Square free words

David Stanovský

Introduction

Combinatorics on words is engaged in looking for regularities in words. For instance van der Waerden's Theorem shows, that every sufficiently long word possesses prescribed arithmetic progression of one letter. We will introduce another situation – not every (long) word contains a square, although it seems on the contrary (try to find any!).

Let $\mathbb{N}$ denote the set of all natural numbers $\{0, 1, \dots\}$. By an *alphabet* we mean a finite nonempty set, its elements are called *letters*. A *word over alphabet* A is a finite sequence of letters from A. Empty word (sequence of length 0) is denoted ε. The set of all words over an alphabet A we denote A^* and $A^+ = A^* \setminus \{\varepsilon\}$.

Concatenation of words u and v is denoted by uv. A *morphism* between A^* and B^* is a map $f : A^* \to B^*$ such that $f(\varepsilon) = \varepsilon$ and $f(uv) = f(u)f(v)$ for every $u, v \in A^*$. A word u is called a *factor* of v, if there exist words x, y such that $v = xuy$. The word u is called a *left factor*, if $x = \varepsilon$. If u is a word, than $|u|$ means length of the word and u^R is a word read in an opposite direction. *Palindrome* is such word, that $u = u^R$.

A *square* is a word of the form uu, where u is some nonempty word. Word contains a square, if one of its factors is square. Otherwise we call the word *square-free*. E.g. $abcacbacbc$ contains the square $acbacb$, but $abcacbabcb$ is square-free (as will be shown later).

We will construct an infinite square-free word over an alphabet with three letters. Clearly, then there exist infintely many finite square-free words. There exists no square-free word over two-letter alphabet of the length more then 3 (the only ones are a, b, ab, ba, aba, bab). The infinite square-free word will be derived from the so-called word of Thue-Morse, which contains no factor of the form $avava$, where a is a letter and v is a word.

Axel Thue, Norwegian mathematician, was first who was interested in this topic. He constructed the same words as we will in his papers written in 1906 and 1912. This was independetly described and improved by M. Morse in 1921. Then many other papers were written on related topics.

Preliminaries

Let A denote an alphabet. Let $u, v \in A^+$, u occures at least twice in v. Then there exist $x, y, x', y' \in A^*$ such that $|x| < |x'|$, $|y| > |y'|$ and $v = xuy = x'uy'$. Occurences of u in v are called

(1) *disjoint*, if $|x'| > |xu|$, i.e. $v = xuzuy'$ for some z.
(2) *adjacent*, if $|x'| = |xu|$, i.e. $v = xuuy'$.
(3) *overlapping*, if $|x'| < |xu|$.

A good description of the third possibility is provided by the following lemma. By an *overlapping factor* we mean a factor of the form $avava$, where $a \in A, v \in A^*$.

Lemma. *A word $w \in A^*$ contains two overlapping occurrences of some nonempty word, iff it contains some overlapping factor.*

Proof.
I. Let $w = xuy = x'uy'$ such that $0 \leq |x| < |x'| < |xu| < |x'u| \leq |w|$. Then $x' = xs, xu = x'z, x'u = xut$ for some nonempty words s, z, t. Then $(*)$ $u = sz = zt$. Denote a the first letter of s, i.e. of z too (by $(*)$). So $s = as', z = az'$. Thus $u = sz = as'az'$, so $w = x'uy' = xsuy' = xas'as'az'y'$ and clearly $as'as'a$ is an overlapping factor in w.

II. If $w = xavavay$, then ava has an overlapping occurrence in w.

By *word* we mean always finite word. Now we will define an *infinite word*. It is an infinite sequence of letters, i.e. a function $\mathbf{a} : \mathbb{N} \to A$, denoted by $\mathbf{a} = a_0a_1a_2\ldots$, where $a_i = \mathbf{a}(i)$ for each $i \in \mathbb{N}$. Let us define $\mathbf{a}^{[k]} = a_0 \ldots a_{k-1}$ and call it a *left factor* of $\mathbf{a}$ of the length k. If $u = \mathbf{a}^{[k]}$, then we shall write $\mathbf{a} = u\mathbf{b}$, where $\mathbf{b}$ is such that $b_i = a_{i+|u|}$, $i \in \mathbb{N}$. A word u we call a *factor* of $\mathbf{a}$, if $\mathbf{a} = xu\mathbf{b}$ for some x and $\mathbf{b}$.

Inifinite words are useful for decription of properties of finite words which are *stable for factors*. It means that if some word possesses this property, then so do all its factors. Clearly square-freeness is stable for factors.

We say, that an infinite word $\mathbf{a}$ has a property P, if all its factors do so. This clarifies the sense of the term "infinite square-free word".

Let us denote L_P the set of all words with the property P. Thus, if P is stable for factors and $w \in L_P$, then all factors of w are in L_P.

Lemma. *Let P be a property of words over A stable for factors. Then L_P is infinite, iff there exist an infinite word over A having the property P.*

Proof.
I. Suppose L_P infinite and A finite. There must exist some $a_0 \in A$ such that infinitely

many words from L_P start with a_0. Let us denote $L_0 = \{b \in L_P : b = a_0 y$ for some $y \in A^*\}$. The same argument allows us to construct by induction sets L_1, L_2, ... of words starting by $a_0a_1, a_0a_1a_2, \dots$ Letters $a_0, a_1, \dots$ form an infinite word $\mathbf{a} = a_0a_1 \dots$ with the property P.

II. Converse direction is quite clear. If $\mathbf{a}$ is an infinite word with the property P, then for every i natural $\mathbf{a}^{[k]} \in L_P$, so L_P is infinite.

The proof shows us an algorithm for derivation of the infinite word with P from infinitely many finite words possessing P.

Let us consider a sequence $w_0, w_1, \dots$ of words over A such that w_n is a left factor of w_{n+1} for all n natural. Denote $\mathbf{a}$ an infinite word satisfying $\mathbf{a}^{[k]} = w_n$ for all $k = |w_n|, n \in \mathbb{N}$. We write $\mathbf{a} = \lim w_n$ and call it a *limit* of sequence $(w_n)_{n=0}^{\infty}$.

Imagine this special case. Let $\alpha : A^* \to A^*$ be a morphism satisfying $\alpha(a) \neq \varepsilon$ for all $a \in A$ and $\exists a_0 \in A$ such that $\alpha(a_0) = a_0 u$ for some $u \in A^+$ (we say that α satisfies ($\heartsuit$) for a_0). Thus for every n natural $\alpha^{n+1}(a_0) = \alpha^n(\alpha(a_0)) = \alpha^n(a_0 u) = \alpha^n(a_0)\alpha^n(u)$, so $\alpha^n(a_0)$ is a left factor of $\alpha^{n+1}(a_0)$. The limit of this sequence is called *limit of iterating α on a_0* and it is denoted $\alpha^\infty(a_0)$.

There is natural extension of a morphism $\alpha : A^* \to A^*$ to infinite words over A. For $\mathbf{b} = b_0b_1 \dots$ is $\alpha(\mathbf{b}) = \alpha(b_0)\alpha(b_1)\dots$ — it is an infinite word because of the first condition in ($\heartsuit$).

Lemma. *Let α satisfies ($\heartsuit$) for a_0 and $\mathbf{a} = \alpha^\infty(a_0)$. Then $\alpha(\mathbf{a}) = \mathbf{a}$.*

Proof.
If u is a left factor of $\alpha(\mathbf{a})$, then so is $\alpha(u)$. Thus every $\alpha^n(a_0)$ is a left factor of $\alpha(\mathbf{a})$. But $\alpha(\mathbf{a})$ starts with a_0 (the second condition in ($\heartsuit$)), so $\alpha(\mathbf{a}) = \lim \alpha^n(a_0) = \alpha^\infty(a_0) = \mathbf{a}$.

Words od Thue-Morse

Let $A = \{a, b\}$ in the rest of the paper. For every $w \in A^*$ we denote $\overline{w}$ the word obtained from w by replacing a to b and vice versa.

Let $\mu : A^* \to A^*$ is a morphism defined by $\mu(a) = ab$ and $\mu(b) = ba$. Clearly μ satisfies ($\heartsuit$) for a and b. Denote

$$\mathbf{t} = \mu^\infty(a) = abbabaabbaababbabaababbaabbabaab\dots$$

$$\overline{\mathbf{t}} = \mu^\infty(b) = baababbaabbabaababbabaabbaababba\dots$$

Lemma. *Let $u_0 = a, v_0 = b, u_{n+1} = u_n v_n, v_{n+1} = v_n u_n$ for all natural n. Then for* *all* $n \in \mathbb{N}$ *hold*

(1) $u_n = \mu^n(a), v_n = \mu^n(b)$.
(2) $v_n = \overline{u_n}$, $u_n = \overline{v_n}$.
(3) u_{2n}, v_{2n} *are palindromes,* $u_{2n+1}^R = v_{2n+1}$

Proof.
By induction on n. The case $n = 0$ is clear.

(1) $u_{n+1} = u_n v_n = \mu^n(a)\mu^n(b) = \mu^n(ab) = \mu^n(\mu(a)) = \mu^{n+1}(a)$. The rest is similar.

(2) $v_{n+1} = v_n u_n = \overline{u_n v_n} = \overline{u_n v_n} = \overline{u_{n+1}}$. The rest is similar.

(3) $u_{2n+2} = u_{2n+1}v_{2n+1} = u_{2n+1}u_{2n+1}^R$ which is a palindrome. For v_n similar. $u_{2n+1}^R = (u_{2n}v_{2n})^R = v_{2n}^R u_{2n}^R = v_{2n}u_{2n} = v_{2n+1}$.

Now we will prove, that $\mathbf{t}$ contains no overlapping factor. Then $\mathbf{t}$ is also cube-free, because if uuu is a factor of $\mathbf{t}$, $u = au'$ for $a \in A$, then $au'au'a$ is an overlapping factor of t providing contradicition.

We will need two lemmas.

Lemma 1. *If* $X = \{ab, ba\}$, $x \in X^*$, *then* $axa \notin X^*$, $bxb \notin X^*$.

Proof.
Let $x \in X^*$. We use induction on $|x|$. For $|x| = 0$ is $aa, bb \notin X^*$. Now let x satisfies $axa \in X^*$ and for all shorter words proposition holds. Let us write $axa = u_0 \dots u_k$, $u_i \in X$. Thus must be $u_0 = ab$ and $u_k = ba$. So $u = u_1 \dots u_{k-1} \in X^*$, u is shorter then x and $bub = x \in X^*$. That is contradiction with an induction assuption. For bxb similarly.

Lemma 2. *If* $w \in A^+$ *contains no overlapping factor, then neither does* $\mu(w)$.

Proof.
Suppose $\mu(w)$ contains an overlapping factor. Then $\mu(w) = xcvcvcy$ for some $x, v, y \in A^*$, $c \in A$. Note, that $\mu(w) \in X^*$ for $X = \{ab, ba\}$ and thus $|\mu(w)|$ is even. But $|cvcvc|$ is odd, so either

(1) $|x|$ is odd, $|y|$ is even and thus $xc, vcvc, y \in X^*$, or
(2) $|x|$ is even, $|y|$ is odd and thus $x, cvcv, cy \in X^*$.

In both cases is $|v|$ odd (if it is even, then $cvc \in X^*$, $v \in X^*$ contradicting lemma 1). So either $vc \in X^*$ or $cv \in X^*$.

(1) We can write $w = rsst$ so that $\mu(r) = xc$, $\mu(s) = vc$, $\mu(t) = y$. Words r, s finish by the same letter $\overline{c}$, so $r = r'\overline{c}, s = s'\overline{c}$. Thus $w = r'\overline{c}s'\overline{c}s'\overline{c}t$ contains an overlapping factor, contradiction.

(2) We can write $w = rsst$ so that $\mu(r) = x$, $\mu(s) = cv$, $\mu(t) = cy$. Words s, t start by the same letter c, so $s = cs', t = ct'$. Thus $w = rcs'cs'ct'$ contains an overlapping factor, contradiction.

Theorem. *An infinite word* $\mathbf{t}$ *contains no overlapping factor.*

Proof.
Let x be an overlapping factor in $\mathbf{t}$. There must exist (sufficiently large) k such that $\mu^k(a)$ has an overlapping factor x. But a doesn't have any overlapping factor, so by lemma 2 also $\mu(a)$ doesn't, so also $\mu^2(a)$, etc., also $\mu^k(a)$ has no overlapping factor. That is contradiction.

Square-free words

We know, that no square-free word longer then 3 occurs over two-letter alphabet. So let $B = \{a, b, c\}$ and

$$\delta : B^* \to A^*, \quad \delta(a) = abb, \quad \delta(b) = ab, \quad \delta(c) = a.$$

If $\mathbf{a}$ is an infinite word without overlapping factors strating with letter a, then there is a unique factorization $\mathbf{a} = y_0 y_1 \ldots$, where $y_n \in \{a, ab, abb\} = \delta(B)$ for all $n \in \mathbb{N}$. It is true, because every a in $\mathbf{a}$ is followed by at most two letters b and then again by a ($\mathbf{a}$ is cube-free). So with every a starts some $w_n \in \delta(B)$ of the length (number of b)+1. Thus it is clear, that there exist a unique infinite word $\mathbf{b}$ over B such that $\delta(\mathbf{b}) = \mathbf{a}$.

Theorem. *If* $\mathbf{a}$ *is an infinite word over* A *starting with* a *without overlapping factors and* $\mathbf{b}$ *is such that* $\delta(\mathbf{b}) = \mathbf{a}$, *then* $\mathbf{b}$ *is square-free.*

Proof.
Let $\mathbf{b}$ contains a square uu, denote d next letter after one of its occurrences (i.e. $\mathbf{b} = xuud\mathbf{c}$ for some $x, \mathbf{c}$). Hence $\delta(uud)$ is a factor of $\mathbf{a}$. Denote v, w words satisfying $\delta(u) = av$, $\delta(d) = aw$. Then $\delta(uud) = \delta(u)\delta(u)\delta(d) = avavaw$ is a factor of $\mathbf{a}$, so $\mathbf{a}$ contains an overlapping factor. Contradiction.

Let us denote $\mathbf{m}$ the square-free word obtained from $\mathbf{t}$ (i.e. $\delta(\mathbf{m}) = \mathbf{t}$). One can check, that

$$\mathbf{m} = abcacbabcbacabcacbacabcbabcacbabcbacabcbabc\ldots$$

The theorem does not hold conversely – there exists a square-free infinite word $\mathbf{b}$ over B, such that $\delta(\mathbf{b})$ has an overlapping factor.

We can also systematically generate finite square-free words by so called *square-*

free morphisms. That are such morphisms $\alpha : A^* \to B^*$ that satisfy $\alpha(A) \neq \{\varepsilon\}$ and for every square-free word w is $\alpha(w)$ square-free. E.g. a morphism

$$\varphi : B^* \to B^*, \quad a \mapsto abcab, \quad b \mapsto acabcb, \quad c \mapsto acbcacb$$

is square-free. An important theorem due to Bean, Ehrenfeucht and McNulty (1979) describes square-free morphisms.

Theorem. *Let $\alpha : A^* \to B^*$ be a morphism satisfying*

(1) $\alpha(A) \neq \{\varepsilon\}$,
(2) *for every square-free word w of the length at most 3 is $\alpha(w)$ square-free,*
(3) *for all $a, b \in A$ no $\alpha(a)$ is a proper factor of $\alpha(b)$.*

Then α is a square-free morphism.

Literature

This is a shortend version of the original paper, which can be found on author's WWW pages `http://www.karlin.mff.cuni.cz/~stanovsk/math`.

The paper was written using the book M. Lothaire, *Combinatorics on Words*, Cambridge University Press, 1983, 1997, chapter 2.