Avoiding Approximate Squares

Dalia Krieger[2], Pascal Ochem[1], Narad Rampersad[2], and Jeffrey Shallit[2]

[1] LaBRI — Université Bordeaux 1
351, cours de la Libération
33405 Talence Cedex FRANCE
ochem@labri.fr

[2] David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, ON, N2L 3G1 CANADA
{d2kriege,nrampersad}@cs.uwaterloo.ca
shallit@graceland.uwaterloo.ca

Abstract. As is well-known, Axel Thue constructed an infinite word over a 3-letter alphabet that contains no squares, that is, no nonempty subwords of the form xx. In this paper we consider a variation on this problem, where we try to avoid *approximate squares*, that is, subwords of the form xx' where $|x| = |x'|$ and x and x' are "nearly" identical.

1 Introduction

A hundred years ago, Norwegian mathematician Axel Thue initiated the study of combinatorics on words [13, 14, 2]. One of his achievements was the construction of an infinite word over a three-letter alphabet that contains no squares, that is, no nonempty subwords of the form xx.

Many variations on this problem have been considered. For example, Brandenburg [3] and Dejean [5] considered the problem of avoiding *fractional powers*. A word w is a *α-power* if it can be written in the form $w = x^n y$, where y is a prefix of x and $\alpha = |w|/|x|$. A word z *contains an α-power* if some subword is a β-power, for $\beta \geq \alpha$; otherwise it *avoids α-powers*. Similarly, a word z *contains an α^+-power* if some subword is a β-power for $\beta > \alpha$; otherwise it *avoids α^+-powers*. We say α-powers (resp. α^+-powers) are *avoidable* over a k-letter alphabet if there exists an infinite word over that alphabet avoiding α-powers (resp., α^+-powers).

Dejean [5] improved Thue's result by showing how to avoid $(7/4)^+$-powers over a 3-letter alphabet; this result is optimal, as every ternary word of length ≥ 39 contains a $7/4$-power. Pansiot [12] showed how to avoid $(7/5)^+$-powers over a 4-letter alphabet. Again, this is optimal, as every quaternary word of length ≥ 122 contains a $7/5$-power.

Dejean also proved that for $k \geq 5$, one cannot avoid $k/(k-1)$-powers over a k-letter alphabet. She conjectured that it was possible to avoid $(k/(k-1))^+$-powers over a k-letter alphabet. This conjecture was proved for $5 \leq k \leq 11$ by Moulin-Ollagnier, for $7 \leq k \leq 14$ by Mohammad-Noori and Currie [9], and for $k \geq 38$ by Carpi [4]. The cases $15 \leq k \leq 37$ remain open.

Another variation is to avoid not all α-powers, but only sufficiently long ones. Entringer, Jackson, and Schatz [6] showed how to construct a word over a 2-letter alphabet that avoids squares xx with $|x| \geq 3$; here the number 3 is best possible.

In this paper we consider yet another variation, but one that seems natural: we consider avoiding *approximate* squares, that is, subwords of the form xx' where x' is "almost the same" as x. The precise definitions are given below. One of our main results is a further strengthening of Dejean's improvement on Thue for 3 letters.

Approximate squares (also known as approximate *tandem repeats* in the biological literature) have been studied before, but from the algorithmic point of view. Landau and Schmidt [7] and Kolpakov and Kucherov [8] both gave efficient algorithms for finding approximate squares in a string.

Notation: we use Σ_k to denote the alphabet of k letters $\{0, 1, 2, \ldots, k-1\}$.

2 Approximate squares

There are at least two natural notions of approximate square. We define them below.

For words x, x' of the same length, define the *Hamming distance* $d(x, x')$ as the number of positions in which x and x' differ. For example, $d(\mathtt{01203}, \mathtt{11002}) = 2$. We say that a word xx' with $|x| = |x'|$ is a *c-approximate square* if $d(x, x') \leq c$. Using this terminology, for example, a 0-approximate square is a square, and a 1-approximate square is either a square or differs from a square in exactly one position.

To avoid c-approximate squares, we would like to enforce the condition $d(x, x') > c$ for all x, x' of the same length, but clearly this is impossible if $|x| \leq c$. To avoid this technicality, we say a word z *avoids c-approximate squares* if for all its subwords xx' where $|x| = |x'|$ we have $d(x, x') \geq \min(c + 1, |x|)$.

This definition is an "additive" version; there is also a "multiplicative" version. Given two words x, x' of the same length, we define their *similarity* $s(x, x')$ as the fraction of the number of positions in which x and x' agree. Formally,

$$s(x, x') := \frac{|x| - d(x, x')}{|x|}.$$

Thus for example, $s(123456, 101406) = 1/2$. The *similarity* of a finite word z is defined to be $\alpha = \max_{\substack{x,x' \text{ subwords of } z \\ |x|=|x'|}} s(x, x')$; we say such a word is α-similar. Thus, a 1-similar finite word contains a square.

For infinite words, the situation is slightly more subtle. We say an infinite word $\mathbf{z}$ is *α-similar* if $\alpha = \sup_{\substack{x,x' \text{ subwords of } \mathbf{z} \\ |x|=|x'|}} s(x, x')$ and there exists at least one subword xx' with $|x| = |x'|$ and $s(x, x') = \alpha$. Otherwise, if $\alpha = \sup_{\substack{x,x' \text{ subwords of } \mathbf{z} \\ |x|=|x'|}} s(x, x')$, but α is not attained by any subword xx' of $\mathbf{z}$, then we say $\mathbf{z}$ is *α^--similar*.

As an example, consider the infinite word over Σ_3,

$$\mathbf{c} = 210201210120210201202\cdots$$

defined to be the length of contiguous blocks of 1's between consecutive 0's in the Thue-Morse sequence $\mathbf{t}$. As is well-known, $\mathbf{c}$ is square-free, so it cannot be 1-similar. However, since $\mathbf{t}$ contains arbitrarily large squares, it follows that $\mathbf{c}$ must contain arbitrarily large 1-approximate squares, and so $\mathbf{c}$ is 1^--similar.

3 Words of low similarity

The main problem of interest is,

Given an alphabet Σ of size k, what is the smallest similarity possible over all infinite words over Σ? We call this the *similarity coefficient* of k.

Answering this question has two aspects. We can explicitly construct an infinite word that is α-similar (or α^--similar). To show that α is best possible, we can construct a tree of all finite words that are β-similar for $\beta < \alpha$. The root of this tree is labeled 0 (which suffices by symmetry), and if a node is labeled w, its children are labeled wa for all $a \in \Sigma$. If a node is β-similar for some $\beta \geq \alpha$, it becomes a leaf and no children are added. We can then use depth-first or breadth-first search to explore this tree. The number of leaves of this tree represents the (finite) number of words beginning with 0 that are β-similar for $\beta < \alpha$, and the height h of the tree is the length of the longest words with this property. The number of leaves at depth h represent the number of maximal words beginning with 0 that are β similar for some $\beta < \alpha$.

We performed this computation for various alphabet sizes k, and the results are reported below in Table 1. For $k = 8$, our method took advantage of some symmetries to speed up the computation, and as a result, we did not compute the number of leaves or maximal strings. For the reported values, these computations represent a proof that the similarity coefficient is at least as large as the α reported.

Alphabet Size k	Similarity Coefficient α	Height of Tree	Number of Leaves	Number of Maximal Words	Lexicographically First Maximal Words
2	1	3	4	1	010
3	3/4	41	2475	36	010201202101201021201210201202101201021 01
4	1/2	9	382	6	012310213
5	2/5	75	3902869	48	012304310342041340120314210412342012403410230420312340321024320410342140243
6	1/3	17	342356	480	01234150325143012
7	?	?	?	?	?
8	1/4	71	—	—	012340567314602516473012756340762135741023640751204356741032715640731 42

Table 1. Similarity bounds

For alphabet size $k = 2$, every infinite word is 1-similar. We now report on larger alphabet sizes.

Theorem 1. *There exists an infinite* $3/4$*-similar word* $\mathbf{w}$ *over* $\{0, 1, 2\}$.

Proof. Let h be the 24-uniform morphism defined by

$$\begin{aligned} 0 &\to 012021201021012102120210 \\ 1 &\to 120102012102120210201021 \\ 2 &\to 201210120210201021012102. \end{aligned}$$

The following lemma may be verified computationally.

Lemma 1. *Let* $a, b, c \in \{0, 1, 2\}$*,* $a \neq b$*. Let* w *be any subword of length* 24 *of* $h(ab)$*. If* w *is neither a prefix nor a suffix of* $h(ab)$*, then* $h(c)$ *and* w *mismatch in at least* 9 *positions.*

Let $\mathbf{w} = h^\omega(0)$. We shall show that $\mathbf{w}$ has the desired property. We argue by contradiction. Suppose that $\mathbf{w}$ contains a subword yy' with $|y| = |y'|$ such that y and y' match in more than $3/4 \cdot |y|$ positions. Let us suppose further that $|y|$ is minimal.

We may verify computationally that $\mathbf{w}$ contains no such subword yy' where $|y| \leq 72$. We therefore assume from now on that $|y| > 72$.

Let $w = a_1 a_2 \cdots a_n$ be a word of minimal length such that $h(w) = xyy'z$ for some $x, z \in \{0, 1, 2\}^*$. By the minimality of w, we have $0 \leq |x|, |z| < 24$.

For $i = 1, 2, \ldots, n$, define $A_i = h(a_i)$. Then if $h(w) = xyy'z$, we can write

$$h(w) = A_1 A_2 \cdots A_n = A'_1 A''_1 A_2 \cdots A_{j-1} A'_j A''_j A_{j+1} \cdots A_{n-1} A'_n A''_n$$

where

$$\begin{aligned} A_1 &= A'_1 A''_1 \\ A_j &= A'_j A''_j \\ A_n &= A'_n A''_n \\ x &= A'_1 \\ y &= A''_1 A_2 \cdots A_{j-1} A'_j \\ y' &= A''_j A_{j+1} \cdots A_{n-1} A'_n \\ z &= A''_n, \end{aligned}$$

and $|A''_1|, |A''_j| > 0$. See Figure 1.

If $|A''_1| > |A''_j|$, then, writing y and y' atop one another, as illustrated in Figure 2, one observes that for $t = j + 1, j + 2, \ldots, n - 1$, each A_t "lines up" with a subword, say B_t, of $A_{t-j} A_{t-j+1}$. We now apply Lemma 1 to conclude that each A_t mismatches with B_t in at least 9 of 24 positions. Consequently, y and y' mismatch in at least $9(j-2)$ positions. Since $j \geq |y|/24 + 1$, we have that

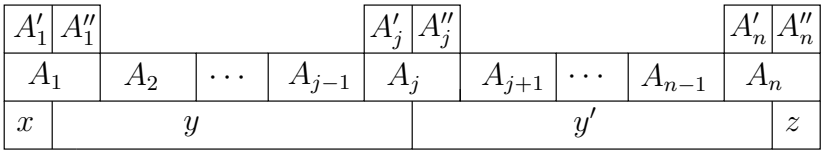

Fig. 1. The string $xyy'z$ within $h(w)$

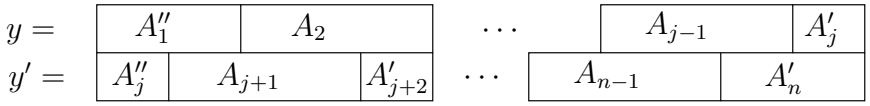

Fig. 2. The case $|A''_1| > |A''_j|$

$9(j-2) \geq 9(|y|/24-1)$. However, $9(|y|/24-1) > |y|/4$ for $|y| > 72$, so that y and y' mismatch in more than $1/4 \cdot |y|$ positions, contrary to our assumption.

If $|A''_1| < |A''_j|$, as illustrated in Figure 3, then a similar argument shows that y and y' mismatch in more than $1/4 \cdot |y|$ positions, contrary to our assumption.

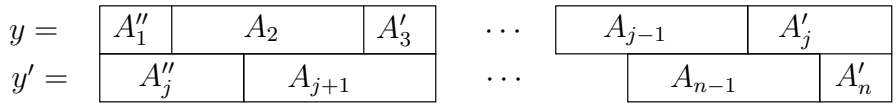

Fig. 3. The case $|A''_1| < |A''_j|$

Therefore $|A''_1| = |A''_j|$. We first observe that any pair of words taken from $\{h(\mathtt{0}), h(\mathtt{1}), h(\mathtt{2})\}$ mismatch at every position. We now consider several cases.

Case 1: $A_1 = A_j = A_n$. Then letting $u = A_1A_2 \cdots A_{j-1}$ and $u' = A_jA_{j+1} \cdots A_{n-1}$, we see that u and u' match in exactly the same number of positions as y and y'.

Case 2: $A_1 = A_j \neq A_n$. Then letting $u = A_1A_2 \cdots A_{j-1}$ and $u' = A_jA_{j+1} \cdots A_{n-1}$, we see that u and u' match in at least as many positions as y and y'.

Case 3: $A_1 \neq A_j = A_n$. Then letting $u = A_2A_3 \cdots A_j$ and $u' = A_{j+1}A_{j+2} \cdots A_n$, we see that u and u' match in at least as many positions as y and y'.

Case 4: $A_1 = A_n \neq A_j$. Then letting $u = A_1A_2 \cdots A_{j-1}$ and $u' = A_jA_{j+1} \cdots A_{n-1}$, we see that u and u' match in exactly the same number of positions as y and y'.

Case 5: A_1, A_j, and A_n are all distinct. Then letting $u = A_1A_2 \cdots A_{j-1}$ and $u' = A_jA_{j+1} \cdots A_{n-1}$, we see that u and u' match in exactly the same number of positions as y and y'.

We finish the argument by considering the word uu'. First observe that either

$$uu' = h(a_1a_2 \cdots a_{j-1})h(a_ja_{j+1} \cdots a_{n-1})$$

or

$$uu' = h(a_2a_3\cdots a_j)h(a_{j+1}a_{j+2}\cdots a_n).$$

Without loss of generality, let us assume that the first case holds.

Recall our previous observation that the words $h(\mathtt{0})$, $h(\mathtt{1})$, and $h(\mathtt{2})$ have distinct letters at every position. Suppose then that there is a mismatch between u and u' occurring within blocks A_t and A_{t+j} for some t, $1 \leq t \leq j$. Then A_t and A_{t+j} mismatch at every position. Moreover, we have $a_j \neq a_{j+t}$. Conversely, if A_t and A_{t+j} match at any single position, then they match at every position, and we have $a_t = a_{t+j}$.

Let $v = a_1a_2\cdots a_{j-1}$ and $v' = a_ja_{j+1}\cdots a_{n-1}$. Let m be the number of matches between u and u'. From our previous observations we deduce that the number of matches m' between v and v' is $m/24$, but since $|v| = |u|/24$, $m'/|v| = m/|u|$. Thus, if $m/|u| > 3/4$, as we have assumed, then $m'/|v| > 3/4$. But the set $\{h(\mathtt{0}), h(\mathtt{1}), h(\mathtt{2})\}$ is a code, so that vv' is the unique pre-image of uu'. The word vv' is thus a subword of $\mathbf{w}$, contradicting the assumed minimality of yy'. We conclude that no such yy' occurs in $\mathbf{w}$, and this completes the argument that $\mathbf{w}$ is 3/4-similar. □

Next, we consider the case $k = 4$.

Theorem 2. *There exists an infinite* 1/2*-similar word* $\mathbf{x}$ *over* $\{\mathtt{0}, \mathtt{1}, \mathtt{2}, \mathtt{3}\}$.

Proof. Let g be the 36-uniform morphism defined by

$$\begin{aligned}
\mathtt{0} &\to \mathtt{012132303202321020123021203020121310}\\
\mathtt{1} &\to \mathtt{123203010313032131230132310131232021}\\
\mathtt{2} &\to \mathtt{230310121020103202301203021202303132}\\
\mathtt{3} &\to \mathtt{301021232131210313012310132313010203}.
\end{aligned}$$

Then $\mathbf{x} = g^\omega(\mathtt{0})$ has the desired property. The proof is entirely analogous to that of Theorem 1 and is omitted. □

In our last result of this section, we show that we can obtain infinite words of arbitrarily low similarity, provided the alphabet size is sufficiently large. The main tool is the following [1, Lemma 5.1.1]:

Lemma 2 (Lovász Local Lemma; asymmetric version). *Let I be a finite set, and let $\{A_i\}_{i\in I}$ be events in a probability space. Let E be a set of pairs $(i,j) \in I \times I$ such that A_i is mutually independent of all the events $\{A_j : (i,j) \notin E\}$. Suppose there exist real numbers $\{x_i\}_{i\in I}$, $0 \leq x_i < 1$, such that for all $i \in I$,*

$$\mathrm{Prob}(A_i) \leq x_i \prod_{(i,j)\in E} (1 - x_j).$$

Then

$$\mathrm{Prob}\left(\bigcap_{i\in I} \overline{A_i}\right) \geq \prod_{i\in I}(1 - x_i) > 0.$$

We now state our result.

Theorem 3. *Let $c > 1$ be an integer. There exists an infinite $1/c$-similar word.*

Proof. Let k and N be positive integers, and let $w = w_1 w_2 \cdots w_N$ be a random word of length N over a k-letter alphabet Σ. Here each letter of w is chosen uniformly and independently at random from Σ.

Let

$$I = \{(t, r) : 0 \le t < N, 1 \le r \le \lfloor (N - t)/2 \rfloor\}.$$

For $i = (t, r) \in I$, write $y = w_t \cdots w_{t+r-1}$ and $y' = w_{t+r} \cdots w_{t+2r-1}$. Let A_i denote the event $s(y, y') > 1/c$. A crude overestimate of $\mathrm{Prob}(A_i)$ is

$$\begin{aligned}\mathrm{Prob}(A_i) &\le \frac{\binom{r}{\lceil r/c \rceil + 1} k^{\lceil r/c \rceil + 1} k^{2r - 2(\lceil r/c \rceil + 1)}}{k^{2r}} \\ &\le \binom{r}{\lceil r/2 \rceil} k^{-r/c} \le 2^r k^{-r/c},\end{aligned}$$

where the last inequality comes from Stirling's approximation.

For all positive integers r, define $\xi_r = 2^{-2r}$. For any real number $\alpha \le 1/2$, we have $(1 - \alpha) \ge e^{-2\alpha}$. Hence, $(1 - \xi_r) \ge e^{-2\xi_r}$. For $i = (t, r) \in I$, define $x_i = \xi_r$. Let E be as in the local lemma. Note that a subword of length $2r$ of w overlaps with at most $2r + 2s - 1$ subwords of length $2s$. Thus, for all $i = (t, r) \in I$, we have

$$\begin{aligned}x_i \prod_{(i,j) \in E} (1 - x_j) &\ge \xi_r \prod_{s=1}^{\lfloor N/2 \rfloor} (1 - \xi_s)^{2r+2s-1} \ge \xi_r \prod_{s=1}^{\infty} (1 - \xi_s)^{2r+2s-1} \\ &\ge \xi_r \prod_{s=1}^{\infty} e^{-2\xi_s (2r+2s-1)} \ge 2^{-2r} \prod_{s=1}^{\infty} e^{-2(2^{-2s})(2r+2s-1)} \\ &\ge 2^{-2r} \exp\left[-2\left(2r \sum_{s=1}^{\infty} \frac{1}{2^{2s}} + \sum_{s=1}^{\infty} \frac{2s-1}{2^{2s}}\right)\right] \\ &\ge 2^{-2r} \exp\left[-2\left(2r\left(\frac{1}{3}\right) + \frac{5}{9}\right)\right] \\ &\ge 2^{-2r} \exp\left(-\frac{4}{3}r - \frac{10}{9}\right).\end{aligned}$$

The hypotheses of the local lemma are met if $2^r k^{-r/c} \le 2^{-2r} \exp\left(-\frac{4}{3}r - \frac{10}{9}\right)$. Taking logarithms, we require $r \log 2 - \frac{r}{c} \log k \le -2r \log 2 - \frac{4}{3}r - \frac{10}{9}$. Rearranging terms, we require $c\left(3 \log 2 + \frac{4}{3} + \frac{10}{9r}\right) \le \log k$. The left side of this inequality is largest when $r = 1$, so we define $d_1 = 3 \log 2 + \frac{4}{3} + \frac{10}{9}$, and insist that $c \cdot d_1 \le \log k$. Hence, for $k \ge e^{c \cdot d_1}$, we may apply the local lemma to conclude that with positive probability, w is $1/c$-similar. Since $N = |w|$ is arbitrary, we conclude that there are arbitrarily large such w. By König's Infinity Lemma, there exists an infinite $1/c$-similar word, as required. □

4 Words avoiding c-approximate squares

In this section we consider the "additive" version of the problem. Table 2 reflects our results using a backtracking algorithm: there is no infinite word over a k-letter alphabet that avoids c-approximate squares, for the k and c given below.

Alphabet Size k	c	Height of Tree	Number of Leaves	Number of Maximal Words	Lexicographically First Maximal Words
2	0	4	3	1	010
3	1	5	23	2	01201
4	2	7	184	6	0123012
5	2	11	3253	24	01234102314
6	3	11	35756	960	01234051230
7	4	13	573019	6480	0123450612340
8	5	15	-	-	012345607123450

Table 2. Lower bounds on avoiding c-approximate squares

Theorem 4. *There is an infinite word over a* 3*-letter alphabet that avoids* 0*-approximate squares, and the* 0 *is best possible.*

Proof. Any ternary word avoiding squares, such as the fixed point, starting with 2, of $2 \to 210$, $1 \to 20$, $0 \to 1$, satisfies the conditions of the theorem. The result is best possible, from Table 2. □

Theorem 5. *There is an infinite word over a* 4*-letter alphabet that avoids* 1*-approximate squares, and the* 1 *is best possible.*

Proof. Let $\mathbf{c}$ be any squarefree word over $\{0, 1, 2\}$, and consider the image under the 48-uniform morphism γ defined by

$$0 \to 012031023120321031201321032013021320123013203123$$
$$1 \to 012031023120321023103213021032013210312013203123$$
$$2 \to 012031023012310213023103210231203210312013203123$$

The resulting word $\mathbf{d} = \gamma(\mathbf{c})$ avoids 1-approximate squares. The result is best possible, from Table 2.

The proof is similar to that of Theorem 1. Suppose to the contrary that $\mathbf{d}$ contains a 1-approximate square yy', $|y| = |y'|$. We may verify computationally that $\mathbf{d}$ contains no such subword yy' where $|y| \leq 96$. We therefore assume from now on that $|y| > 96$.

Let $w = a_1 a_2 \cdots a_n$ be a word of minimal length such that $\gamma(w) = xyy'z$ for some $x, z \in \{0, 1, 2, 3\}^*$. By the minimality of w, we have $0 \leq |x|, |z| < 48$.

For $i = 1, 2, \ldots, n$, define $A_i = \gamma(a_i)$. Just as in the proof of Theorem 1, we write

$$\gamma(w) = A_1 A_2 \cdots A_n = A'_1 A''_1 A_2 \cdots A_{j-1} A'_j A''_j A_{j+1} \cdots A_{n-1} A'_n A''_n,$$

so that the situation illustrated in Figure 1 applies to $xyy'z$ within $\gamma(w)$. We now make the following observations regarding the morphism γ:

1. Let $a, b, c \in \{\mathtt{0}, \mathtt{1}, \mathtt{2}\}$, $a \neq b$. Let u be any subword of length 48 of $\gamma(ab)$. If u is neither a prefix nor a suffix of $\gamma(ab)$, then $\gamma(c)$ and u mismatch in at least 18 positions.
2. Let $a, b \in \{\mathtt{0}, \mathtt{1}, \mathtt{2}\}$, $a \neq b$. Then $\gamma(a)$ and $\gamma(b)$ mismatch in at least 18 positions.
3. Let u, u', v, v' be words satisfying the following:
   - $|u| = |u'|$, $|v| = |v'|$, and $|uv| = |u'v'| = 48$;
   - each of u and u' is a suffix of a word in $\{\gamma(\mathtt{0}), \gamma(\mathtt{1}), \gamma(\mathtt{2})\}$; and
   - each of v and v' is a prefix of a word in $\{\gamma(\mathtt{0}), \gamma(\mathtt{1}), \gamma(\mathtt{2})\}$.

   Then either $uv = u'v'$ or uv and $u'v'$ mismatch in at least 18 positions.
4. Let $a \in \{\mathtt{0}, \mathtt{1}, \mathtt{2}\}$. Then $\gamma(a)$ is uniquely determined by either its prefix of length 17 or its suffix of length 17.

From the first observation, we deduce, as in the proof of Theorem 1, that the cases illustrated by Figures 2 and 3 cannot occur. In particular, we have that $|A''_1| = |A''_j|$ and $|A'_j| = |A'_n|$.

From the second observation, we deduce that for $i = 2, 3, \ldots, j-1$, $A_i = A_{i+j-1}$, and consequently, $a_i = a_{i+j-1}$.

From the third observation, we deduce that $A''_1 = A''_j$ and $A'_j = A'_n$.

From the fourth observation, we deduce that either $A_1 = A_j$ or $A_j = A_n$. If $A_1 = A_j$, then $a_1 = a_j$; if $A_j = A_n$, then $a_j = a_n$. In the first case, $a_1 a_2 \cdots a_{j-1} a_j a_{j+1} \cdots a_{n-1}$ is a square in $\mathbf{c}$, contrary to our assumption. In the second case, $a_2 a_3 \cdots a_j a_{j+1} a_{j+2} \cdots a_n$ is a square in $\mathbf{c}$, contrary to our assumption.

We conclude that $\mathbf{d}$ contains no 1-approximate square yy', as required. □

Theorem 6. *There is an infinite word over a* 6*-letter alphabet that avoids* 2*-approximate squares, and the* 2 *is best possible.*

Proof. Let $\mathbf{c}$ be any squarefree word over $\{\mathtt{0}, \mathtt{1}, \mathtt{2}\}$, and consider the image under the 6-uniform morphism β defined by

$$\mathtt{0} \to \mathtt{012345}; \quad \mathtt{1} \to \mathtt{012453}; \quad \mathtt{2} \to \mathtt{012534}.$$

The resulting word avoids 2-approximate squares. The result is best possible, from Table 2.

The proof is similar to that of Theorem 5, so we only note the properties of the morphism β needed to derive the result:

1. Let $a, b, c \in \{\mathtt{0}, \mathtt{1}, \mathtt{2}\}$, $a \neq b$. Let u be any subword of length 6 of $\beta(ab)$. If u is neither a prefix nor a suffix of $\beta(ab)$, then $\beta(c)$ and u mismatch in at least 3 positions.

2. Let $a, b \in \{0, 1, 2\}$, $a \neq b$. Then $\beta(a)$ and $\beta(b)$ mismatch in at least 3 positions.
3. Let u, u', v, v' be words satisfying the following:
   - $|u| = |u'|$, $|v| = |v'|$, and $|uv| = |u'v'| = 6$;
   - each of u and u' is a suffix of a word in $\{\beta(0), \beta(1), \beta(2)\}$; and
   - each of v and v' is a prefix of a word in $\{\beta(0), \beta(1), \beta(2)\}$.

   Then either $uv = u'v'$ or uv and $u'v'$ mismatch in at least 3 positions.
4. Let $a \in \{0, 1, 2\}$. Then $\beta(a)$ is uniquely determined by either its prefix of length 4 or its suffix of length 1.

□

Further results on additive similarity are summarized in the next theorem.

Theorem 7. *For each k, n, d given below, there is an infinite word over a k-letter alphabet that avoids n-approximate squares, and in each case such an infinite word can be generated by applying the given d-uniform morphism to any infinite squarefree word over $\{0, 1, 2\}$. (Note that we have used the coding $A = 10$, $B = 11$, etc.)*

k	n	d	Morphism
7	3	14	0 → 01234056132465
			1 → 01234065214356
			2 → 01234510624356
8	4	16	0 → 0123456071326547
			1 → 0123456072154367
			2 → 0123456710324765
9	5	36	0 → 012345607821345062718345670281346578
			1 → 012345607182346750812347685102346578
			2 → 012345607182346510872345681702346578
11	6	20	0 → 012345670A812954768A
			1 → 0123456709A1843576A9
			2 → 01234567089A24365798
12	7	24	0 → 012345678091AB2354687A9B
			1 → 012345678091A3B4257689AB
			2 → 012345678091A2B3465798AB
13	8	26	0 → 01234567890A1BC24635798BAC
			1 → 01234567890A1B3C4257689ABC
			2 → 01234567890A1B2C354687A9BC
14	9	28	0 → 0123456789A0B1DC32465798BDAC
			1 → 0123456789A0B1DC243576A98DBC
			2 → 0123456789A0B1CD325468A79CBD
15	10	30	0 → 0123456789AB0D1CE3246579B8ACDE
			1 → 0123456789AB0D1CE2435768A9DCBE
			2 → 0123456789AB0CED32154687BA9DEC

In each case, the proof is similar to the ones given previously, and is omitted.

Theorem 8. *For all integers $n \geq 3$, there is an infinite word over an alphabet of $2n$ letters that avoids $(n-1)$-approximate squares.*

Proof. Consider the $2n$-uniform morphism $h : \Sigma_3^* \to \Sigma_{2n}^*$ defined as follows:

$$\begin{aligned}
0 &\to 012\cdots(n-1)n\cdots(2n-1)\\
1 &\to 012\cdots(n-1)(n+1)(n+2)\cdots(2n-1)n\\
2 &\to 012\cdots(n-1)(n+2)(n+3)\cdots(2n-1)n(n+1)
\end{aligned}$$

We claim that if $\mathbf{w}$ is any squarefree word over Σ_3, then $h(\mathbf{w})$ has the desired properties. The proof is a simple generalization of Theorem 6. □

5 Another variation

Yet another variation we can study is trying to avoid xx' where x is very similar to x', but only for sufficiently large x. Let us say that a finite word is (n, α)-similar if $\alpha = \sup_{\substack{x,x' \text{subwords of } z \\ |x|=|x'|\geq n}} s(x, x')$, and analogous definitions for infinite z.

Exercise 5.8.1 of Alon and Spencer [1] asks the reader to show, using the Lovász local lemma that, (in our language) for every $\epsilon > 0$, there exists an infinite binary word $\mathbf{z}$ and an integer c such that $\mathbf{z}$ is (c, α)-similar for some $\alpha \leq \frac{1}{2} + \epsilon$.

Theorem 9. $\frac{1}{2}$ *is best possible in the previous result.*

Proof. Suppose $\frac{1}{2}$ is not best possible. Then there exists an infinite binary word $\mathbf{z}$ and a positive integer c, such that any subword xx' of $\mathbf{z}$ with $|x| = |x'| \geq c$ satisfies $s(x, x') < \frac{1}{2}$. Consider a subword of $\mathbf{z}$ of the form $xx'yy'$, with $|x| = |x'| = |y| = |y'| = c$. By our assumption, $s(x, x') < \frac{1}{2}$ and $s(x', y) < \frac{1}{2}$; hence, since $\mathbf{z}$ is defined over a binary alphabet, necessarily $s(x, y) > \frac{1}{2}$. Similarly, we must have $s(x', y') > \frac{1}{2}$. But then by definition of s,

$$2c \cdot s(xx', yy') = c \cdot s(x, y) + c \cdot s(x', y') > \frac{c}{2} + \frac{c}{2} = c,$$

and so $s(xx', yy') > \frac{1}{2}$, a contradiction to our assumption. □

6 Edit distance

There are many definitions of *edit distance*, but for our purposes, we say the edit distance $e(x, y) = c$ if x can be transformed into y by a sequence of c insertions, deletions, or replacements, and no sequence of $c - 1$ insertions, deletions, or replacements suffices.

We can expand our notion of approximate square to avoid all words that are within edit distance c of all squares.

For example, consider the case $c = 1$. Then every word of length 1, say a, is within edit distance 1 of a square, as we can simply insert a to get aa. Similarly, every word of length 2, say ab, is within edit distance 1 of a square, as we can simply replace the b by a to get aa. Thus we need to restrict our attention to avoiding words that are within edit distance c of all *sufficiently large* squares.

Theorem 10. *There is an infinite word over* 5 *letters such that all subwords* x *with* $|x| \geq 3$ *are neither squares, nor within edit distance* 1 *of any square. There is no such word over* 4 *letters.*

Proof. The usual tree traversal technique shows there is no such word over 4 letters Over 5 letters we can use the 5-uniform morphism h defined by

$$0 \to 01234; \quad 1 \to 02142; \quad 2 \to 03143.$$

We claim the image of every square-free word under h has the desired property. Details will appear in the final paper. □

References

1. N. Alon and J. Spencer. *The Probabilistic Method*. 2nd edition. Wiley, 2000.
2. J. Berstel. *Axel Thue's Papers on Repetitions in Words: a Translation*. Number 20 in Publications du Laboratoire de Combinatoire et d'Informatique Mathématique. Université du Québec à Montréal, February 1995.
3. F.-J. Brandenburg. Uniformly growing k-th power-free homomorphisms. *Theoret. Comput. Sci.* **23** (1983), 69–82.
4. A. Carpi. On the repetition threshold for large alphabets. In R. Královič and P. Urzyczyn, eds., *Proc. MFCS 2006*, Lect. Notes in Comput. Sci. #3162, Springer-Verlag, 2006, pp. 226–237.
5. F. Dejean, Sur un théorème de Thue, *J. Comb. Theory. Ser. A* **13** (1972), 90–99.
6. R. C. Entringer, D. E. Jackson, and J. A. Schatz. On nonrepetitive sequences. *J. Combin. Theory. Ser. A* **16** (1974), 159–164.
7. G. M. Landau and J. P. Schmidt. An algorithm for approximate tandem repeats. In A. Apostolico, M. Crochemore, Z. Galil, and U. Manber, eds., *Combinatorial Pattern Matching, 4th Annual Symposium, CPM 93*. Lect. Notes in Comput. Sci. # 684, Springer-Verlag, 1993, pp. 120–133.
8. R. Kolpakov and G. Kucherov. Finding approximate repetitions under Hamming distance. *Theor. Comput. Sci.* **303** (2003), 135–156.
9. M. Mohammad-Noori and J. D. Currie. Dejean's conjecture and Sturmian words. *European J. Combin.* **28** (2007), 876–890.
10. J. Moulin-Ollagnier. Proof of Dejean's conjecture for alphabets with $5, 6, 7, 8, 9, 10$ and 11 letters. *Theoret. Comput. Sci.* **95** (1992), 187–205.
11. T. Nagell (ed.). *Selected Mathematical Papers of Axel Thue*. Universitetsforlaget, Oslo, 1977.
12. J.-J. Pansiot. A propos d'une conjecture de F. Dejean sur les répétitions dans les mots. *Discrete Appl. Math.* **7** (1984), 297–311.
13. A. Thue. Über unendliche Zeichenreihen. *Norske vid. Selsk. Skr. Mat. Nat. Kl.* **7** (1906), 1–22. Reprinted in [11, pp. 139–158].
14. A. Thue. Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. *Norske vid. Selsk. Skr. Mat. Nat. Kl.* **1** (1912), 1–67. Reprinted in [11, pp. 413–478].