum_{k=0}^{2^N-1} (-1)^{a_k} k^n = 0$$

for all $n \in \{0, \ldots, N-1\}$.

*3.2.8.** Find out where the name of the paperfolding word comes from.

*3.2.9.** (Programming.) Write a program that takes as input a letter $a \in \Sigma$, a morphism $h : \Sigma^* \to \Sigma^*$ prolongable on a, and a number $n \in \mathbb{Z}_{\geq 0}$, and returns the prefix of length n of $h^\omega(a)$.

3.3 Repetition-freeness

Let $\alpha \in \mathbb{R}$, $\alpha > 1$. A finite or infinite word is *α-free* if it does not have a nonempty factor that is a q-power for any rational number $q \geq \alpha$. A finite or infinite word is *α^+-free* if it does not have a nonempty factor that is a q-power for any rational number $q > \alpha$. 2-free words can be called *square-free*, 3-free words *cube-free*, and 2^+-free words *overlap-free*.

A morphism h is *α-free*, if $h(w)$ is α-free for all α-free finite words w. Similarly, h is *α^+-free*, if $h(w)$ is α^+-free for all α^+-free finite words w.

Lemma 3.3.1. *Let $\alpha \in \mathbb{R}$, $\alpha > 1$. Let $h : \Sigma^* \to \Sigma^*$ be an α-free (α^+-free) morphism prolongable on $a \in \Sigma$. Then $h^\omega(a)$ is α-free (α^+-free, respectively).*

Proof. The word $h^0(a) = a$ is α-free. If h is α-free, then it follows by induction that $h^n(a)$ is α-free for all n. Every factor of $h^\omega(a)$ is a factor of $h^n(a)$ for some n, so also $h^\omega(a)$ is α-free. The claim about α^+-freeness can be proved in the same way. □

Example 3.3.2. Let $\alpha > 1$ and $n = \lceil \alpha \rceil$. The word $1^{n-1}0$ is α-free, but the word $\phi(1^{n-1}0) = 0^n 1$ is not α-free. This shows that ϕ is not α-free for any α. However, the Fibonacci word is known to be $(2 + \varphi)$-free, where $\varphi = (1 + \sqrt{5})/2$ is the golden ratio (this is more difficult to prove than the repetition-freeness results we prove in this section). It follows that the converse of Lemma 3.3.1 does not hold, that is, $h^\omega(a)$ can be α-free even if h is not.

For which numbers α does there exist an infinite α-free word? The answer depends on the size of the alphabet. First, we consider the binary case. Every word of length 4 over $\{0, 1\}$ contains one of the factors 00, 11, 0101, 1010, so a binary infinite word cannot be square-free. The next theorem shows that there are cube-free binary infinite words.

Theorem 3.3.3. *The morphism $h : \{0, 1\}^* \to \{0, 1\}^*$, $h(0) = 010$, $h(1) = 011$, and therefore also the infinite word $h^\omega(0)$, is cube-free.*

Proof. We have to prove that if the image $h(w)$ of some $w \in \{0, 1\}^*$ contains a nonempty cube u^3 as a factor, then w is not cube-free. Let $w = a_1 \cdots a_n$ and

$$h(w) = h(a_1) \cdots h(a_n) = 01a_1 \cdots 01a_n = b_1 \cdots b_{3n},$$

where $a_1, \ldots, a_n, b_1, \ldots, b_{3n} \in \{0, 1\}$. Then $b_i = 0$ if $i \equiv 1 \pmod 3$, $b_i = 1$ if $i \equiv 2 \pmod 3$, and $b_i = a_{i/3}$ if $i \equiv 3 \pmod 3$. Let $|u| = m$ and $u^3 = b_j \cdots b_{j+3m-1}$.

If $3 \nmid m$, then j, $j + m$, $j + 2m$ are pairwise distinct modulo 3 and thus at least one of b_j, b_{j+m}, b_{j+2m} is 0 and at least one is 1. On the other hand, $\mathrm{pref}_1(u) = b_j = b_{j+m} = b_{j+2m}$, which is a contradiction.

If $3|m$, $k = m/3$, and $i = \lceil j/3 \rceil$, then

$$u = xa_i y \cdots xa_{i+k-1} y = xa_{i+k} y \cdots xa_{i+2k-1} y = xa_{i+2k} y \cdots xa_{i+3k-1} y,$$

where

$$(x, y) = \begin{cases} (01, \varepsilon) & \text{if } j \equiv 1 \pmod 3 \\ (1, 0) & \text{if } j \equiv 2 \pmod 3 \\ (\varepsilon, 01) & \text{if } j \equiv 3 \pmod 3. \end{cases}$$

It follows that $a_i \cdots a_{i+3k-1}$ is a cube, so w is not cube-free. □

Example 3.3.4. The word $h^\omega(0)$ in Theorem 3.3.3 is not α-free for any $\alpha < 3$. We can prove by induction that, for all $n \geq 0$, there exists a word u of length $3^n - 1$ such that $u0u0u$ is a factor of $h^\omega(0)$: This is true for $n = 0$ because 00 is a factor, and if $u0u0u$ is a factor and $|u| = n$, then $h^\omega(0)$ has the factor

$$h(u0u0u)01 = h(u)010h(u)010h(u)01 = v0v0v,$$

where $v = h(u)01$ and thus $|v| = 3|u| + 2 = 3^{n+1} - 1$. This means that, for all $n \geq 0$, $h^\omega(0)$ contains a $(3^{n+1} - 1)/3^n$-power as a factor, so it cannot be α-free for any $\alpha < 3$.

Even though binary infinite words cannot be square-free, they can be overlap-free.

Theorem 3.3.5. *The morphism τ, and therefore also the Thue–Morse word, is overlap-free.*

Proof. We have to prove that if the image $\tau(w)$ of some $w \in \{0,1\}^*$ has a factor $auaua$, where $a \in \{0,1\}$ and $u \in \{0,1\}^*$, then w is not overlap-free. Let $w = a_1 \cdots a_n$ and

$$h(w) = \tau(a_1) \cdots \tau(a_n) = a_1\overline{a_1} \cdots a_n\overline{a_n} = b_1 \cdots b_{2n},$$

where $a_1, \ldots, a_n, b_1, \ldots, b_{2n} \in \{0,1\}$. Let $|u| = m - 1$.

If $2 \nmid m$, then there exist i, j, k such that

$$aua = \tau(a_i) \cdots \tau(a_{i+k}) = a\tau(a_j) \cdots \tau(a_{j+k-1})a.$$

Because aua has even length, it must have a factor bb for some $b \in \{0,1\}$. This is a contradiction because $\tau(0) \neq bb \neq \tau(1)$.

If $2|m$ and $k = m/2$, then there exists an index i such that

$$au = \tau(a_i) \cdots \tau(a_{i+k-1}) = \tau(a_{i+k}) \cdots \tau(a_{i+2k-1}), \qquad a_i = a_{i+k} = a_{i+2k},$$

or

$$ua = \tau(a_{i+1}) \cdots \tau(a_{i+k}) = \tau(a_{i+k+1}) \cdots \tau(a_{i+2k}), \qquad a_i = a_{i+k} = a_{i+2k},$$

so w has the factor $a_iva_iva_i$, where $v = a_{i+1} \cdots a_{i+k-1}$. □

In the case of a ternary alphabet, there exist square-free infinite words.

Theorem 3.3.6. *Let $h : \{0,1,2\}^* \to \{0,1\}^*$, $h(0) = 0$, $h(1) = 01$, $h(2) = 011$ be a morphism. There exists a unique infinite word $w \in \{0,1,2\}^\omega$ such that $h(w)$ is the Thue–Morse word. This word w is square-free.*

Proof. It is clear that the Thue–Morse word (like any other infinite word in $\{0,1\}^\omega$ that begins with 0) can be written uniquely as a product $u_1u_2u_3 \cdots$, where $u_i \in 01^*$ for all i. Because the Thue–Morse word is cube-free, $u_i \in \{0, 01, 011\}$ for all i. Thus $T = h(w)$ for some unique w. If w has a nonempty factor u^2, then $h(w)$ has the factor $h(u)^2 0$, so $h(w)$ is not overlap-free, which is a contradiction. □

The *repetition threshold* for k-ary alphabets, denoted by $RT(k)$, is the infimum of the numbers α such that there exists an α-free k-ary infinite word. We have seen that $RT(2) = 2$. The exact value of $RT(k)$ is known for all other values of k as well. The following result was an open conjecture, known as *Dejean's conjecture*, for a long time before it was proved. We state it here without proof.

Theorem 3.3.7.

$$RT(k) = \begin{cases} k/(k-1) & \text{if } k = 2 \text{ or } k \geq 5 \\ 7/4 & \text{if } k = 3 \\ 7/5 & \text{if } k = 4. \end{cases}$$

A finite or infinite word w *avoids* a language L if w has no nonempty factors in L. A language L is *avoidable* on an alphabet Σ if there exists an infinite word over Σ that avoids L. So we have proved that the set of cubes is avoidable on a binary alphabet, and the set of squares is avoidable on a ternary alphabet. A huge number of different avoidability questions have been studied. As an example, we consider the avoidability of so-called abelian powers in the next section. Some other questions that have been studied are mentioned below.

- Let Γ, Σ be alphabets. For some word $u \in \Gamma^*$, is the set of images of u under all nonerasing morphisms $\Gamma^* \to \Sigma^*$ avoidable? For example, if $\Gamma = \Sigma = \{0,1\}$, the answer is known to be negative for $u = 0011$ and positive for $u = 00110$.

- If the alphabet is a subset of $\mathbb{Z}$, a word $a_1 \cdots a_{2n}$ such that $a_1 + \cdots + a_n = a_{n+1} + \cdots + a_{2n}$ is called an *additive square*. Is the set of additive squares avoidable on some alphabet? This is a famous open question. It is known that the analogously defined additive cubes are avoidable on some ternary alphabets.

- For some avoidable language L, what is the growth rate of the number of words of length n that avoid L? For example, it is known that the number of cube-free binary and square-free ternary words grows exponentially, but the number of overlap-free binary words grows only polynomially.

- For some $\alpha > 1$, how can we check whether a given morphism is α-free? For example, a simple algorithm is known for $\alpha = 2$.

Exercises

3.3.1. Show that the morphism $h : \{0,1\}^* \to \{0,1\}^*$, $h(0) = 001$, $h(1) = 011$, is cube-free.

3.3.2. Show that the Thue–Morse word does not have nonempty square prefixes.

3.3.3. Show that there exists a binary infinite word that does not have nonunary square factors.

3.3.4. Show that $RT(k) \geq k/(k-1)$ for all $k \geq 2$ (without using Theorem 3.3.7).

3.3.5. Find the shortest cube that is a factor of the Fibonacci word.

*3.3.6.** Show that the Fibonacci word is not α-free for any $\alpha < (2 + \varphi)$, where $\varphi = (1 + \sqrt{5})/2$ is the golden ratio.

*3.3.7.** (Programming.) Write a program that counts the number of cube-free words in $\{0,1\}^{20}$ and square-free words in $\{0,1,2\}^{20}$.

3.4 Abelian repetition-freeness

Let $n \geq 2$ be an integer. An *abelian n-power* is a finite word of the form $u_1 \cdots u_n$, where $u_1, \ldots, u_n$ are abelian equivalent words. Abelian 2-powers can be called *abelian squares* and abelian 3-powers can be called *abelian cubes.*

Example 3.4.1. All squares are abelian squares. Of all the words in $\{0,1\}^4$, the only abelian squares which are not squares are 0110 and 1001.

Before moving on to abelian repetition-freeness, we introduce a couple of notions that can be useful when studying abelian equivalence, and also in other contexts. To simplify notation, let $k \geq 1$ and $\Sigma = \{0, \ldots, k-1\}$ for the rest of the section.

The *Parikh vector* of a word $u \in \Sigma^*$, denoted by $\Pi(u)$, is the k-dimensional column vector $(|u|_0, \ldots, |u|_{k-1})^T$. Clearly, $\Pi(uv) = \Pi(u) + \Pi(v)$, and words are abelian equivalent if and only if they have the same Parikh vector.

The *incidence matrix* of a morphism $h : \Sigma^* \to \Sigma^*$, denoted by $M(h)$, is the $k \times k$ matrix with columns $\Pi(h(0)), \ldots, \Pi(h(k-1))$, that is,

$$M(h) = \begin{pmatrix} |h(0)|_0 & \cdots & |h(k-1)|_0 \\ \cdots & \cdots & \cdots \\ |h(0)|_{k-1} & \cdots & |h(k-1)|_{k-1} \end{pmatrix}$$

It is easy to see that $\Pi(h(u)) = M(h)\Pi(u)$.

Example 3.4.2. Let $f_{-2} = 0$, $f_{-1} = 1$, and $f_n = f_{n-1} + f_{n-2}$ for all $n \geq 0$. Then $|F_n| = f_n$ and $\Pi(F_n) = (f_{n-1}, f_{n-2})^T$ for all $n \geq 0$. Because $F_{n+1} = \phi(F_n)$, we get

$$\Pi(F_{n+1}) = M(\phi)\Pi(F_n) = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} f_{n-1} \\ f_{n-2} \end{pmatrix} = \begin{pmatrix} f_{n-1} + f_{n-2} \\ f_{n-1} \end{pmatrix} = \begin{pmatrix} f_n \\ f_{n-1} \end{pmatrix},$$

as expected.

Lemma 3.4.3. *Let $h : \Sigma^* \to \Sigma^*$ be a morphism and $u, v \in \Sigma^*$. If u and v are abelian equivalent, then $h(u)$ and $h(v)$ are abelian equivalent. If $h(u)$ and $h(v)$ are abelian equivalent and $M(h)$ is invertible, then u and v are abelian equivalent.*

Proof. If $\Pi(u) = \Pi(v)$, then

$$\Pi(h(u)) = M(h)\Pi(u) = M(h)\Pi(v) = \Pi(h(v)),$$

so the first claim is true. If $\Pi(h(u)) = \Pi(h(v))$ and $M(h)$ is invertible, then

$$\Pi(u) = M(h)^{-1}\Pi(h(u)) = M(h)^{-1}\Pi(h(v)) = \Pi(v),$$

so the second claim is true. □

A finite or infinite word is *abelian n-free* if it does not have a nonempty factor that is an abelian n-power. Abelian 2-free words can be called *abelian square-free* and abelian 3-free words *abelian cube-free.*

If the alphabet size is fixed, then what is the smallest integer n for which there exists an infinite abelian n-free word? It can be proved that the answer is $n = 4$ in the binary case, $n = 3$ in the ternary case, and $n = 2$ in the 4-ary case. We concentrate mostly on the binary case.

Theorem 3.4.4. *The fixed point $h^\omega(0)$ of the morphism*

$$h : \{0,1\}^* \to \{0,1\}^*,\ h(0) = 011,\ h(1) = 0001,$$

is abelian 4-free.

Proof. We assume that $h^\omega(0)$ is not abelian 4-free and derive a contradiction. Let $x_1x_2x_3x_4$, where x_1, x_2, x_3, x_4 are abelian equivalent, be the shortest nonempty factor of $h^\omega(0)$ that is an abelian 4-power. We can easily check that $|x_1| \le 3$ is not possible. Because $h^\omega(0)$ is a fixed point of h, it has a factor $a_1w_1a_2w_2a_3w_3a_4w_4a_5$ such that $w_i \in \{0,1\}^*$ for $i \in \{1,2,3,4\}$, $a_i \in \{0,1\}$, $h(a_i) = y_iz_i$, $y_i \in \{0,1\}^*$, $z_i \in \{0,1\}^+$ for $i \in \{1,2,3,4,5\}$, and $x_i = z_ih(w_i)y_{i+1}$ for $i \in \{1,2,3,4\}$ (see Figure 3.1).

Let us define a function $f : \{0,1\}^* \to \mathbb{Z}$, $f(u) = |u|_0 + 2|u|_1$. Then $f(uv) = f(u) + f(v)$ for all words u, v. If u and v are abelian equivalent, then $f(u) = f(v)$, so $f(x_i) = f(x_1)$ for all i. From $f(h(0)) = f(h(1)) = 5$ it follows that $f(h(u)) \equiv 0 \pmod 5$ for all words u.

For $i \in \{1, \dots, 4\}$, we have

$$\begin{aligned} f(y_{i+1}) &= f(x_i) - f(h(w_i)) - f(z_i) \\ &= f(x_i) - f(h(w_i)) - f(h(a_i)) + f(y_i) \\ &\equiv f(x_1) + f(y_i) \pmod 5. \end{aligned}$$

Thus the sequence $f(y_1), f(y_2), f(y_3), f(y_4), f(y_5)$ is an arithmetic progression modulo 5. For all i, y_i is a proper prefix of $h(0)$ or $h(1)$, so $y_i \in \{\varepsilon, 0, 00, 01, 000\}$, and therefore $f(y_i) \in \{0,1,2,3\}$. In particular, $f(y_i) \not\equiv 4 \pmod 5$, so the only possible arithmetic progression modulo 5 is a constant one, that is,

$$f(y_1) = f(y_2) = f(y_3) = f(y_4) = f(y_5).$$

It follows that if $f(y_1) \in \{0,1,2\}$, then it must be $y_1 = \dots = y_5$, and if $f(y_1) = 3$, then it must be $z_1 = \dots = z_5 = 1$.

Let us assume that $y_1 = \dots = y_5$ (the case $z_1 = \dots = z_5$ is similar). For all $i \in \{1,2,3,4\}$, $y_ix_i = h(a_i)h(w_i)y_{i+1}$ and $y_i = y_{i+1}$, and then x_i and $h(a_i)h(w_i)$ are abelian equivalent. Thus the words $h(a_iw_i)$ are abelian equivalent for $i \in \{1,2,3,4\}$. The incidence matrix of h is invertible, so it follows from Lemma 3.4.3 that the words a_iw_i are abelian equivalent for $i \in \{1,2,3,4\}$. But then $a_1w_1a_2w_2a_3w_3a_4w_4$ is an abelian 4-power and a factor of $h^\omega(0)$, which contradicts the minimality of $x_1x_2x_3x_4$. This contradiction completes the proof. □

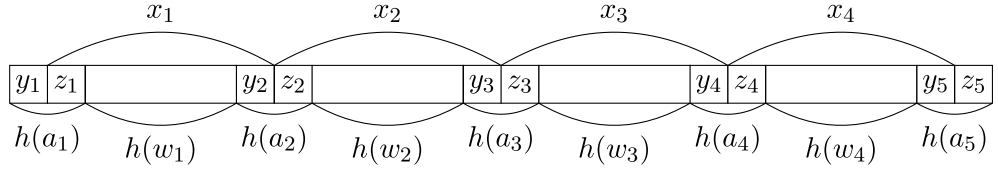


Figure 3.1: Illustration related to the proof of Theorem 3.4.4.

Example 3.4.5. Let us determine the longest binary abelian cube-free words. By symmetry, we can assume that the first letter is 0. Then we can continue by 0 or 1. After 00, we must continue by 1, because 000 is a cube. After 001, we can continue by 0 or 1. After 0010, we can continue by 0 or 1. After 00100, we must continue by 1. After 001001, we can continue by 0 or 1. After 0010010, we must continue by 0, because 100101 is an abelian cube. The word 00100100 cannot be extended, so we must backtrack to 0010011, which also cannot be extended, so we must backtrack to 00101. After 00101, we can continue by 0 or 1. Continuing the search gives the tree in Figure 3.2, from which we see that the longest words are of length 9: 001101100, 001101101, 010010011.

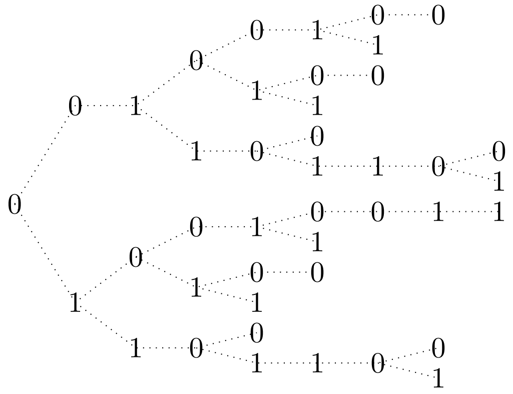


Figure 3.2: Abelian cube-free binary words starting with 0.

We state two other results without proof.

Theorem 3.4.6. *The fixed point* $h^\omega(0)$ *of the morphism*

$$h : \{0,1,2\}^* \to \{0,1\}^*,\ h(0) = 0012,\ h(1) = 112,\ h(2) = 022,$$

is abelian cube-free.

Theorem 3.4.7. *The fixed point* $h^\omega(0)$ *of the morphism*

$$\begin{aligned} &h : \{0,1,2,3\}^* \to \{0,1,2,3\}^*,\\ &h(0) = 0120232123203231301020103101213121021232 02\\ &\qquad\quad 1013010203212320231210212320232132303132120,\\ &h(1) = 1231303230310302012131210212320232132303 13\\ &\qquad\quad 2120121310323031302321323031303203010203231,\\ &h(2) = 2302010301021013123202321323031303203010 20\\ &\qquad\quad 3231232021030102013032030102010310121310302,\\ &h(3) = 3013121012132120230313032030102010310121 31\\ &\qquad\quad 0302303132101213120103101213121021232021013, \end{aligned}$$

is abelian square-free.

Exercises

3.4.1. How long can a ternary abelian square-free word be?

*3.4.2.** Prove Theorem 3.4.6 using the same strategy as in the proof of Theorem 3.4.4.

*3.4.3.** Let n be a positive integer. Show that every infinite word has a nonempty factor that is abelian equivalent to an n-power.

3.5 Factor complexity

For an infinite word w, let $\operatorname{Fact}_n(w)$ be the set of factors of w of length n. The *factor complexity* of an infinite word w is the function

$$P_w : \mathbb{Z}_+ \to \mathbb{Z}_+,\ P_w(n) = |\operatorname{Fact}_n(w)|.$$

Example 3.5.1. Let $w = 0(01)^\omega$. Then

$$\begin{aligned}\operatorname{Fact}_{2n-1}(w) &= \{0(01)^{n-1}, 0(10)^{n-1}, 1(01)^{n-1}\},\\ \operatorname{Fact}_{2n}(w) &= \{00(10)^{n-1}, (01)^n, (10)^n\}\end{aligned}$$

for all $n \geq 1$, so $P_w(1) = 2$ and $P_w(n) = 3$ for all $n \geq 2$.

A finite word x is a *right special factor* of an infinite word w if there exist at least two letters a such that xa is a factor of w (note that if x is a factor of w, then there always exists at least one such letter).

Example 3.5.2. For the Thue–Morse word T,

$$\operatorname{Fact}_2(T) = \{00, 01, 10, 11\}, \qquad \operatorname{Fact}_3(T) = \{001, 010, 011, 100, 101, 110\},$$

so $P_T(2) = 4$ and $P_T(3) = 6$, and the right special factors of T of length 2 are 01 and 10.

The next lemma gives a connection between right special factors and factor complexity. In particular, it shows that $P_w(n) \leq P_w(n+1)$ for all infinite words w and for all positive integers n.

Lemma 3.5.3. *Let w be an infinite word and n a positive integer. Let w have exactly k right special factors of length n. Then $P_w(n+1) \geq P_w(n) + k$. Moreover, if $k = 0$ or if w is binary, then $P_w(n+1) = P_w(n) + k$.*

Proof. For $u \in \operatorname{Fact}_n(w)$, let

$$A_u = \{a \in \Sigma \mid ua \in \operatorname{Fact}_{n+1}(w)\}.$$

Let R_n be the set of right special factors of length n. Then $|A_u| \geq 2$ for all $u \in R_n$, and $|A_u| = 1$ for all $u \in \operatorname{Fact}_n(w) \smallsetminus R_n$. We have

$$\operatorname{Fact}_{n+1}(w) = \bigcup_{u \in \operatorname{Fact}_n(w)} uA_u,$$

and therefore

$$\begin{aligned}P_w(n+1) &= \sum_{u \in \operatorname{Fact}_n(w)} |A_u|\\ &= |\operatorname{Fact}_n(w)| + \sum_{u \in \operatorname{Fact}_n(w)} (|A_u| - 1)\\ &= P_w(n) + \sum_{u \in R_n} (|A_u| - 1)\\ &\geq P_w(n) + |R_n| = P_w(n) + k.\end{aligned}$$

If $k = 0$, then $R_n = \varnothing$, and if w is binary, then $|A_u| - 1 = 1$ for all $u \in R_n$, so in these cases we have $P_w(n+1) = P_w(n) + k$. □

The following theorem is known as the theorem of Morse and Hedlund.

Theorem 3.5.4. *For an infinite word w, the following are equivalent:*

1. *w is ultimately periodic.*
2. *$P_w(n) = c$ for some constant c and all sufficiently large n.*
3. *$P_w(n) \leq n$ for some n.*

Proof. $1 \implies 2$: Let $w = uv^\omega$. Every factor of w is a prefix of tv^ω for some nonempty suffix t of uv. Thus $P_w(n) \leq |uv|$. The claim follows, because $P_w(n) \leq P_w(n+1)$ for all n.

$2 \implies 3$: Trivial.

$3 \implies 1$: If $P_w(n) \leq n$, then it must be $P_w(m) = P_w(m+1)$ for some $m \leq n$. This means that w has no right special factor of length m by Lemma 3.5.3. Let $w = a_1 a_2 a_3 \cdots$, where all a_i are letters. Some factor of length m must occur twice in w, say, $a_{i+1} \cdots a_{i+m} = a_{j+1} \cdots a_{j+m}$, where $i < j$. But then $a_{i+m+1} = a_{j+m+1}$, because otherwise $a_{i+1} \cdots a_{i+m}$ would be right special. By induction, $a_{i+k} = a_{j+k}$ for all $k \geq 1$, so w is ultimately periodic. □

In Section 3.6, we see that there exist words w such that $P_w(n) = n + 1$ for all $n \geq 1$. It follows from Theorem 3.5.4 that this is the smallest possible factor complexity function of an aperiodic word. The largest possible factor complexity function, on the other hand, is $P_w(n) = k^n$, where k is the size of the alphabet. Proving this is left as an exercise.

Often, we are not interested in the exact values of a factor complexity function, but only in its growth rate. For that purpose, let us define some commonly used notation. Let $f : \mathbb{Z}_+ \to \mathbb{Z}_+$ and $g : \mathbb{R}_+ \to \mathbb{R}$ be functions.

- The notation $f(n) = O(g(n))$ means that there exist $\alpha \in \mathbb{R}_+$ and $N \in \mathbb{Z}_+$ such that $f(n) \leq \alpha g(n)$ for all $n \geq N$.
- The notation $f(n) = \Theta(g(n))$ means there exist $\alpha, \beta \in \mathbb{R}_+$ and $N \in \mathbb{Z}_+$ such that $\alpha g(n) \leq f(n) \leq \beta g(n)$ for all $n \geq N$.

We can make a remark about the factor complexities of morphic words. It is known that $P_w(n) = O(n^2)$ for all morphic words w. For pure morphic words, the following more precise result, known as *Pansiot's theorem*, is known.

Theorem 3.5.5. *If w is a pure morphic word, then one of the following holds:*

1. *$P_w(n) = \Theta(1)$.*
2. *$P_w(n) = \Theta(n)$.*
3. *$P_w(n) = \Theta(n \log \log n)$.*
4. *$P_w(n) = \Theta(n \log n)$.*
5. *$P_w(n) = \Theta(n^2)$.*

In addition to studying how many factors of a certain length an infinite word has, we can study how often the factors occur in the word. We conclude this section with some defitions related to this question. These definitions are studied a little bit in the exercises. They are not otherwise needed in the remaining part of these lecture notes.

An infinite word w is *recurrent* if every factor of w has infinitely many occurrences in w. An infinite word w is *uniformly recurrent* if for every factor x of w, there exists a number n such that x is a factor of every word in $\mathrm{Fact}_n(w)$.

Let x be a factor of an infinite word w. If the limit

$$\lim_{n\to\infty} \frac{|\,\mathrm{pref}_n(w)|_x}{n}$$

exists, it is called the *frequency* of x in w.

Exercises

3.5.1. Give an example of an infinite word w such that $P_w(n) = 5$ for all $n \geq 4$ but not for $n = 3$.

3.5.2. Find all right special factors of the Fibonacci word of length less than 7.

3.5.3. Find all right special factors of the word $\prod_{i=1}^{\infty} 0^i 1 = 01001000100001\cdots$ and estimate the growth rate of its factor complexity.

3.5.4. Show that for all infinite words w and for all $m, n \geq 1$,

$$P_w(m+n) \leq P_w(m)P_w(n).$$

3.5.5. Give an example of a k-ary infinite word w such that $P_w(n) = k^n$ for all $n \geq 1$.

*3.5.6.** Let w be a k-ary infinite word. Show that either $P_w(n) = k^n$ for all $n \geq 1$ or $P_w(n) = O(\alpha^n)$ for some $\alpha < k$.

3.5.7. Show that $P_T(n) = \Theta(n)$.

*3.5.8.** Try to find an interesting class of morphisms such that the following holds: If h is a morphism in this class and $w = h^\omega(0)$, then $P_w(n) = O(n)$.

3.5.9. Show that if every factor of an infinite word w has at least two occurrences in w, then w is recurrent.

3.5.10. Show that every periodic word is uniformly recurrent. Show that an ultimately periodic word that is not periodic is not recurrent.

3.5.11. Show that the Thue–Morse word is uniformly recurrent. Show that the Sierpinski word is recurrent but not uniformly recurrent.

3.5.12. Give an example of an infinite word w and its factor x such that the frequency of x in w does not exist.

3.5.13. Find the frequency of 0 in the Thue–Morse word.

*3.5.14.** Find the frequencies of 0 and 00 in the period-doubling word.

3.6 Sturmian words

An infinite word w is *Sturmian* if $P_w(n) = n+1$ for all n. Note that if w is Sturmian, then $P_w(1) = 2$, which means that w is binary. From now on, we concentrate on infinite words over the alphabet $\{0,1\}$.

It is not immediately clear from the definition whether there exist any Sturmian words. We see later that there are uncountably many Sturmian words in $\{0,1\}^\omega$, and the Fibonacci word is one of them.

An infinite word $w \in \{0,1\}^\omega$ is *balanced* if for all $n \geq 1$ and all $u, v \in \mathrm{Fact}_n(w)$,

$$||u|_1 - |v|_1| \leq 1.$$

Example 3.6.1. The set of factors of length six of the Fibonacci word is

$$\mathrm{Fact}_6(F) = \{001001, 001010, 010010, 010100, 100100, 100101, 101001\},$$

so $|u|_1 \in \{2,3\}$ for all $u \in \mathrm{Fact}_6(F)$. This means that F satisfies the balance condition at least for $n = 6$. We show later that F is balanced. The Thue–Morse word is not balanced, because it has factors 00 and 11.

Lemma 3.6.2. *An infinite word $w \in \{0,1\}^\omega$ is not balanced if and only if it has factors $0z0$ and $1z1$ for some word z.*

Proof. The "if"-direction is clear. To prove the "only if"-direction, let w be not balanced. Let u, v be the shortest factors of w such that $|u| = |v|$ and $|u|_1 - |v|_1 \geq 2$. Let $u = ax$ and $v = by$, where $a, b \in \{0,1\}$. If $a = b$, then $|x| = |y|$ and $|x|_1 - |y|_1 = |u|_1 - |v|_1 \geq 2$, which contradicts the minimality of u and v, so it must be $a \neq b$. Let z be the maximal common prefix of x and y. If $z = x = y$, then $||u|_1 - |v|_1| = 1$, which is a contradiction, so it must be $|z| < |u| = |v|$. Then $u = azcs$ and $v = bzdt$, where c and d are distinct letters. If $a = d$ and $b = c$, then $|s| = |t|$ and $|s|_1 - |t|_1 = |u|_1 - |v|_1 \geq 2$, which contradicts the minimality of u and v, so it must be $a = c$ and $b = d$. Then $0z0$ and $1z1$ are factors of w. □

Theorem 3.6.3. *If $w \in \{0,1\}^\omega$ is balanced and aperiodic, then it is Sturmian.*

Proof. Let w be aperiodic and not Sturmian. Then w has two right special factors u, v of the same length, which means that $u0, u1, v0, v1$ are factors of w. Let x be the longest common suffix of u and v. Because $u \neq v$, it must be $|x| < |u| = |v|$. This means that one of $0x$ and $1x$ is a suffix of u and the other is a suffix of v. Then w has the factors $0x0$ and $1x1$, so w is not balanced. This proves the theorem. □

In the next two theorems, we find examples of Sturmian words.

Theorem 3.6.4. *The Fibonacci word is Sturmian.*

Proof. The Fibonacci word F is aperiodic by Theorem 3.1.11, so by Theorem 3.6.3, it is sufficient to show that it is balanced. We assume that F is not balanced and derive a contradiction. Let $0z0$ and $1z1$ be factors of F and let $|z|$ be minimal. It is easy to see that $|z| \geq 2$. Because 11 is not a factor of F, $z = 0x0$, and $00x00$ and $010x01$ are factors of F. Because $F = \phi(F)$, we see that $00x0 = \phi(1)\phi(u)\phi(1)$ and $010x01 = \phi(0)\phi(u)\phi(0)$, where $1u1$ and $0u0$ are factors of F. This contradicts the minimality of z. □

An infinite word $w = a_1 a_2 a_3 \cdots$ is *mechanical* if there exist real numbers $\alpha, \beta \in [0, 1]$ such that

$$a_n = \lfloor \alpha(n+1) + \beta \rfloor - \lfloor \alpha n + \beta \rfloor$$

for all $n \geq 0$ or

$$a_n = \lceil \alpha(n+1) + \beta \rceil - \lceil \alpha n + \beta \rceil$$

for all $n \geq 0$.

Theorem 3.6.5. *If $w \in \{0, 1\}^\omega$ is mechanical and aperiodic, then it is Sturmian.*

Proof. Left as an exercise. □

It could be proved that also the other directions of Theorems 3.6.3 and 3.6.5 are true. In other words, the following are equivalent for an infinite word $w \in \{0, 1\}^\omega$:

1. w is Sturmian.
2. w is balanced and aperiodic.
3. w is mechanical and aperiodic.

Sturmian words have also many other equivalent definitions, and they come up in many different places.

Exercises

3.6.1. Give examples of balanced and nonbalanced periodic words.

3.6.2. Show that Sturmian words are recurrent.

*3.6.3.** Show that all suffixes of a Sturmian word are Sturmian. Show that for every Sturmian word w, there exists a letter a such that aw is Sturmian. Can you find some kind of a generalization of this for some larger class of infinite words?

*3.6.4.** Prove Theorem 3.6.5. Prove that a mechanical word is aperiodic if and only if the number α in the definition is irrational.

*3.6.5.** How is the picture on the cover related to this section?

Hints and answers to selected exercises

1.1.1. $3^n - 3 \cdot 2^n + 3$.

1.1.2. 12, 11.

1.1.4. Show that a k-ary word of length n can have at most $\sum_{i=0}^{n} \min\{k^i, n-i+1\}$ factors. Find a binary word of length 10 that matches this bound, either by hand, by a computer search, or by using De Bruijn words.

1.1.9. Use the function N_B of Example 1.1.11, or the alphabetical order (it is defined in Section 1.3, but you probably know how it works).

1.2.1. Notice that $|xy| = |yx|$.

1.2.2. Use Lemma 1.2.1.

1.2.3. Use Theorem 1.2.4.

1.2.4. Use Lemma 1.2.5.

1.2.5. Use Theorem 1.2.6 and Theorem 1.2.8.

1.2.6. Let $p_k(n)$ be the number of primitive words in Σ^n. Show that

$$k^n = \sum_{d|n} p_k(d)$$

and use the Möbius inversion formula.

1.3.2. $k^{n/2}$ if n is even and $k^{(n+1)/2}$ if n is odd.

1.3.3. Show that if w is a word and a is a letter, then wa can have at most one palindromic factor that w does not have. Conclude that the answer is $n+1$.

1.3.4. As an intermediate result, show that v and v^R are conjugates.

1.3.10. Show that $uuv \leq_{\text{lex}} uvu$ or $vuu \leq_{\text{lex}} uvu$.

1.4.4. Use Theorem 1.4.4 and the fact that if k is a period, then nk is a period for all $n \geq 1$. Alternatively, use Theorem 1.2.6.

1.4.6. Use Theorem 1.4.6.

2.1.2. $(X, Y, Z) \mapsto (pq, p, qpq)$, $(X, Y, Z) \mapsto (pq, pqp, q)$.

2.1.3. $(X, Y, Z) \mapsto (p^i, p^j, p^k)$, $(X, Y, Z) \mapsto (p, \varepsilon, q)$.

2.1.4. Try to fit $h(T)^3$ inside $h(Y)$.

2.1.6. At least length 14 is possible.

2.1.7. Make $h(Y)$ short, $h(X)$ long, and $h(Z)$ even longer.

2.1.8. Show first that if h is a solution, then $h(X) = h(YZ)$.

2.2.1. $(X, Y) \mapsto ((pa)^i p, (pa)^j p)$.

2.2.2. $(S, T, X, Y) \mapsto (p, p, q^i, q^j)$.

2.2.3. $(X, Y, Z) \mapsto ((pq)^i p, (pq)^{i+1} p, qp^2 q)$, $(X, Y, Z) \mapsto (p, p, \varepsilon)$.

2.2.4. $(X, Y, Z, T) \mapsto (a^i, a^j, a^k, a^l)$.

2.2.5. $(X) \mapsto (a)$, $(X) \mapsto (aaba)$.

2.2.6. $(X, Y, Z) \mapsto (a^i, a, a^{i+1})$.

2.2.7. $(X, Y, Z) \mapsto (((pq)^{i+1} p)^j pq, (pq)^i p, qp((pq)^{i+1} p)^k)$, $(X, Y, Z) \mapsto (p^i, p^j, p^k)$.

2.2.8. Start by writing the equation in the form $(U_0 b_1 U_1 \cdots b_m U_m, V_0 c_1 V_1 \cdots c_n V_n)$, where all $b_i, c_i \in \Sigma \smallsetminus \{a\}$ and all $U_i, V_i \in \{X, a\}^*$. Show that $m = n$ and $b_i = c_i$ for all i. Study the equations (U_i, V_i).

2.2.9. To prove the first claim, show first that $(UaU'UbU', VaV'VbV')$ is equivalent to the pair of equations $\{(UaU', VaV'), (UbU', VbV')\}$. Show that if h is a solution of this pair, then $h(U)$ and $h(V)$ cannot be of different length. Conclude that h is a solution of the original pair of equations $\{(U, V), (U', V')\}$.

2.3.2. $(ab)^* a$, $\{w \in \{a, b\}^* \mid 3 \leq |w| \leq 5\}$.

2.3.4. $b^* a^+ b^+ a^*$, no.

2.3.7. Show that it is a right unitary monoid.

2.3.10. The case of the submonoid $\{\varepsilon\}$ is trivial. Otherwise, let k be the smallest positive integer such that a^k is in the submonoid. Show that the minimal generating set cannot contain two words a^m and a^n where $m \equiv n \pmod{k}$.

2.4.2. The equations can be very simple.

2.4.3. Find a morphism that is a solution of the first equation but not of the second, and a morphism that is a solution of the second equation but not of the first.

2.4.4. $(X, Y, Z) \mapsto (\varepsilon, p, q)$, $(X, Y, Z) \mapsto (p^i, p^j, p^k)$.

3.1.2. Recall some result from Chapter 1.

3.1.8. Use the previous exercise.

3.1.9. Use the words G_n from a previous exercise.

3.2.1. Show that $h^n(0) \in (01)^*0$ for all n and thus $h^\omega(0) = (01)^\omega$.

3.2.2. In the case of the period-doubling word, show that $h^{2n}(0) = P_n0$ and $h^{2n}(1) = P_n1$ for all n. In the case of the Sierpinksi word, show that $h^n(0) = S_n$ for all n.

3.2.5. $u \in 0^*$, $u \in 1^*$.

3.2.6. First, look at the fixed point of the morphism defined by $c \mapsto cab$, $a \mapsto a$, $b \mapsto ab$.

3.3.1. Similar to the proof of Theorem 3.3.3.

3.3.2. Show that if u^2 is a nonempty prefix of T, then $|u|$ cannot be odd, and if $|u|$ is even, then T has a shorter nonempty square prefix. Alternatively, show that if u^2 is a nonempty prefix of T, then $0u^2$ and $1u^2$ are factors of T, and one of them is an overlap.

3.3.3. $abaabbaaabbb\cdots$.

3.3.6. Show first that $F_{n+1}^3 \operatorname{pref}_{|F_n|-2}(F_n)$ is a factor of the Fibonacci word for all $n \geq 1$.

3.3.7. 5324, 2388.

3.4.2. You can use the function $f : \{0,1,2\}^* \to \mathbb{Z},\ f(u) = |u|_0 + 2|u|_1 + 3|u|_2$.

3.4.3. Look at the Parikh vectors of prefixes modulo n.

3.5.3. 0^n and 0^n10^{n+k} where $n \geq 0$ and $k \geq 1$, $\Theta(n^2)$.

3.5.13. 1/2.

3.5.14. 2/3, 1/3.

3.6.2. Show first that if some factor of a Sturmian word w has only one occurrence in w, then some suffix of w has less factors than w.

Bibliography

[1] Jean-Paul Allouche and Jeffrey Shallit. *Automatic sequences. Theory, applications, generalizations.* Cambridge University Press, 2003. `doi:10.1017/CBO9780511546563`.

[2] Jean Berstel, Dominique Perrin, and Christophe Reutenauer. *Codes and Automata.* Cambridge University Press, 2010.

[3] Christian Choffrut and Juhani Karhumäki. Combinatorics of words. In Grzegorz Rozenberg and Arto Salomaa, editors, *Handbook of Formal Languages*, volume 1, pages 329–438. Springer, 1997. `doi:10.1007/978-3-642-59136-5_6`.

[4] M. Lothaire. *Combinatorics on Words.* Addison-Wesley, 1983.

[5] M. Lothaire. *Algebraic Combinatorics on Words.* Cambridge University Press, 2002. URL: `http://www-igm.univ-mlv.fr/~berstel/Lothaire/AlgCWContents.html`.

[6] M. Lothaire. *Applied Combinatorics on Words.* Cambridge University Press, 2005. URL: `http://www-igm.univ-mlv.fr/~berstel/Lothaire/AppliedCW/AppCWContents.html`.

[7] Michel Rigo. *Formal languages, automata and numeration systems 1.* ISTE; John Wiley & Sons, Inc., 2014.