4-tangrams are 4-avoidable

Pascal Ochem[a] Théo Pierron[b]

Submitted: TBD; Accepted: TBD; Published: TBD

Abstract

A tangram is a word in which every letter occurs an even number of times. Thus it can be cut into parts that can be arranged into two identical words. The *cut number* of a tangram is the minimum number of required cuts in this process. Tangrams with cut number one corresponds to squares. For $k \geqslant 1$, let $t(k)$ denote the minimum size of an alphabet over which an infinite word avoids tangrams with cut number at most k. The existence of infinite ternary square-free words shows that $t(1) = t(2) = 3$. We show that $t(3) = t(4) = 4$, answering a question from Dębski, Grytczuk, Pawlik, Przybyło, and Śleszyńska-Nowak.

Mathematics Subject Classifications: 68R15

1 Introduction

A *tangram* is a word in which every letter occurs an even number of times, possibly zero. In particular, tangrams are a generalization of squares. In this article, we consider a classification of tangrams depending on how close they are from being a square. This relies on the so-called *cut number* of a tangram, recently introduced by Dębski, Grytczuk, Pawlik, Przybyło, and Śleszyńska-Nowak [3]. The *cut number* of a tangram is defined as the minimum number of cuts needed so that the parts can be rearranged into two identical words. Tangrams with cut number at most k are called k-tangrams. Note that 1-tangrams are exactly squares, and the larger the cut number, the farther the tangram is from a square.

Let $\Sigma_q = \{1, 2, \ldots, q-1\}$ denote the q-letter alphabet. It is straightforward to check that every binary word of length 4 contains a square, while a famous theorem from Thue in 1906 asserts that there exist infinite words avoiding squares over Σ_3. In [3], the authors consider a similar question by investigating (infinite) words without tangrams: since every infinite word must contain some tangram, they consider the relation between the size of the alphabet and the cut number of the excluded tangrams. For $k \geqslant 1$, the authors thus

[a]LIRMM, CNRS, Université de Montpellier, France. ochem@lirmm.fr
[b]Univ Lyon, UCBL, CNRS, INSA Lyon, LIRIS, UMR5205, F-69622 Villeurbanne, France theo.pierron@univ-lyon1.fr

define $t(k)$ as the minimum alphabet size such that there exists an infinite word avoiding k-tangrams. By definition, $t(k)$ is non-decreasing, and since every 2-tangram contains a square, the result of Thue shows that $t(1) = t(2) = 3$.

The other results from [3] are summarized in the following.

Theorem 1 ([3])**.**

- $t(k) \leqslant 1024 \lceil \log_2 k + \log_2 \log_2 k \rceil$ *for every* $k \geqslant 3$*.*
- $t(k) \leqslant k + 1$ *for every* $k \geqslant 4$*.*
- $4 \leqslant t(3) \leqslant t(4) \leqslant 5$

Moreover, the authors leave as an open problem the exact value of $t(3)$. In this article, we prove the following.

Theorem 2. $t(3) = t(4) = 4$*.*

2 Preliminaries

To obtain Theorem 2, we heavily use the following relation observed in [3] between k-tangrams and patterns. A pattern P is a finite word over the alphabet $\Delta = \{A, B, \ldots\}$, whose letters are called *variables*. An *occurrence* of a pattern P in a word $w \in \Sigma^*$ is a non-erasing morphism $h : \Delta^* \to \Sigma^*$ such that $h(P)$ is a factor of w, and a word w avoids a pattern P if it contains no occurrence of P.

As noticed in [3], a k-tangram is an occurrence of some pattern with at most k variables such that every variable occurs exactly twice. So for every $k \geqslant 1$, there exists a minimum set S_k of such patterns such that avoiding S_k is equivalent to avoiding k-tangrams. Obviously, $S_k \subset S_{k+1}$ for every $k \geqslant 1$. A small case analysis gives the first four sets S_k:

- $S_1 = S_2 = \{AA\}$
- $S_3 = \{AA, ABACBC, ABCACB, ABCBAC\}$
- $S_4 = \{AA, ABACBC, ABCACB, ABCBAC, ABACBDCD, ABACDBDC, ABACDCBD, ABCACDBD, ABCADBDC, ABCADCBD, ABCADCDB, ABCBADCD, ABCBDACD, ABCBDADC, ABCBDCAD, ABCDACBD, ABCDADCB, ABCDBADC, ABCDBDAC, ABCDCADB, ABCDCBAD\}$,

In the next section, we prove Theorem 2 by constructing infinite words over Σ_4 avoiding all patterns in S_4. But first, let us show the weaker result $t(3) \leqslant 4$ as a straightforward (and computer-free) consequence of well-know results in pattern avoidance. Following Cassaigne [2], we associate to each pattern a *formula*, by replacing each variable appearing only once by a dot (such variables are called *isolated*). For example, the formula associated to the pattern $ABBACABADAA$ is $ABBA.ABA.AA$. The factors between the dots are called *fragments*. Similarly to patterns, an *occurrence* of a formula f in a word $w \in \Sigma^*$ is a non-erasing morphism $h : \Delta^* \to \Sigma^*$ such that every fragment of f is mapped under h to a factor of w (note that the order of the fragments does not matter). A word w *avoids* a formula f if it contains no occurrence of f.

Consider the formula $F_3 = AB.BA.AC.CA.BC$. Notice that AA contains an occurrence of F_3. Moreover, $ABACBC$, $ABCACB$, and $ABCBAC$ also contain an occurrence of F_3 since they have 5 distinct factors of length 2. So every pattern in S_3 contains an occurrence of F_3. Baker, McNulty, and Taylor [1] have considered that the fixed point $b_4 \in \Sigma_4^\omega$ of the morphism defined by $\texttt{0} \mapsto \texttt{01}$, $\texttt{1} \mapsto \texttt{21}$, $\texttt{2} \mapsto \texttt{03}$, $\texttt{3} \mapsto \texttt{23}$ (that is, $b_4 = \texttt{01210321012303210121}\ldots$) and shown that b_4 avoids F_3. Then b_4 avoids every pattern in S_3. So b_4 avoids 3-tangrams, which implies that $t(3) \leqslant 4$.

3 Proof of $t(4) \leqslant 4$

Unfortunately, the word b_4 contains the factor $\texttt{03210123}$ which is a 4-tangram. Moreover, backtracking shows that every infinite word over Σ_4 avoiding 4-tangrams must contain a factor aba for some letters a and b. In particular, every $\frac{7}{5}^+$-free word over Σ_4 contains a 4-tangram. More generally, we have not been able to find a word that might witness $t(4) \leqslant 4$ in the literature. Thus we use an ad-hoc construction. The proof will need the following notions. Given a square-free word w, a *repetition* in w is a factor of w of the form uvu. Its *period* is $|uv|$ and its *exponent* is $\frac{|uvu|}{|uv|}$. Given $\alpha \in \mathbb{Q}$ and $n \in \mathbb{N}$, a word w is (α^+, n)*-free* if it does not contain any repetition with period at least n and exponent strictly greater than α. We say that w is α^+*-free* if it is $(\alpha^+, 1)$-free.

Consider the 312-uniform morphism $h : \Sigma_6^* \to \Sigma_4^*$ below. We will show that for every $\frac{6}{5}^+$-free word w over Σ_6, $h(w)$ avoids every pattern in S_4. Together with the result of Kolpakov and Rao [5] that there exist exponentially many $\frac{6}{5}^+$-free infinite words over Σ_6, this implies that there exist exponentially many words over Σ_4 avoiding 4-tangrams.

First, we show that $h(w)$ is $\left(\frac{5}{4}^+, 9\right)$-free by using the following lemma from [6]. A morphism $f : \Sigma^* \to \Delta^*$ is q*-uniform* if $|f(a)| = q$ for every $a \in \Sigma$, and is called *synchronizing* if for all $a, b, c \in \Sigma$ and $u, v \in \Delta^*$, if $f(ab) = uf(c)v$, then either $u = \varepsilon$ and $a = c$, or $v = \varepsilon$ and $b = c$.

Lemma 3. *Let $\alpha, \beta \in \mathbb{Q}$, $1 < \alpha < \beta < 2$ and $n \in \mathbb{N}^*$. Let $h : \Sigma_s^* \to \Sigma_e^*$ be a synchronizing q-uniform morphism (with $q \geqslant 1$). If $h(w)$ is (β^+, n)-free for every α^+-free word w such that $|w| < \max\left(\frac{2\beta}{\beta-\alpha}, \frac{2(q-1)(2\beta-1)}{q(\beta-1)}\right)$, then $h(t)$ is (β^+, n)-free for every (finite or infinite) α^+-free word t.*

We have checked that h is synchronizing and that the h-image of every $\frac{6}{5}^+$-free word of length smaller than $\frac{2\times\frac{5}{4}}{\frac{5}{4}-\frac{6}{5}} = 50$ is $\left(\frac{5}{4}^+, 9\right)$-free. Therefore $h(w)$ is $\left(\frac{5}{4}^+, 9\right)$-free by Lemma 3.

Now we show that every occurrence m of a pattern $P \in S_4$ in a $\left(\frac{5}{4}^+, 9\right)$-free word is such that $|m(P)|$ is bounded (see Table 1). As an example, let us detail the case of $ABCDACBD$. To lighten notations, we write $y = |m(Y)|$ for every variable Y.

Lemma 4. *Let z be a $\left(\frac{5}{4}^+, 9\right)$-free word. Then if z contains an occurrence m of $ABCDACBD$, then $|m(ABCDACBD)| \leqslant 24$.*

Proof. Consider an occurrence m of $ABCDACBD$ in z. The factor $m(ABCDA)$ of z is a repetition with period $|m(ABCD)|$ and exponent $\frac{|m(ABCDA)|}{|m(ABCD)|}$.

Since z is $\left(\frac{5}{4}^+, 9\right)$-free, then $a+b+c+d \leqslant 8$ or $\frac{2a+b+c+d}{a+b+c+d} \leqslant \frac{5}{4}$. The latter inequality gives $\frac{a}{a+b+c+d} \leqslant \frac{1}{4}$ and then

$$3a \leqslant b+c+d. \tag{1}$$

Similarly, the repetition $m(BCDACB)$ implies that $a+b+2c+d \leqslant 8$ or

$$3b \leqslant a+2c+d. \tag{2}$$

$m(CDAC)$ implies that $a+c+d \leqslant 8$ or

$$3c \leqslant a+d. \tag{3}$$

$m(DACBD)$ implies that $a+b+c+d \leqslant 8$ or

$$3d \leqslant a+b+c. \tag{4}$$

Suppose that $a+c+d \geqslant 9$. Then the combination $6 \times \textbf{(1)} + 4 \times \textbf{(2)} + 7 \times \textbf{(3)} + 6 \times \textbf{(4)}$ gives $a+c+d \leqslant 0$, a contradiction. Therefore

$$a+c+d \leqslant 8. \tag{5}$$

This implies

$$c \leqslant 6. \tag{6}$$

Now suppose that $b \geqslant 5$, so that $a+b+2c+d \geqslant 9$. Then the combination $\textbf{(2)}+\textbf{(5)}+\textbf{(6)}$ gives $3b \leqslant 14$, which contradicts $b \geqslant 5$. Therefore

$$b \leqslant 4. \tag{7}$$

By **(5)** and **(7)**, we get that $a+b+c+d \leqslant 12$, hence $|m(ABCDACBD)| \leqslant 24$. $\square$

Now, notice that every pattern in S_4 is *doubled*, that is, every variable appears at least twice [4, 7]. Alternatively, a doubled pattern is a formula with exactly one fragment. The *avoidability exponent* $AE(f)$ of a pattern or a formula f is the largest real x such that every x-free word avoids f. By Lemma 10 in [8], the avoidability exponent of a doubled pattern with 4 variables is at least $\frac{6}{5}$. This bound is not good enough, so we have computed the avoidability exponent of every pattern in S_4, see Table 1. Notice that these avoidability exponents are greater than $\frac{5}{4}$. Then the $\left(\frac{5}{4}^+, 9\right)$-freeness of $h(w)$ ensures that there is no "large" occurrence of a pattern in S_4, that is, such that the period of every repetitions is at least 9. This is witnessed by the combination $6 \times \textbf{(1)} + 4 \times \textbf{(2)} + 7 \times \textbf{(3)} + 6 \times \textbf{(4)}$ in the proof of Lemma 4. Now, to bound the length of the other occurrences, we do not rely on a tedious analysis by hand as in Lemma 4. Instead, the bound in the last column of Table 1 is computed as the maximum of $2(a+b+c+d)$ such that $1 \leqslant a,b,c,d < 100$ and $(a+b+c+d \leqslant 8 \vee 3a \leqslant b+c+d) \wedge (a+b+2c+d \leqslant 8 \vee 3b \leqslant a+2c+d) \wedge (a+c+d \leqslant 8 \vee 3c \leqslant a+d) \wedge (a+b+c+d \leqslant 8 \vee 3d \leqslant a+b+c)$, again with the example of $P = ABCDACBD$ of Lemma 4.

Finally, for every pattern $P \in S_4$, we check exhaustively by computer that $h(w)$ contains no occurrence of P of length at most the corresponding bound. [1] So $h(w)$ avoids every $P \in S_4$. So $h(w)$ avoids 4-tangrams. So $t(4) \leqslant 4$.

Pattern P	P^R	$AE(P)$	Bound on $\lvert m(P)\rvert$
AA	self-reverse	2	16
$ABACBC$	self-reverse	$1.414213562 = \sqrt{2}$	30
$ABCACB$	$ABCBAC$	1.361103081	26
$ABACBDCD$	self-reverse	1.381966011	32
$ABACDBDC$	$ABCBADCD$	$1.333333333 = \frac{4}{3}$	40
$ABACDCBD$	$ABCACDBD$	1.340090632	32
$ABCADBDC$	$ABCBDACD$	1.292893219	32
$ABCADCBD$	self-reverse	1.295597743	28
$ABCADCDB$	$ABCBDCAD$	1.327621756	32
$ABCBDADC$	self-reverse	1.302775638	32
$ABCDACBD$	self-reverse	1.258055872	24
$ABCDADCB$	$ABCDCBAD$	1.288391893	42
$ABCDBADC$	self-reverse	1.267949192	24
$ABCDBDAC$	$ABCDCADB$	1.309212406	44

Table 1: The patterns in S_4, their avoidability exponent, and the upper bound for the length of their occurrences in a $\left(\frac{5}{4}^+, 9\right)$-free word.

4 Concluding remarks

Notice that our words $h(w)$ contain the factor `012130212321` which is a 5-tangram since `0|1|213021|2|3|21` can be rearranged as `213021|2|1|3|0|21`. The exact value of $t(k)$ remains unknown for every $k \geqslant 5$. In particular, we only known that $4 \leqslant t(5) \leqslant 6$. Improving the upper bound on $t(5)$ using the approach in this paper might be tedious, as we expect the set S_5 to be quite large.

[1] The C code to check the properties of the morphism h and the bounds on $|m(P)|$ is available at http://www.lirmm.fr/~ochem/morphisms/tangram4.htm.

0 → 010201321021231201213020103101203013032131031323010302130320312023020103212
0123121021320102031201213021232101312030130230310320131012301021032120213012010
3102101320212303230102031210213212012310201032102123012131203213230231232013121
0301032313021231201213210231323012320321030230313201031021013012031303210310123
1 → 010201312103231302123201020312023021303203010231301321310312301030213013120
3132310121302012032021023012101320103102313031201301021031012303132131032312320
1213212023132301232032103020312302321320323102030121012320213032031030230132023
2103203012303130201032120231323012320321030230313201031021013012031303210310123
2 → 010201312103231302030123031032101310230130320310302131012010312102123020120
2321012130212312031321310123031032030130231013120310302130323010203121013202123
0323102321320131231303210130120103102130313201301023101210302013230231232032130
2030123031032101310230130320310302131012321230313201031021013012031303210310123
3 → 010201312103231302030123031032101301020312102132120123102010321021230121312
0232103023013032031023202130230312030103231302123201312103010231012013202102030
1232032132302312030201320323102321230132131031231302123201020131210323123023213
2031213102312320132303212023020123203231030230313201031021013012031303210310123
4 → 010201312102123101201032021020301232032132302312030201320323102321230313201
0302120231020320132302321203132130131231032123201321310231303121013202102302012
0321012130210203120232010302131012303132023210323123012131203213230231232013121
0301032313021231201213210231323012320321030203123023213203231020301303210310123
5 → 010201312102123101201032021020301232032132302312030201320210231012010302131
0312313013210301023103132013121302312320321323012131020102321203132130131231032
1232013213102313031201301021031012303132023201030213103123130132103010231012013
2021203102101301201032120231323012320321030203123023213203231020301303210310123

References

[1] K. A. Baker, G. F. McNulty, and W. Taylor. Growth problems for avoidable words. *Theoret. Comput. Sci.*, **69(3)** (1989), 319–345.

[2] J. Cassaigne. *Motifs évitables et régularité dans les mots.* PhD thesis, Université Paris VI, 1994.

[3] M. Dębski, J. Grytczuk, B. Pawlik, J. Przybyło, and M. Śleszyńska-Nowak. Words avoiding tangrams. *Annals of Combinatorics* (2024).

[4] A. Domenech and P. Ochem. Doubled patterns with reversal and square-free doubled patterns. *Electron. J. Combin.*, **30(1)** (2023), #P1.50.

[5] R. Kolpakov and M. Rao. On the number of Dejean words over alphabets of 5, 6, 7, 8, 9 and 10 letters. *Theoretical Computer Science* **412** (2011), 6507–6516.

[6] P. Ochem. A generator of morphisms for infinite words. *RAIRO - Theoret. Informatics Appl.*, 40:427–441, 2006.

[7] P. Ochem. Doubled patterns are 3-avoidable. *Electron. J. Combin.*, **23(1)** (2016), #P1.19.

[8] P. Ochem and M. Rosenfeld. Avoidability of palindrome patterns. *Electron. J. Combin.* **28(1)** (2021), #P1.4.