Theoretical Computer Science ••• (••••) •••–•••

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Abelian-square-rich words ☆

Gabriele Fici [a,*], Filippo Mignosi [b], Jeffrey Shallit [c]

[a] *Dipartimento di Matematica e Informatica, Università di Palermo, Palermo, Italy*
[b] *Dipartimento di Ingegneria e Scienze dell'Informazione e Matematica, Università dell'Aquila, L'Aquila, Italy*
[c] *School of Computer Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada*

ARTICLE INFO

Article history:
Received 24 May 2016
Received in revised form 7 February 2017
Accepted 9 February 2017
Available online xxxx

Keywords:
Abelian square
Thue–Morse word
Sturmian word

ABSTRACT

An abelian square is the concatenation of two words that are anagrams of one another. A word of length n can contain at most $\Theta(n^2)$ distinct factors, and there exist words of length n containing $\Theta(n^2)$ distinct abelian-square factors, that is, distinct factors that are abelian squares. This motivates us to study infinite words such that the number of distinct abelian-square factors of length n grows quadratically with n. More precisely, we say that an infinite word w is *abelian-square-rich* if, for every n, every factor of w of length n contains, on average, a number of distinct abelian-square factors that is quadratic in n; and *uniformly abelian-square-rich* if every factor of w contains a number of distinct abelian-square factors that is proportional to the square of its length. Of course, if a word is uniformly abelian-square-rich, then it is abelian-square-rich, but we show that the converse is not true in general. We prove that the Thue–Morse word is uniformly abelian-square-rich and that the function counting the number of distinct abelian-square factors of length $2n$ of the Thue–Morse word is 2-regular. As for Sturmian words, we prove that a Sturmian word s_α of angle α is uniformly abelian-square-rich if and only if the irrational α has bounded partial quotients, that is, if and only if s_α has bounded exponent.

1. Introduction

A fundamental topic in combinatorics on words is the study of repetitions. A *repetition* in a word is a factor that is formed by the concatenation of two or more identical blocks. The simplest kind of repetition is a *square*, that is, the concatenation of two copies of the same block, such as the English word `hotshots`. A famous conjecture of Fraenkel and Simpson [20] states that a word of length n contains fewer than n distinct square factors. Experiments strongly suggest that the conjecture is true, but a theoretical proof of the conjecture seems difficult. In [20], the authors proved a bound of $2n$. In [25], Ilie improved this bound to $2n - \Theta(\log n)$, and recently Deza et al. showed the current best bound of $\frac{11}{6}n$ [12], but the conjectured bound is still out of reach.

Other variations on counting squares include counting squares in partial words (e.g., [5]) and pseudo-repetitions (e.g., [22]).

Among the different generalizations of the notion of repetition, a prominent one is that of an abelian repetition. An *abelian repetition* in a word is a factor that is formed by the concatenation of two or more blocks that have the same number of occurrences of each letter in the alphabet. Of course, the simplest kind of abelian repetition is an *abelian square*,

☆ Some of the results contained in this paper were presented (without the third author) at the 10th International Conference on Words, WORDS 2015 [19].
* Corresponding author.
E-mail addresses: gabriele.fici@unipa.it (G. Fici), filippo.mignosi@univaq.it (F. Mignosi), shallit@cs.uwaterloo.ca (J. Shallit).

http://dx.doi.org/10.1016/j.tcs.2017.02.012
0304-3975/© 2017 Elsevier B.V. All rights reserved.

that is, the concatenation of a word with an anagram of itself, such as the English word `intestines`. Abelian squares were considered in 1961 by Erdős [16], who conjectured that there exist infinite words avoiding abelian squares. This conjecture was later confirmed, and the smallest possible size of an alphabet for which it holds is known to be 4 [26].

We focus on the maximum number of distinct abelian squares that a word can contain. In contrast to the case of ordinary squares, a word of length n can contain $\Theta(n^2)$ distinct abelian-square factors (see [27]). Since the total number of factors in a word of length n is quadratic in n, this means that there exist words in which a constant fraction of all factors are abelian squares. So we turn our attention to infinite words, and we ask whether there exist infinite words such that for every n the factors of length n contain, on average, a number of distinct abelian-square factors that is quadratic in n. We call such an infinite word *abelian-square-rich*. Since a random binary word of length n contains $\Theta(n\sqrt{n})$ distinct abelian-square factors [10], the existence of abelian-square-rich words is not immediate. We also introduce *uniformly abelian-square-rich* words; these are infinite words such that for every n, every factor of length n contains a quadratic number of distinct abelian squares. Of course, if a word is uniformly abelian-square-rich, then it is abelian-square-rich, but the converse is not true in general — we provide in this paper an example of a word that is abelian-square-rich but not uniformly abelian-square-rich. However, we show that for linearly recurrent words the two definitions are equivalent. Moreover, we prove that if an infinite word w is uniformly abelian-square-rich, then w has bounded exponent (that is, there exists an integer $k \geq 2$ such that w does not contain any repetition of order k as a factor).

We then prove that the famous Thue–Morse word is uniformly abelian-square-rich. Furthermore, we look at the function that counts the number of distinct abelian squares of length $2n$ in the Thue–Morse word and prove that this function is 2-regular.

Then we look at the class of Sturmian words; these are aperiodic infinite words with the lowest possible factor complexity. In this case, we prove that a Sturmian word has bounded exponent if and only if it is uniformly abelian-square-rich, and leave open the question of determining whether a Sturmian word is not abelian-square-rich in the case when it does not have bounded exponent.

2. Notation and background

Let $\Sigma = \{a_1, a_2, \ldots, a_\sigma\}$ be an ordered σ-letter alphabet. Let Σ^* stand for the free monoid generated by Σ, whose elements are called *words* over Σ. The *length* of a word w is denoted by $|w|$. The *empty word*, denoted by ε, is the unique word of length zero and is the neutral element of Σ^*. We also define $\Sigma^+ = \Sigma^* \setminus \{\varepsilon\}$.

A *prefix* (respectively, a *suffix*) of a word w is a word u such that $w = uz$ (respectively, $w = zu$) for some word z. A *factor* of w is a prefix of a suffix (or, equivalently, a suffix of a prefix) of w. The set of prefixes, suffixes and factors of the word w are denoted, respectively, by $\mathrm{Pref}(w)$, $\mathrm{Suff}(w)$ and $\mathrm{Fact}(w)$. From the definitions, we have that ε is a prefix, a suffix and a factor of every word.

A word w is a *k-power* (also called a *repetition of order k*), for an integer $k \geq 2$, if there exists a nonempty word u such that $w = u^k$. A 2-power is called a *square*. The *period* of a word $w = w_1 w_2 \cdots w_{|w|}$ is the minimal integer p such that $w_{i+p} = w_i$ for every $1 \leq i \leq |w| - p$. The *exponent* $e(w)$ of a word w is the ratio between its length $|w|$ and its period p. For example, the period of $w = abaab$ is $p = 3$, hence $e(w) = 5/3$. Of course, if a word w avoids k-powers (that is, no factor of w is a k-power), then the supremum of the exponents of factors of w is smaller than k.

For a word w and a letter $a_i \in \Sigma$, we let $|w|_{a_i}$ denote the number of occurrences of a_i in w. The *Parikh vector* (sometimes called the *composition vector*) of a word w over $\Sigma = \{a_1, a_2, \ldots, a_\sigma\}$ is the vector $P(w) = (|w|_{a_1}, |w|_{a_2}, \ldots, |w|_{a_\sigma})$. An *abelian k-power* is a nonempty word of the form $v_1 v_2 \cdots v_k$ where all the v_i have the same Parikh vector (and therefore in particular the same length). An abelian 2-power is called an *abelian square*; an example in English is the word `reappear`.

An *infinite word* w over Σ is an infinite sequence of letters from Σ, that is, a function $w : \mathbb{N} \mapsto \Sigma$. An infinite word is *recurrent* if each of its factors occurs infinitely often. Given an infinite word w, the *recurrence index* $R_w(n)$ of w is defined to be the least integer m such that every factor of w of length m contains all factors of w of length n, or $+\infty$ if such an integer does not exist. If the recurrence index is finite for every n, the infinite word w is called *uniformly recurrent* and the function $R_w(n)$ the *recurrence function* of w. A uniformly recurrent word is of course recurrent, but the converse is not always true. For example, the *Champernowne word* $w = 011011100101\cdots$, obtained by concatenating the base-2 representations of the natural numbers, is recurrent but not uniformly recurrent (to see this, it is sufficient to observe that it contains arbitrarily large consecutive blocks of the same letter). A uniformly recurrent word w is called *linearly recurrent* if the ratio $R_w(n)/n$ is bounded by a constant. Given a linearly recurrent word w, the real number $r_w = \limsup_{n\to\infty} R_w(n)/n$ is called the *recurrence quotient* of w. The *factor complexity function* (sometimes called *subword complexity*) of an infinite word w is the integer function $p_w(n)$ defined by $p_w(n) = |\mathrm{Fact}(w) \cap \Sigma^n|$. An infinite word w has *linear complexity* if $p_w(n) = O(n)$. In particular, if a word is linearly recurrent, then it has linear complexity (see, for example, [15]).

A *substitution* over the alphabet Σ is a map $\tau : \Sigma \mapsto \Sigma^+$. A substitution τ over Σ can be naturally extended to a (non-erasing) morphism from Σ^* to Σ^*. A substitution can be iterated: for every substitution τ and every $n > 0$, using the extension to a morphism, one can define the substitution τ^n. A substitution τ is *r-uniform* if there exists an integer $r \geq 1$ such that for all $a \in \Sigma$, $|\tau(a)| = r$. A substitution is called *uniform* if it is r-uniform for some $r \geq 1$. A substitution τ is *primitive* if there exists an integer $n \geq 1$ such that for every $a \in \Sigma$, the word $\tau^n(a)$ contains every letter of Σ at least once. In this paper, we will only consider primitive substitutions such that $\tau(a_1) = a_1 v$ for a letter a_1 and some nonempty