A SHORT PROOF THAT SHUFFLE SQUARES ARE
7-AVOIDABLE

GUILLAUME GUÉGAN¹ AND PASCAL OCHEM²

Abstract. A shuffle square is a word that can be partitioned into two identical words. We obtain a short proof that there exist exponentially many words over the 7 letter alphabet containing no shuffle square as a factor. The method is a generalization of the so-called power series method using ideas of the entropy compression method as developed by Gonçalves, Montassier, and Pinlou.

1991 Mathematics Subject Classification. 68R15.

Introduction

Entropy compression has been used to avoid squares [5] and patterns [9] in infinite words over a small alphabet. The proofs require many features (an algorithm, a record, an analysis of the size the record,...). Gonçalves, Montassier, and Pinlou [4] have recently obtained a generic way of using the entropy compression method in the context of graph coloring that avoids a lot of these technicalities.

In a recent paper [8], we have used ideas from the entropy compression method to generalize the power series method as used in combinatorics on words by Bell and Goh [1], Rampersad [10], and Blanchet-Sadri and Woodhouse [2]. We describe this method in Section 1 to make the paper self-contained.

A shuffle square is a word that can be partitioned into two identical words. For example, every square is a shuffle square, \(aabbcc\) and \(abacbc\) are shuffle squares of \(abc\), and \(ccbcbc\) is a shuffle square of \(cbca\).

Recently, Currie [3] has answered a question of Karhumäki by showing that there exist infinite words over a finite (but large) alphabet containing no shuffle square as a factor using the Lovász local lemma. Then Müller has lowered the alphabet size to 10 in his thesis [7] and has also proved that shuffle cubes are[

Keywords and phrases: Combinatorics on words, Shuffle square, Entropy compression

¹ Univ. Montpellier 2, LIRMM, guegan@lirmm.fr
² CNRS, LIRMM, ochem@lirmm.fr
© EDP Sciences 1999
avoidable over the 6 letter alphabet. We apply the method in Section 1 to obtain the following result in Section 2.

**Theorem 0.1.** There exist at least $5.59^n$ words of length $n$ over the 7 letter alphabet containing no shuffle square as a factor.

Grytczuk, Kozik, and Zaleski [6] have an independent proof of the list version of Theorem 0.1 using another flavor of entropy compression and different parameters. Notice that words avoiding shuffle squares avoid in particular the patterns $AA$ and $ABACBC$. We have checked that words over 3 letters avoiding $AA$ and $ABACBC$ have finite length, so at least 4 letters are needed to avoid shuffle squares. Thus, the minimum alphabet size for an infinite word avoiding shuffle squares remains an open problem and is between 4 and 7.

1. **Description of the method**

Let $\Sigma_m = \{0, 1, \ldots, m - 1\}$ be the $m$-letter alphabet and let $L \subset \Sigma_m^*$ be a factorial language defined by a set $F$ of forbidden factors of length at least 2. We denote the factor complexity of $L$ by $n_i = L \cap \Sigma_m^i$. We define $L'$ as the set of words $w$ such that $w$ is not in $L$ and the prefix of length $|w| - 1$ of $w$ is in $L$. For every forbidden factor $f \in F$, we choose a number $1 \leq s_f \leq |f|$. Then, for every $i \geq 1$, we define an integer $a_i$ such that

$$a_i \geq \max_{w \in L} |\{v \in \Sigma_m^i \mid uv \in L', uv = bf, f \in F, s_f = i\}|. \quad (1)$$

We consider the formal power series $P(x) = 1 - mx + \sum_{i \geq 1} a_i x^i$. If $P(x)$ has a positive real root $x_0$, then $n_i \geq x_0^{-i}$ for every $i \geq 0$.

Let us rewrite that $P(x_0) = 1 - mx_0 + \sum_{i \geq 1} a_i x_0^i = 0$ as

$$m - \sum_{i \geq 1} a_i x_0^{i-1} = x_0^{-1} \quad (2)$$

Since $n_0 = 1$, we will prove by induction that $\frac{n_i}{n_{i-1}} \geq x_0^{-1}$ in order to obtain that $n_i \geq x_0^{-i}$ for every $i \geq 0$. By using (2), we obtain the base case: $\frac{n_1}{n_0} = n_1 = m \geq x_0^{-1}$. Now, for every length $i \geq 1$, there are:

- $m^i$ words in $\Sigma_m^i$,
- $n_i$ words in $L$,
- at most $\sum_{1 \leq j \leq i} n_i - j a_j$ words in $L'$,
- $m(m^{i-1} - n_{i-1})$ words in $\Sigma_m^i \setminus \{L \cup L'\}$.

This gives $n_i + \sum_{1 \leq j \leq i} n_i - j a_j + m(m^{i-1} - n_{i-1}) \geq m^i$, that is, $n_i \geq mn_{i-1} - \sum_{1 \leq j \leq i} n_i - j a_j$. 
$$\frac{m_i}{n_{i-1}} \geq m - \sum_{1 \leq j \leq n_i} a_j \frac{m_i}{n_{i-1}} \geq m - \sum_{1 \leq j \leq n_i} a_j x_j^{n_i-1} \quad \text{By induction}$$

$$\geq m - \sum_{j \geq 1} a_j x_0^{j-1} = x_0^{-1} \quad \text{By (2)}$$

2. Avoiding shuffle squares

We apply the method of the previous section to the avoidance of shuffle squares. The $q$-prefix (resp. $q$-suffix) of a word is its prefix (resp. suffix) of length $q$. A shuffle square is minimal if it does not contain a smaller shuffle square as a factor. A shuffle square is small if its length is two and is large otherwise. The set $F$ of forbidden factors contains every minimal shuffle square. We set $s_f = 1$ if $f \in F$ is small and $s_f = |f| - 2$ otherwise.

We set $a_1 = 1$ because $s_f = 1$ only for small shuffle squares and there is only one way to extend a prefix by one letter to obtain a suffix $xx$. Thus, there are at most $m$ possibilities for $w$ of length $i$, we associate the height function $h$: $[0, \ldots, 2i] \rightarrow \mathbb{Z}$ defined as follows:

- $h(0) = 0$.
- For $0 < j < 2i$, $h(j) = h(j - 1) + 1$ if the $j$-th letter of $f$ belongs to the subword $w$ containing the first letter of $f$, and $h(j) = h(j - 1) - 1$ otherwise.

Since $f$ is a shuffle square, we have $h(2i) = 0$. Moreover, if $h(j) = 0$ for some $0 < j < 2i$, then the prefix of length $j$ of $f$ is a shuffle square. So, if $h$ is the height function of a minimal shuffle square, then $h(j) > 0$ for every $0 < j < 2i$. Thus, every height function of a minimal shuffle square is associated to a unique Dyck word of length $2i - 2$. The number of height functions is thus at most $\frac{(2i-2)!}{i(i-1)!}$. According to [1], we need to bound the number of solutions to $uv = bf$ such that $u$ is fixed and $|v| = s_f = |f| - 2 = 2i - 2$. The 2-prefix of $f$ is fixed since it corresponds to the 2-suffix of $u$. Notice that the 2-prefix of a large minimal shuffle square of a word $w$ is equal to the 2-prefix of $w$, so the 2-prefix of $w$ is also fixed. Thus, there are at most $m^{i-2}$ possibilities for $w$. Since $f$ is determined by $w$ and its height function, there are at most $m^{i-2} \frac{(2i-2)!}{i(i-1)!}$ possibilities for $f$. So we set $a_{2i-2} = m^{i-2} \frac{(2i-2)!}{i(i-1)!}$ and consider the polynomial

$$P(x) = 1 - mx + x + \sum_{i \geq 2} m^{i-2} \frac{(2i-2)!}{i(i-1)!} x^{2i-2} = 1 - (m - 1)x + \left(\frac{2x}{1 + \sqrt{1 - 4mx}}\right)^2.$$

For $m = 6$, $P(x)$ has no positive root. For $m = 7$, we have $P(x_0) = 0$ with $x_0 = 0.1788487593 \ldots$. So there exists at least $\alpha^n$ words of length $n$ over $\Sigma_7$ that avoid shuffle squares, where $\alpha = x_0^{-1} = 5.5913163944 \ldots$
References


Communicated by (The editor will be set by the publisher).