Chapter 2
Words that Can Be Avoided

We start with an old problem from number theory whose solution leads to the first example of an infinite cube-free word, thus showing that the word x^3 is avoidable. Then we shall describe all avoidable words.

Theorems proved in this Chapter include

- Theorem 2.1.2 of Thue, Hedlund, Morse and Arshon about cube-free words.
- Theorem 2.3.4 of Thue characterizing square-free substitutions.
- Theorem 2.5.14 of Bean, Ehrenfeucht, McNulty and Zimin characterizing avoidable words.

2.1 An Old Example

In 1851 Eugène Prouhet [271] (see also [5]) studied the following number theory problem, which on the first glance has nothing to do with combinatorial algebra.

Question. Are there arbitrary big numbers m such that some intervals $[1, M]$, M depends on m, of natural numbers can be divided into 2 disjoint parts P_1, P_2 such that the sum of all elements in P_1 is the same as in P_2, the sum of all squares of elements P_1 is the same as in $P_2, \ldots$, the sum of m-th powers of elements of P_1 is the same as in P_2.

This problem has long history. In particular, Gauss and Euler studied some variations of this problem.

Prouhet came up with a solution. His solution may be interpreted in the following way. Let us consider an example. Take $m = 2$. Consider the word $p_2 \equiv abbabaab$, and produce the following table:

$$\begin{array}{cccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ a & b & b & a & b & a & a & b \end{array}$$

M.V. Sapir, *Combinatorial Algebra: Syntax and Semantics*,
Springer Monographs in Mathematics, DOI 10.1007/978-3-319-08031-4_2

Now let P_a be the set of numbers from 1 to 8, that are above a in this table, P_b be the set of numbers, that are above b: $P_a = \{1,4,6,7\}$, $P_b = \{2,3,5,8\}$ (recall the connection between words and partitions from Example 1.2.2). Let us check the Prouhet condition:

$$1+4+6+7 = 18 = 2+3+5+8,$$

$$1^2+4^2+6^2+7^2 = 102 = 2^2+3^2+5^2+8^2.$$

If we want to construct the Prouhet decomposition for $m = 3$ we have to take this word $p_2 \equiv abbabaab$, change a by b and b by a (we'll get $baababba$), and concatenate these two words:

$$p_3 \equiv abbabaabbaababba.$$

By induction one can easily define the Prouhet word for every m: p_m is obtained by concatenating p_{m-1} and p'_{m-1} where p'_{m-1} is obtained from p_{m-1} by the substitution $a \mapsto b, b \mapsto a$.

Exercise 2.1.1. Prove that for every m the partition corresponding to the Prouhet word p_m satisfies the Prouhet condition for the sums of nth powers for every $n \le m$.

The word p_m was rediscovered several times after Prouhet. Axel Thue rediscovered this word in 1906 [314] and he was the first to prove the following result:

Theorem 2.1.2. *The word p_m does not contain subwords of the form www where w is any nonempty word. Thus words p_m are* cube-free*.*

Arshon, Hedlund and Morse proved the same result in the late 1930s [14, 238]. Now it belongs to some collections of problems for high school students.

2.2 Proof of Thue's Theorem

Let us consider the following substitution:

$$\phi(a) \equiv ab, \ \phi(b) \equiv ba.$$

Words ab and ba will be called *blocks*.

Let $t_1 = a, \ \ldots, \ t_n = \phi(t_{n-1})$.

Exercise 2.2.1. Prove that $t_n = p_{n-1}$ for every $n \ge 2$.

Lemma 2.2.2. *If w is cube-free, then $\phi(w)$ is also cube-free.*

Proof. Suppose that $\phi(w)$ contains a cube ppp.

Case 1. The length of p is even.

Case 1.1. The first occurrence of p starts with the first letter of a block. Since $\phi(w)$ is a product of blocks of length 2, and $|p|$ is even, p ends with the second letter of a block. Then the second and the third occurrence of p also start with the first letter of a block and end with the second letter of a block. Thus p is a product of blocks, so $p = \phi(q)$ for some word q. Now let us substitute every block in $\phi(w)$ by the corresponding letter. Then $\phi(w)$ will turn into w and $ppp = \phi(q)\phi(q)\phi(q)$ will turn into qqq. Therefore w contains a cube qqq.

Case 1.2. The first occurrence of p starts with the second letter of a block. Then it ends with the first letter of a block, and the same is true for the second and the third occurrence of p. Without loss of generality assume that p starts with a. This a is second letter of the block ba. Since the second occurrence of p also starts with a and this a is the second letter of a block, we can conclude that p ends with b. Therefore $p \equiv ap'b$. Then we have:

$$\phi(w) \equiv \ldots b\ ap'b\ ap'b\ ap'b\ \ldots.$$

Consider the word bap'. This word has even length, starts with the beginning of a block and repeats 3 times in $\phi(w)$, which is impossible by the previous case.

Case 2. The word p has odd length. If the first occurrence of p starts with the first letter of a block, then the second occurrence of p starts with the second letter of a block. If the first occurrence of p starts with the second letter of a block, then the second occurrence of p starts with the first letter of a block and the third occurrence of p starts with the second letter of a block.

In any case there are two consecutive occurrences of p such that the first one starts with the first letter of a block and the second one starts with the second letter of a block. Let us denote these copies of p by p' and p''.

It is clear that $|p| \geq 2$.

Suppose that p' starts with ab. Then p'' also starts with ab. This b is the first letter of a block. Therefore the third letter in p'' is a. Therefore the third letter in p' is also a. This a is the first letter of a block. The second letter of this block is b. Therefore the fourth letter of p' is b. Then the fourth letter of p'' is b, which is the first letter of a block, the fifth letter of p'' is a, same as the fifth letter in p', and so on. Every odd letter in p is a, every even letter in p is b. Since p' has odd number of letters, the last letter in p' is a. This a is the first letter of a block. The second letter of this block is the first letter of p'', which is a – a contradiction (we found a block aa). □

Exercise 2.2.3. Prove that t_n does not contain subwords of the form $qpqpq$ for any words p and q.

2.3 Square-Free Words

The words constructed in the previous section do not contain cubes. Now we will consider words that do not contain squares, the *square-free* words.

A word is called *square-free* if it does not contain a subword of the form uu.

Exercise 2.3.1. Prove that every square-free word over an alphabet with 2 letters has length at most 3.

Theorem 2.3.2 (Thue, [314]). *There exist arbitrary long square-free words over a 3-letter alphabet.*

Consider the following substitution:

$$\phi(a) \equiv abcab,\ \phi(b) \equiv acabcb,\ \phi(c) \equiv acbcacb.$$

Theorem 2.3.2 easily follows from the following

Lemma 2.3.3. *For every square-free word w, $\phi(w)$ is square-free.*

A substitution satisfying the condition of Lemma 2.3.3 will be called *square-free.*

In turn, Lemma 2.3.3 will be a corollary of the following powerful theorem of Thue.

Theorem 2.3.4. *Let M and N be alphabets and let ϕ be a substitution from M to N^+. If*

(1) *$\phi(w)$ is square-free whenever w is a square-free word from M^+ of length no greater than 3,*
(2) *$a = b$ whenever $a, b \in M$ and $\phi(a)$ is a subword of $\phi(b)$,*

then ϕ is a square-free substitution.

Proof of Theorem 2.3.4. Let ϕ satisfy (1) and (2). First of all let us prove the following "rigidity" statement:

Claim. If $a, e_1, \dots, e_n$ are letters from M, $E \equiv e_1 \dots e_n$ is a square-free word, and $\phi(E) \equiv X\phi(a)Y$ then $a = e_j$, $X = \phi(e_1 \dots e_{j-1})$, $Y = \phi(e_{j+1} \dots e_n)$ for some j.

Suppose this is not true. Since $\phi(e_i)$ cannot be a subword of $\phi(a)$ (by (2)), $\phi(a)$ intersects with at most 2 factors $\phi(e_i)$. Since $\phi(a)$ cannot be a subword of $\phi(e_i)$ (again by (2)), $\phi(a)$ intersects with exactly 2 factors, say $\phi(e_j e_{j+1})$. Then $\phi(e_j) \equiv pq$, $\phi(e_{j+1}) \equiv rs$, and $\phi(a) \equiv qr$. Now $\phi(ae_ja) \equiv qrpqqr$ is not square-free. By condition (1) the word ae_ja is not square-free, thus $a = e_j$. On the other hand: $\phi(ae_{j+1}a) \equiv qrrsqr$ also is not square-free. Thus $a = e_{j+1}$. Therefore $e_j = e_{j+1}$, which contradicts the fact that $e_1 \dots e_n$ is a square-free word.

Now suppose that w is a square-free word from M^+ and $\phi(w) \equiv xyyz$ for some nonempty word y. Let $w \equiv e_0 \dots e_n$. Let us denote $\phi(e_i)$ by E_i. We have

$$E_0 E_1 \dots E_n \equiv xyyz.$$

If E_0 is contained in x or E_n is contained in z then we can shorten w (delete e_0 or e_n). Therefore we can suppose that

$$E_0 \equiv xE_0',\ E_n = E_n'z,\ yy \equiv E_0' E_1 \dots E_{n-1} E_n'.$$

By condition (0) we have that $n \geq 3$.

The word y is equal to $E_0' E_1 \dots E_j' \equiv E_j'' E_{j+1} \dots E_n'$ with $E_j' E_j'' \equiv E_j$. If $j = 0$ then $E_1 E_2$ must be a subword of E_0, which is impossible. Similarly, $j \neq n$.

Now by the rigidity statement,

$$E_0' \equiv E_j'',\ E_1 \equiv E_{j+1},\ \dots, E_j' \equiv E_n',$$

and, in particular, $n = 2j$. Therefore

$$\begin{aligned}&\phi(e_0 e_j e_n) \equiv E_0 E_j E_n \equiv xE_0' E_j' E_j'' E_n' z \equiv \\ &xE_0' E_j' E_0' E_j' z.\end{aligned}$$

By condition (1) either $e_0 = e_j$ or $e_j = e_n$. Without loss of generality let $e_0 = e_j$. We also know that $E_1 \equiv E_{j+1}, \dots, E_{j-1} \equiv E_{2j-1}$. Condition (1) implies that ϕ is one-to-one. Therefore $e_0 = e_j, e_1 = e_{j+1}, \dots$. Hence w is not square-free: it is equal to $e_0 e_1 \dots e_{j-1} e_0 \dots e_{j-1} e_n$. □

Theorem 2.3.4 implies Lemma 2.3.3 (check it!) and Theorem 2.3.2.

A complete algorithmic characterization of square-free substitutions was found first by Berstel [45, 46]. The best (in the computational sense) characterization was found by Crochemore.

Theorem 2.3.5 (Crochemore [75])**.** *Let ϕ be a substitution, M be the maximal size of a block, m be the minimal size of a block, k be the maximum of* 3 *and the number* $1 + [(M-3)/m]$. *Then ϕ is square-free if and only if for every square-free word w of length $\leq k$ $\phi(w)$ is square-free.*

Using Theorem 2.3.4 and a computer, one can establish a substitution from an infinite alphabet to $\{a, b, c\}^+$ that is square-free. We will present here a substitution from $\{x_1, x_2 \dots\}$ to $\{a, b, c, d, e\}^+$. Let $w_1, w_2, \dots$ be an infinite sequence of distinct square-free words on a three-letter alphabet $\{a, b, c\}$. Consider the following substitution from $x_1, x_2, \dots$ to $\{a, b, c, d, e\}^+$:

$$x_i \to dw_i e w_i.$$

This substitution is square-free by Theorem 2.3.4. Indeed it is clear that a word $dw_i e w_i$ cannot be a subword of $dw_j e w_j$ because w_i and w_j do not contain d and e. Now if $dw_iew_idw_jew_jdw_kew_k$ contains a square uu then the numbers of d's and e's in uu must be even. Neither of them may be 0, because otherwise one of the words w_i, w_j, w_k would contain a square. So each of them is 2. Therefore each of the copies of u has one d and one e. The first d cannot participate, otherwise u must be equal to dw_iew_i, and $w_i \neq w_j$. Therefore u must start at the middle of the first w_i, and end in the middle of the first w_j. Then u must contain the subword ew_id. But this subword occurs in our product only once, so the second copy of u cannot contain it, a contradiction.

A square-free substitution from $\{a_1, a_2, a_3 \ldots\}$ to $\{a, b, c\}^+$ has been found by Bean, Ehrenfeucht and McNulty [28].

2.4 *k*th Power-Free Substitutions

For every natural $k \geq 2$, a substitution ϕ is called *kth power free* if the word $\phi(w)$ is k-power free whenever w is.

The following theorem, also proved by Bean–Ehrenfeucht–McNulty, gives a sufficient condition for a substitution to be kth power-free for $k > 2$.

Theorem 2.4.1. *Let M and N be alphabets and let ϕ be a substitution $M \to N^+$ that satisfies the following three conditions*

(1) *$\phi(w)$ is kth power-free whenever w is a k-power free word of length no greater than $k + 1$.*
(2) *$a = b$ whenever $\phi(a)$ is a subword of $\phi(b)$.*
(3) *If $a, b, c \in M$ and $x\phi(a)y \equiv \phi(b)\phi(c)$ then either x is empty and $a = b$ or y is empty and $a = c$.*

Then ϕ is kth power-free.

Exercise 2.4.2. Prove Theorem 2.4.1.

In particular, Theorem 2.4.1 implies existence of cube-free substitutions from the infinite set $\{a_1, a_2, \ldots,\}$ to $\{a, b\}^+$. An explicit substitution has been constructed in [28].

2.5 Avoidable Words

2.5.1 Examples and Simple Facts

We say that a word u *avoids* a word v if u does not contain the value of v under any substitution replacing letters by nonempty words (in that case we also call the word u *v-free*). A word v is called *k-avoidable* if there exists an infinite word in a k-letter alphabet avoiding v (equivalently, if there are infinitely many (finite) words in a k-letter alphabet avoiding v). If v is k-avoidable for some k, then v is called *avoidable*. We already know that x^k, $k \geq 2$, are avoidable words. It is clear that x^2yx and $xyxy$ are also avoidable words because, in general,

Exercise 2.5.1. Prove that if u is a k-avoidable word and ϕ is any substitution then any word containing $\phi(u)$ is also k-avoidable.

The following exercise is also useful.

Exercise 2.5.2. Prove that for every words u and w, if w avoids u, then every subword of w avoids u.

On the other hand, the word xyx is *unavoidable*. Indeed, let k be any natural number. Suppose that there exists an infinite set of words W in a k-letter alphabet that avoid xyx. Then W must contain a word w of length $> 2k$. Since we have only k letters in the alphabet, one of these letters, say a, occurs twice in w, that is w contains a subword apa for some nonempty word p. This word is of the form xyx ($\phi : \; x \mapsto a, \; y \mapsto p$).

Let us consider the following words Z_n, which will be called *Zimin words*:

$$Z_1 \equiv x_1, Z_2 \equiv x_1 x_2 x_1, \ldots, Z_n \equiv Z_{n-1} x_n Z_{n-1}. \tag{2.5.1}$$

Notice that these words were studied before Zimin. They appear in [28] as "maximal" unavoidable words. Before that they were considered by Coudrain and Schützenberger [74], and even before that they were considered by Levitzki [201] also in connection to Burnside-type problems (for rings).

The next exercise shows that prehistoric humans discovered Zimin words as soon as they learned how to count and divide by 2.

Exercise 2.5.3. Count the numbers i from 1 to $2^n - 1$ and write down the maximal exponent of 2 dividing i. The word we obtain is $01020103\ldots$. Show that this word is the value of Z_n under the substitution $x_j \mapsto j - 1$.

Zimin was first to understand the real role of Zimin words in the theory of avoidable words and Burnside problems for semigroups.

Exercise 2.5.4 (Bean–Ehrenfeucht–McNulty [28], Zimin [337, 338]). Z_n is an unavoidable word (for every n).

We shall show that the Zimin word Z_n is a "universal" unavoidable word in an n-letter alphabet.

Let us list some properties of these words.

Exercise 2.5.5. (1) $|Z_n| = 2^n - 1$,
(2) Every odd letter in Z_n is x_1, every even letter in Z_n is x_i for $i > 1$,
(3) For every $k \leq n$, $Z_n \equiv Z_k(Z_{n-k+1}, x_{n-k+2}, \ldots, x_n)$,
(4) For every k, n, $Z_n(Z_{k+1}, x_{k+2}, \ldots, x_{k+n}) \equiv Z_{n+k}(x_1, \ldots, x_{n+k})$,
(5) For every k, n, $Z_k \cdot Z_n(x_{k+1} Z_k, \ldots, x_{k+n} Z_k) \equiv Z_{n+k}(x_1, \ldots, x_{n+k})$,
(6) If we delete the letter x_1 from Z_n then we get a word that differs from Z_{n-1} only by names of the letters.

(7) Every word in the n-letter alphabet $\{x_1, \dots, x_n\}$, which properly contains Z_n, contains a square, and so is avoidable.

2.5.2 Fusions, Free Sets and Free Deletions

Definition 2.5.6. Let u be a word in an alphabet X, let B and C be subsets of X. We call the pair (B, C) a *fusion* in u if for every two-letter subword xy in u

$$x \in B \text{ if and only if } y \in C.$$

We will call B and C *components* of the fusion.

For example the sets $\{x\}$ and $\{y\}$ form a fusion in the word $yxyzxytx$ and in the word $xyxy$. Sets $\{x, z\}$ and $\{z, t\}$ form a fusion in the word $xtxzztxz$.

Definition 2.5.7. If B, C is a fusion in u, then any subset $A \subseteq B \backslash C$ is called a *free set* of u.

For example, $\{x\}$ is a free set in $xtxzztxz$.

Exercise 2.5.8. Find all fusions and all free sets in Z_n.

Let u be a word, Y be a subset of letters of u. Consider the word obtained from u by deleting all letters that belong to Y. This word will be denoted by u_Y.

Definition 2.5.9. A deletion of a free set A in a word is called a *free deletion* and is denoted by σ_A.

Definition 2.5.10. A sequence of deletions $\sigma_Y, \sigma_Z, \dots$ in a word u is called a *sequence of free deletions* if Y is a free set in u, Z is a free set in $u_Y \equiv \sigma_Y(u)$, etc.

Let us fix some notation. Let $u = x_1 x_2 \dots x_n$ be a word, a be a letter. For $\alpha, \beta \in \{0, 1\}$ we denote by $[u, a]^\alpha_\beta$ the word u with letter a inserted between every two consecutive letters and possibly at the beginning and at the end. More precisely

$$[u, a]^\alpha_\beta = a^\alpha x_1 a x_2 a \dots a x_n a^\beta \tag{2.5.2}$$

Definition 2.5.11. Let u be a word, B, C be a fusion in u, $A \subseteq B \backslash C$ and a be a letter not in $\mathrm{cont}(u)$. For every substitution ϕ of the word $\sigma_A(u)$ we can define a substitution ϕ^* of u by the following rules:

$$\phi^*(x) = \begin{cases} a & \text{if } x \in A \\ [\phi(x), a]^0_0 & \text{if } x \in C \backslash B \\ [\phi(x), a]^0_1 & \text{if } x \in B \cap C \\ [\phi(x), a]^1_1 & \text{if } x \in B \backslash (C \cup A) \\ [\phi(x), a]^1_0 & \text{otherwise.} \end{cases} \tag{2.5.3}$$

We will call ϕ^* the substitution induced by ϕ relative to the triple (A, B, C).

For example, if $u \equiv xyzt$, $B = \{x, z\}$, $C = \{y, t\}$, $A = \{x\}$, and $\phi(x) \equiv x, \phi(y) \equiv y, \phi(z) \equiv z, \phi(t) \equiv t$, then $\phi^*(x) \equiv a$, $\phi(y) \equiv y$, $\phi(z) \equiv aza$, $\phi(t) \equiv t$. Therefore $\phi^*(u) \equiv ayazat$.

The following lemma contains the main property of free deletions.

Lemma 2.5.12. *For every word u, $\phi^*(u) = [\phi(\sigma_A(u)), a]^\alpha_\beta$ for some α and β. If u starts and ends with a letter from $B \backslash C$ then $\alpha = \beta = 1$.*

Exercise 2.5.13. Prove Lemma 2.5.12.

2.5.3 The Bean–Ehrenfeucht–McNulty and Zimin Theorem

2.5.3.1 The Formulation

Theorem 2.5.14 (Bean–Ehrenfeucht–McNulty, Zimin)**.** *The following conditions are equivalent for every word u:*

(1) *u is unavoidable.*
(2) *Z_n contains a value of u, where n is the number of distinct letters in u.*
(3) *There exists a sequence of free deletions that reduces u to a 1-letter word.*

Exercise 2.5.15. Deduce the following two facts from Theorem 2.5.14

(1) Every word of length $\geq 2^n$ in a n-letter alphabet is avoidable.
(2) Every unavoidable word u contains a *linear letter*, i.e., a letter that occurs in u only once.

Exercise 2.5.16. Z_n has a sequence of free deletions of length $n-1$ which reduces Z_n to x_n. **Hint:** The set $\{x_1\}$ is a free set in Z_n. Deleting x_1, we obtain a word that differs from Z_{n-1} only by names of letters (part (6) of Exercise 2.5.5).

Our proof of Theorem 2.5.14 is simpler than the original proofs of Bean–Ehrenfeucht–McNulty and Zimin. We shall follow [277]. It is also more general: the same ideas will be used later in a more complicated situation.

2.5.3.2 Proof of (3) → (2)

Exercise 2.5.17. Prove this implication. **Hint:** Suppose that a one letter word p is obtained from u by a sequence of free deletions $\sigma_1, \ldots, \sigma_k$: $u \to u_1 \to \ldots u_k \equiv p$. Note that $k \leq n-1$. Let ϕ be the identity substitution $p \mapsto p$. Then apply induced substitutions (2.5.3) and Lemma 2.5.12 k times using letters $x_k, x_{k-1}, \ldots, x_1$ and notice that $\phi^*(u_{k-1})$ is a subword of $x_k p x_k$, $\phi^{**}(u_{k-2})$ is a subword of $x_{k-1}x_k x_{k-1} p x_{k-1} x_k x_{k-1}$, etc.

2.5.3.3 Proof of (2) → (1)

This implication immediately follows from Exercises 2.5.1 and 2.5.2.

2.5.3.4 Proof of (1) → (3)

Suppose that there is no sequence of free deletions that reduces u to a 1-letter word. We shall prove that u is avoidable.

In fact we will construct a substitution γ such that for some letter a words $\gamma(a)$, $\gamma^2(a)$, ... are all different and avoid u.

The substitution γ is constructed as follows. Let r be a natural number. Let $\mathcal{A}$ denote the alphabet of r^2 letters $a_{ij},\ 1 \le i, j \le r$. Consider the $r^2 \times r$-matrix M, in which every odd column is equal to

$$(1, 1, \dots, 1, 2, 2, \dots, 2, \dots, r, r, \dots, r),$$

and each even column is equal to

$$(1, 2, \dots, r, 1, 2, \dots, r, \dots, 1, 2, \dots, r).$$

Replace every number i in the j-th column by the letter a_{ij}. The resulting matrix is denoted by M'. The rows of M' can be considered as words in the alphabet $\mathcal{A}$. Denote these words by $w_1, \dots, w_{r^2}$ counting from top to bottom. Now define the substitution γ by the following rule:

$$\gamma(a_{ij}) = w_{(j-1)r+i}.$$

For every $j = 1, 2, \dots$ let $\mathcal{A}_j$ be the set $\{a_{ij} \mid i = 1, 2, \dots, r\}$.

Exercise 2.5.18. γ satisfies the following properties:

(1) The length of each block is r.
(2) No two different blocks have common 2-letter subwords.
(3) All letters in each block are different.
(4) The j-th letter in each block belongs to $\mathcal{A}_j$ (that is its second index is j).

Theorem 2.5.19. *Suppose that the word u cannot be reduced to a 1-letter word by a sequence of free deletions. Let $r > 6n + 1$ (where n is the number of letters in u). Then $\gamma^m(a_{11})$ avoids u for every m.*

Proof. By contradiction, suppose there exists m such that $\gamma^m(a_{11})$ contains $\delta(u)$ for some substitution δ. We may suppose that m is minimal possible. Clearly $m > 0$.

The idea of getting a contradiction is the following. First of all we will modify δ and find a substitution ϵ such that $\epsilon(u)$ has "long" sequences of free deletions. Then we prove that if a value of a word has "long" sequences of free deletions, then so does the word itself.

Let us call images of letters under γ *blocks*, and let us call products of blocks *integral words*.

Let w be an integral word. Let v be a subword of w. Then v is equal to a product of three words $p_1p_2p_3$ where p_1 is a suffix of a block, p_2 is a product of blocks or lies strictly inside a block, p_3 is a prefix of a block. Some of these words may be empty. The following property of integral words is important. It easily follows from Exercise 2.5.18.

Lemma 2.5.20. *The decomposition $v = p_1p_2p_3$ does not depend on the particular occurrence of v in an integral word: if we consider another occurrence of v, the decomposition $v = p_1p_2p_3$ will be the same.*

Now let $w = \gamma^m(a_{11})$. We have that $\delta(u)$ is a subword of w. For every letter $x \in \mathrm{cont}(u)$ let $\delta(x) = p_{x1}p_{x2}p_{x3}$ be the decomposition described above.

Now let us take any block B in w. This block can appear in many different places of w. Let us consider all words p_{xi} that are contained in all possible occurrences of B. There are no more than $3n$ such words p_{xi}. Each of them may occur in B at most once because all letters in B are different. Therefore we may consider B as an interval, which contains at most $3n$ other intervals. The length of B is at least $6n + 2$. Therefore there exists a 2-letter subword t_B of B that satisfies the following condition:

(T) For every subword p_{xi} inside B, the word t_B either is contained in p_{xi} or does not have common letters with it.

This easily follows from the trivial 1-dimensional geometry fact that k intervals on a line divide this line into at most $2k + 1$ subintervals two of which are infinite (why?).

Now let us replace the subword t_B in every block B in w by a new letter y_B. We shall get a new word w_1 in the alphabet $\mathcal{A} \cup \{y_B \mid B = \gamma(a_{ij})\}$. This word has the following form:

$$P_0Q_1y_1P_1Q_2y_2P_2\dots Q_ky_kP_kQ_{k+1},$$

where Q_i is a prefix of a block, P_i is a suffix of a block, P_i's and Q_i's do not overlap, and

$$\text{if } y_i = y_j \text{ then } Q_i = Q_j \text{ and } P_i = P_j. \tag{2.5.4}$$

Every word with these properties will be called a *quasi-integral* word.

Let us consider the substitution ϵ that is obtained from δ by the following procedure: take each $\delta(x)$ and replace each occurrence of t_B there by y_B. The property (T) of words t_B implies that occurrences of words t_B in $\delta(x)$ do not intersect, so our definition of ϵ is consistent. This property also implies that $\epsilon(u)$ is a subword of the quasi-integral word w_1.

Now we shall show that w_1 (and any other quasi-integral word) has "long" sequences of free deletions.

Exercise 2.5.21. In any quasi-integral word the sequence of deletions $\sigma_{\mathcal{A}_1}$, $\sigma_{\mathcal{A}_2}\dots$, $\sigma_{\mathcal{A}_r}$ is a sequence of free deletions.

The result of this sequence of deletions is, of course, the deletion of all letters from $\mathcal{A}$. Now we have to understand how free deletions in $\epsilon(u)$ relate to the deletions in u.

First of all we have the following two simple observations.

Let θ be any substitution and let v be any word. Let D be a set of letters of $\theta(v)$. Let D' be a subset of $\mathrm{cont}(v)$ defined as follows:

$$D' = \{x \mid \mathrm{cont}(\theta(x)) \subseteq D\}.$$

Now let us define a substitution θ_D of $v_{D'}$:

$$\theta_D(x) = \theta(x)_D.$$

Exercise 2.5.22. $\theta_D(v_{D'}) = \theta(v)_D$.

The following exercise, while simple, is a yet another key property of free sets.

Exercise 2.5.23. If D is a free set in $\theta(v)$ then D' is a free set in v.

Let us apply Exercises 2.5.21, 2.5.22 and 2.5.23 to our situation: $\epsilon(u) \leq w_1$. We deduce that there exists a sequence of free deletions $\sigma_1 \ldots, \sigma_r$ in u. Let $u_{\mathcal{A}'}$ be the result of these deletions. Then by Exercise 2.5.22

$$\epsilon_{\mathcal{A}}(u_{\mathcal{A}'}) \leq (w_1)_{\mathcal{A}} = y_1 y_2 \ldots y_k.$$

Now, by definition, each y_B determines the block B. Let us consider the substitution α that takes each y_B to B. Then $\alpha((w_1)_{\mathcal{A}}) = w$. Therefore $\alpha\epsilon_{\mathcal{A}}(u_{\mathcal{A}'})$ is a subword of $w = \gamma^m(a_{11})$. But the image of each letter from $u_{\mathcal{A}'}$ under $\alpha\epsilon_{\mathcal{A}}$ is a product of blocks. Therefore we can apply γ^{-1} to $\alpha\epsilon_{\mathcal{A}}(u_{\mathcal{A}'})$. The result, $\gamma^{-1}\alpha\epsilon_{\mathcal{A}}(u_{\mathcal{A}'})$, is a subword of $\gamma^{m-1}(a_{11})$.

Now we can complete the proof. The word $u_{\mathcal{A}'}$ contains at most the same number of letters as u, and $m-1$ is strictly less than m. By the assumption (we assumed that m is minimal) there exists a sequence of free deletions, which reduces $u_{\mathcal{A}'}$ to a 1-letter word. If we combine this sequence of free deletions with the sequence that we used to get $u_{\mathcal{A}'}$ from u, we will get a sequence of free deletions that reduces u to a 1-letter word, which is impossible.

The theorem is proved. □

2.5.4 Simultaneous Avoidability

Definition 2.5.24. A set of words W is said to be (k-)avoidable if there exists an infinite set of words in a finite alphabet (a k-letter alphabet), each of which avoids each word from W.

Exercise 2.5.25. Prove that a finite system of words W is avoidable if and only if each word in W is avoidable. **Hint:** Use Theorem 2.5.19. Alternatively, if X_w is the finite alphabet of an infinite word avoiding $w \in W$, find an infinite word in the alphabet $\prod_{w \in W} X_w$ (which is the Cartesian product of all X_w) avoiding all words from W simultaneously.

2.6 Further Reading and Open Problems

More on avoidable words, Prouhet–Morse–Thue words and related topics, applications to dynamics and arithmetics can be found in the books by Lothaire [206, 207] and N. Pytheas Fogg [103].

2.6.1 Square- and Cube-Free Words

The subject of k-power free words and certain nice generalizations of this concept is fast growing. Some of the recent papers are surveyed by Shur in [304].

One of the interesting problems was: how to decide, given a substitution $\phi: M \to M^+$ and $a \in M$ whether $\phi^n(a)$ is k-th power free for all n. Some sufficient conditions are in Thue's Theorem 2.3.4. Berstel [46] showed that over an alphabet with three letters, there is an algorithm for $k = 2$. Then Karhumäki [171] showed that over a binary alphabet, there is an algorithm for $k = 3$. The problem was solved in general by Mignosi and Séébold [231], who showed that there exists an algorithm for this problem for all alphabet sizes and all k.

There are several interesting generalizations of k-power free word studied, in particular, by Capri and Currie (see [63] and a survey [77]). For example, we say that a word u *does not contain Abelian k-th powers* if it does not contain products $w_1 w_2 \dots w_k$ where all subwords w_i are permutations of each other. One of the typical results about such words is the following. Let $a(k)$ be the smallest alphabet size, on which Abelian k-powers are avoidable. Then $a(4) = 2, a(3) = 3, a(2) = 4$ (see [77]).

2.6.2 Avoidable Words

Open problems about avoidable words are discussed in McNulty's talks [227]. Since the slides of [227] are not generally available, we present a few open problems here. Some of these (and other) problems can be found in the survey by Currie [77] as well.

Problem 2.6.1. *For each natural number n find out how many unavoidable words there are in the n letter alphabet.*

If a word u is avoidable, then there is the smallest natural number $\mu(u)$ such that there is an infinite word on $\mu(u)$ letters, which avoids u. We can extend this function μ to all words by putting $\mu(u) = 1$ when u is unavoidable.

Problem 2.6.2. *Let u be an avoidable word with $\mu(u) = m$. Can every u-free word on m letters be extended to a maximal u-free word on m letters?*

The answer is yes if $m = 1$. The answer is not known even for $m = 2$.

Problem 2.6.3. *Is μ a* recursive function*? That is, is there an algorithm that computes μ?*

The proof of Theorem 2.5.14 in Chapter 2 gives an upper bound on μ as a quadratic polynomial in the number $c = c(u)$ of different letters in u. A linear upper bound was found by I. Mel'nichuk [229]: $\mu(u) \leq 3c/2 + 3$. She also announced (in 1988) that $\mu(u) \leq c + 6$ (unpublished). On the other hand we still do not know the answer to the following problem, which is open since [28, 337]

Problem 2.6.4. *Is $\mu(u)$ bounded from above by a constant?*

It is not even known if that constant is not equal to 5. By Exercise 2.3.1 and Theorem 2.3.4, $\mu(x^2) = 3$. In [17], Baker, McNulty and Taylot found the first word $u = abx_1bcx_2cax_3bax_4ac$, for which $\mu(u) = 4$. Clark [67] constructed a somewhat similar looking word $u = abx_1bax_2acx_3bcx_4cdax_5dcd$, for which $\mu(u) = 5$. This is all we know about the lower bound of μ. Note that the words from [17] and [67] contain linear letters, i.e., letters appearing only once.

Problem 2.6.5. *Suppose that u does not contain linear letters. Is it true that $\mu(u) \leq 3$?*

One can also consider the following Abelian version of avoidability, which generalizes the notion of Abelian k-power free words from Section 2.6.1. Let ϕ be a substitution. We say that a word W *contains an Abelian image of a word w under ϕ* if it contains a word w' obtained from w by replacing each letter x by a permutation of the word $\phi(x)$. We say that w is *avoidable in the Abelian sense* if there are infinitely many words in a finite alphabet that do not contain Abelian images of w. The following nice problem from [77] was solved (in the affirmative) for $n = 1, 2, 3$ in Currie and Linek [78].

Problem 2.6.6. *Let w be a word in an alphabet with n letters. Is it true that w is avoidable in the Abelian sense if and only if the Zimin word Z_n does not contain Abelian images of w?*

http://www.springer.com/978-3-319-08030-7

Combinatorial Algebra: Syntax and Semantics
Sapir, M.V.
2014, XVI, 355 p. 66 illus., Hardcover
ISBN: 978-3-319-08030-7