Avoidability in Words: New Results and Open Problems

Jeffrey Shallit
School of Computer Science
University of Waterloo
Waterloo, Ontario N2L 3G1
Canada
shallit@graceland.uwaterloo.ca
http://www.cs.uwaterloo.ca/~shallit

This talk represents joint work with N. Rampersad, M.-w. Wang, J. Karhumäki, L. Ilie, and P. Ochem.

<u>Squares</u>

- This talk is about words (strings of symbols) over a finite alphabet.
- A nonempty word is called a **<u>square</u>** if it is of the form xx, where x is a word.
- For example, here are some squares in English:
  - `atlatl`
  - `murmur`
  - `tartar`
  - `beriberi`
  - `hotshots`
- A word is **<u>squarefree</u>** if it contains no square subwords. (A subword is a block of contiguous symbols inside another word.)
- It is easy to see that every word of length ≥ 4 over the alphabet $\Sigma = \{0, 1\}$ contains a square.

Squarefree Words

- Are there arbitrarily large squarefree words over an alphabet of size 3?
- The Norwegian mathematician Axel Thue proved in 1906 that there are arbitrarily large squarefree words (and hence infinite squarefree words) over an alphabet of size 3.
- One such word begins

  $2102012101202102012021012102012101 20 \cdots$

- This word is the **fixed point** of the **morphism** g, which sends $2 \to 210$, $1 \to 20$, and $0 \to 1$.
- His construction was rediscovered many times, for example, by Marston Morse in 1921 and by the Dutch chess master Max Euwe in 1929.

Cubes and Overlaps

- A nonempty word is called a **cube** if it is of the form xxx, where x is a word.
- The English sort-of-word `shshsh` is a cube, as is the Finnish word `kokoko`.

- A word is an **overlap** if it is of the form $axaxa$, where a is a single letter and x is a (possibly empty) word.
- The English words
  - `alfalfa`
  - `entente`
  - `kinnikinnik`

  are overlaps. You can think of an overlap as a "$2 + \epsilon$" or just "2^+" power, since it is just slightly larger than a square.

Overlap-free Words

- Thue also proved that there exists an infinite word over a 2-letter alphabet that avoids overlaps (and hence cubes).
- His example begins

  0110100110010110100101100110100110 $\cdots$

  and is now known as the Thue-Morse word.
- It is the fixed point of the morphism μ, which sends $0 \to 01$ and $1 \to 10$.

<u>Fractional Powers</u>

- The generalization to higher powers of words should be clear
- How about rational powers?
- We say a word w is an e'th power (e rational) if there exist words $y, y' \in \Sigma^*$ such that $w = y^n y'$ and y' is a prefix of y with
$$e = n + \frac{|y'|}{|y|}.$$
- Examples:
  - `tormentor` is a $\frac{3}{2}$-power.
  - `educated` is a $\frac{4}{3}$-power.
  - `onion` is a $\frac{5}{3}$-power.
- A word **<u>avoids e'th powers</u>** if it contains no subwords that are e' powers for $e' \geq e$.
- A word **<u>avoids e^+'th powers</u>** if it contains no subwords that are e' powers for $e' > e$.

Enumerating Words Avoiding Patterns

- We can count the number of words of length n avoiding squares, overlaps, cubes, etc.
- Enumeration is hard! Gaps between upper and lower bounds are often frustratingly large
- For the squarefree words over a 3-letter alphabet, it is known that

$$1.118419^n < s(n) < 1.302128^n$$

  for all sufficiently large n.
- Enumeration is hard for the squarefree words because we don't understand their structure very well.

Enumerating Words Avoiding Patterns

Open Problem 1 *Find a simple description for the lexicographically least squarefree word*

$$0102012021012010201 2\cdots$$

over $\{0,1,2\}$*.*

- By contrast, we do understand the structure of overlap-free words over a 2-letter alphabet very well.
- For example, the lexicographically least overlap-free word over $\{0,1\}$ is known to be

  $001001\,\overline{\mathbf{t}} = 001001100101100110100 1\cdots$

Enumerating Words Avoiding Patterns

- Restivo and Salemi proved that there are only polynomially many overlap-free binary words of length n.
- The current best upper bound is $O(n^{1.37})$, due to Lepistö
- On the other hand, the gap for the sequence $(u_n)_{n\geq 0}$ counting the number of overlap-free words over $\{0, 1\}$ is more fundamental, as Cassaigne proved in 1993 that

$$\sup\{r \; : u_n = \Omega(n^r)\} < \inf\{s \; : u_n = O(n^s)\}.$$

Enumerating Words Avoiding Patterns

- Restivo and Salemi proved that there are only polynomially many overlap-free binary words of length n.
- Brandenburg proved there are exponentially many cubefree binary words of length n.
- Open question of Kobayashi (1986):

> At what exponent $2 < e < 3$ does the number of words avoiding e'th powers jump from polynomial to exponential?

- The surprising answer is $e = 7/3$. There are only polynomially many words of length n avoiding $\frac{7}{3}$ powers, but exponentially many avoiding $\frac{7}{3}^{+}$ powers.

The Upper Bound

Decomposition Theorem. Let x be a word avoiding α-powers, with $2 < \alpha \leq 7/3$. Let μ be the Thue-Morse morphism, sending $0 \to 01$, $1 \to 10$. Then there exist binary words u, v, y with

$$u, v \in \{\epsilon, 0, 1, 00, 11\}$$

such that $x = u\mu(y)v$.

Corollary. Let $2 < \alpha \leq \frac{7}{3}$. There are $O(n^{\log_2 25}) = O(n^{4.644})$ binary words of length n that avoid α-powers.

Proof. Let $x = x_0$ be a nonempty binary word that is α-power-free, with $2 < \alpha \leq \frac{7}{3}$. Then by the decomposition theorem we can write

$$x_0 = u_1 \mu(x_1) v_1$$

with $|u_1|, |v_1| \leq 2$. If $|x_1| \geq 1$, we can repeat the process, writing

$$x_1 = u_2 \mu(x_2) v_2.$$

The Upper Bound

Continuing in this fashion, we obtain the decomposition

$$x_i = u_i \mu(x_i) v_i$$

until $|x_{t+1}| = 0$ for some t. Then

$$\begin{aligned} x_0 \;=\; & u_1 \mu(u_2) \cdots \mu^{t-1}(u_{t-1}) \mu^t(x_t) \\ & \mu^{t-1}(v_{t-1}) \cdots \mu(v_2) v_1. \end{aligned}$$

Then from the inequalities

$$1 \leq |x_t| \leq 4$$

and

$$2|x_i| \leq |x_{i-1}| \leq 2|x_i| + 4,$$

for $1 \leq i \leq t$, an easy induction gives

$$2^t \leq |x| \leq 2^{t+3} - 4.$$

Thus $t \leq \log_2 |x| < t + 3$, and so

$$\log_2 |x| - 3 < t \leq \log_2 |x|. \qquad (1)$$

The Upper Bound

- There are at most 5 possibilities for each u_i and v_i, and there are at most 22 possibilities for x_t (since $1 \leq |x_t| \leq 4$ and x_t is α-power-free).
- Inequality (1) shows there are at most 3 possibilities for t.
- Letting $n = |x|$, we see there are at most $3 \cdot 22 \cdot 5^{2 \log_2 n} = 66 n^{\log_2 25}$ words of length n that avoid α-powers.

Arbitrarily Large Squares

Theorem. Every infinite $\frac{7}{3}$-power-free binary word contains arbitrarily large squares.

Proof.

- Let $\mathbf{w}$ be an infinite $\frac{7}{3}$-power-free binary word.
- By the decomposition theorem and Eq. (1), any prefix of $\mathbf{w}$ of length 2^{n+5} contains $\mu^{n+2}(0)$ as a factor.
- But $\mu^{n+2}(0) = \mu^{n}(0110)$, so any prefix of length 2^{n+5} contains the square factor xx with $x = \mu^{n}(1)$.

The Magic Number $7/3$

- Narad Rampersad has also found another related property of $\frac{7}{3}$:
- The Thue-Morse word $\mathbf{t}$ and its complement $\overline{\mathbf{t}}$ are the only infinite binary words avoiding $\frac{7}{3}$-powers that are fixed points of a non-trivial morphism.
- This improves a 1982 theorem due to Séébold.
- Once again the number $\frac{7}{3}$ is best possible, since the morphism sending

$$0 \rightarrow \mathtt{0110100110110010110}$$
$$1 \rightarrow \mathtt{1001011001001101001}$$

has a fixed point that is $\frac{7}{3}^{+}$-power-free.

Avoiding Large Squares

- As we have seen, it is impossible for infinite binary words to avoid all squares
- But is it possible to avoid arbitrarily large squares?
- Yes! Entringer, Jackson, and Schatz proved in 1974 that there exists an infinite binary word containing no squares yy with $|y| \geq 3$.
- Their strategy: start with any word over three letters that avoids squares, such as

  $210201210120210201202101210201210120 \cdots$

  Replace each letter by applying the morphism h, as follows:

  $$\begin{aligned} 0 &\rightarrow 1010 \\ 1 &\rightarrow 1100 \\ 2 &\rightarrow 0111 \end{aligned}$$

  The resulting word

  $$\mathbf{w} = 011111001010011110101100 \cdots$$

  has the desired properties.

Avoiding Large Squares

- The proof is rather technical, but here are the basic ideas.
- We divide the proof into two cases: $\mathbf{w}$ avoids small squares xx, with $3 \leq |x| \leq 8$, and $\mathbf{w}$ avoids large squares, $|x| > 8$.
- To see that it avoids small squares, it suffices to check the image of all squarefree strings of length ≤ 5.
- To see that it avoids large squares, we argue by contradiction. It suffices to check certain properties of the morphism.

Avoiding Large Squares

- The fact that 3 is best possible can be proved purely mechanically.
- Given a set of forbidden patterns P, we create a tree T as follows:
  - The root of T is labeled ϵ (the empty string).
  - If a node is labeled w and avoids P, then it is an internal node with two children, where the left child is labeled $w0$ and the right child is labeled $w1$.
  - If it does not avoid P, then it is an external node (or "leaf").
- No infinite word avoiding P exists if and only if T is finite.
- Breadth-first search can be used to verify that T is finite.

Avoiding Large Squares

- Furthermore, certain parameters of T correspond to information about the finite words avoiding P:
  - the number of leaves n is one more than the number of internal nodes, and so $n-1$ represents the total number of finite words avoiding P;
  - if the height of the tree (i.e., the length of the longest path from the root to a leaf) is h, then h is the smallest integer such that there are no words of length $\geq h$ avoiding P;
  - the internal nodes at depth $h-1$ gives the all words of maximal length avoiding P;
- In the case of Entringer-Jackson-Schatz, let P be the set of all squares of length ≥ 3.
- The resulting tree is finite. It has height 19, and contains 478 leaves. The longest label is 010011000111001101 and its complement.

Avoiding Both Powers and Large Squares

- Dekking considered avoiding both cubes and large squares over $\{0, 1\}$.
- He proved that there exists an infinite binary word avoiding both cubes xxx and squares yy with $|y| \geq 4$.
- Furthermore, the bound 4 is best possible.
- This suggests the following natural problem. For each length $l \geq 1$, determine the fractional exponent e such that
  - There is no infinite binary word simultaneously avoiding squares yy with $|y| \geq l$ and e'th powers
  - There is an infinite binary word simultaneously avoiding squares yy with $|y| \geq l$ and e^+'th powers

Avoiding Both Powers and Large Squares

Summary of Results

minimum length l of square avoided	avoidable power	unavoidable power
2	none	all
3	3^+	3
$4, 5, 6$	$(5/2)^+$	$5/2$
≥ 7	$(7/3)^+$	$7/3$

- The unavoidability results are proved using the tree-traversal technique
- The avoidability results are proved using a strategy similar to Entringer-Jackson-Schatz: we start with a word over $\{0, 1, 2\}$ avoiding squares, and then replace each symbol by an appropriately-chosen binary string.

Avoiding Both Powers and Large Squares

- To show there is an infinite binary word avoiding 3^+ powers and squares yy with $|y| \geq 3$, we use the map

$$\begin{aligned} 0 &\rightarrow 0010111010 \\ 1 &\rightarrow 0010101110 \\ 2 &\rightarrow 0011101010 \end{aligned}$$

- To show there is an infinite binary word avoiding $\frac{5}{2}^+$ powers and squares yy with $|y| \geq 4$, we use a map sending each letter to a string of 1560 letters. (!)
- To show there is an infinite binary word avoiding $\frac{7}{3}^+$ powers and squares yy with $|y| \geq 7$, we use a map sending each letter to a string of 252 letters.

Example of the Morphism for Length 7

0 $\to$ 001101001011001001101100101101001100101100100110110010110011010
010110010011011001011010011001011001101001101100100110100110010
110100110110010011010010110010011011001011010011001011001101001
101100100110100101100100110110010110011010011001011010011011001

1 $\to$ 001101001011001001101100101101001100101100100110110010110011010
011001011010011011001001101001011001101001101100100110100110010
110100110110010011010010110010011011001011010011001011001101001
101100100110100101100100110110010110011010011001011010011011001

2 $\to$ 001101001011001001101100101101001100101100110100110110010011010
011001011010011011001001101001011001001101100101101001100101100
100110110010110011010010110010011011001011010011001011001101001
101100100110100101100100110110010110011010011001011010011011001

How Were the Morphisms Found?

- How were these morphisms found?
- In the first case, we iteratively generated all words of length $1, 2, 3, \ldots$ (up to some bound) that avoid both 3^+ powers and squares yy with $|y| \geq 3$.
- We then guessed such words were the image of a k-uniform morphism applied to a squarefree word over $\{\mathtt{0}, \mathtt{1}, \mathtt{2}\}$.
- For values of $k = 2, 3, \ldots$, we broke up each word into contiguous blocks of size k, and discarded any word for which there were more than 3 blocks.
- For certain values of k, this procedure eventually resulted in 0 words fitting the criteria.
- At this point we knew a k-uniform morphism cannot work, so we increased k and started over.

How Were the Morphisms Found?

- Eventually a k was found for which the number of such words appeared to increase without bound.
- We then examined the possible sets of 3 k-blocks to see if any of them were suitable. This gave our candidate morphism.
- Pascal Ochem has found similar morphisms using a more automated approach. See his talk in this conference.

The Shuffle Problem

- Prodinger and Urbanek in 1983 studied the avoidance of arbitrarily large squares in binary words.
- They were unable to answer the following question: is there a pair of infinite binary words, avoiding arbitrarily large squares, such that their perfect shuffle contains arbitrarily large squares?
- Here, by the perfect shuffle of two infinite words $\mathbf{w} = a_0 a_1 a_2 \cdots$ and $\mathbf{x} = b_0 b_1 b_2 \cdots$ we mean the word

$$a_0 b_0 a_1 b_1 a_2 b_2 \cdots .$$

The Shuffle Problem

The answer is yes. Consider the morphism f defined by

$$0 \to 001$$
$$1 \to 110.$$

The fixed point

$$f^{\omega}(0) = 001001110001001110110110 \cdots$$

begins with arbitrarily large squares of the form $f^n(0)f^n(0)$. It is the shuffle of two words

$$0\ 1\ 0\ 1\ 0\ 0\ \cdots$$

and

$$0\ 0\ 1\ 1\ 0\ 1 \cdots$$

each of which avoids squares xx with $|x| \geq 4$.

Generalized Repetition Threshold

- Brandenburg and Dejean considered the problem of determining the *repetition threshold*: the least exponent $\alpha = \alpha(k)$ such that there exist infinite words over a k-letter alphabet that avoid α^+-powers.
- Dejean proved that $\alpha(3) = \frac{7}{4}$.
- She also conjectured that $\alpha(4) = \frac{7}{5}$ and $\alpha(k) = \frac{k}{k-1}$ for $k \geq 5$.
- Pansiot proved that $\alpha(4) = \frac{7}{5}$
- Moulin-Ollagnier proved that Dejean's conjecture holds for $5 \leq k \leq 11$.
- Dejean's conjecture is still open.

Generalized Repetition Threshold

- With Ilie and Ochem, I generalized the repetition threshold of Dejean to handle avoidance of all *sufficiently large* fractional powers.
- Pansiot also suggested looking at this generalization at the end of his paper, but to the best of my knowledge no one else has pursued this question.
- We say that an infinite word is (α, ℓ)-free if it contains no powers x^β for $\beta \geq \alpha$ and $|x| \geq \ell$.
- The *generalized repetition threshold* $R(k, \ell)$ is defined to be the least α such that there exist infinite words over a k-letter alphabet that are (α, ℓ)-free.
- Thus $R(k, 1)$ is the repetition threshold of Dejean and Brandenburg.

- We were able to show that

$$R(3,2) = \frac{3}{2}^{+}$$
$$R(3,3) = \frac{4}{3}^{+}$$
$$R(2,4) = \frac{3}{2}^{+}$$

- However, many open problems remain.

Open Problem 2 *Is* $R(2,3) = \frac{8}{5}^{+}$*?*

Avoiding Reversed Factors

- Let us consider avoiding reversed factors, that is, creating infinite words where if w is a factor, then its reversal w^R is not.
- Evidently it is not possible to avoid *all* reversed factors, since factors of length 1 cannot be avoided.
- But we could consider avoiding all sufficiently large reversed factors.
- Over a 2-letter alphabet, we can avoid reversed factors of length ≥ 5, and the number 5 is best possible.
- All such infinite words are ultimately periodic, and they have a simple description.
- Over a 3-letter alphabet, we can avoid reversed factors of length ≥ 2, and evidently 2 is best possible.
- Again, all such infinite words are ultimately periodic.

- Alon, Grytczuk, Haluszczak, and Riordan proved that there exists an infinite squarefree word over a 4-letter alphabet that avoids palindromes x with $|x| \geq 2$.

- Over a 5-letter alphabet, however, there are infinite squarefree words with an even stronger property: they also avoid *all* reversed factors of length ≥ 2.

Avoidability and Repetition Complexity

Ilie, Yu, and Zhang (2002) defined the *repetition complexity* of a string w to be the smallest number of symbols needed to represent w using concatenation and exponentation where

- exponents are represented in *decimal*
- exponentiation and concatenation can be nested
- parentheses, concatenation symbol, and exponentiation symbol not counted

They showed the existence of a family of binary strings with

$$R(w) \geq c\frac{|w| \log \log |w|}{\log |w|}.$$

Avoidability and Repetition Complexity

We can prove

Theorem. There exist arbitrarily long words $w \in \{0,1\}^*$ with $R(w) \geq |w|/2$.

Proof. Apply the morphism

$$\begin{aligned} 0 &\rightarrow 1010 \\ 1 &\rightarrow 1100 \\ 2 &\rightarrow 0111 \end{aligned}$$

to the squarefree word that is a fixed point of

$$\begin{aligned} 2 &\rightarrow 210 \\ 1 &\rightarrow 20 \\ 0 &\rightarrow 1 \end{aligned}$$

Avoidability and Repetition Complexity

Theorem. For all words $w \in \{0,1\}^*$ we have

$$R(w) \leq 8 \left\lceil \frac{|w|}{9} \right\rceil .$$

Proof. Every binary word of length 9 contains either a square xx with $|x| \geq 2$, or 000, or 111. Each of these leads to a compression by at least 1 symbol.

Additional Open Problems

Open Problem 3 *Is there an infinite word over some finite subset of $\mathbb{Z}$ that avoids all subwords of the form ww' with*

$$|w| = |w'|$$

and

$$\sum w = \sum w' \ ?$$

Open Problem 4 *Is there an infinite word over $\{0, 1, 2\}$ avoiding ww' with w a permutation of w', for $|w| = |w'| \geq 2$?*

For Further Reading

1. R. C. Entringer, D. E. Jackson, and J. A. Schatz. On nonrepetitive sequences. *J. Combin. Theory. Ser. A* **16** (1974), 159–164.

2. F. M. Dekking. On repetitions of blocks in binary sequences. *J. Combin. Theory. Ser. A* **20** (1976), 292–299.

3. J. Karhumäki and J. Shallit. Polynomial versus exponential growth in repetition-free binary words. *J. Combinat. Theory Ser. A* **105** (2004), 335–347.

4. H. Prodinger and F. J. Urbanek. Infinite 0–1-sequences without long adjacent identical blocks. *Discrete Math.* **28** (1979), 277–289.

5. N. Rampersad, J. Shallit, and M.-w. Wang. Avoiding large squares in infinite binary words. Preprint available at `http://www.arxiv.org/abs/math.CO/0306081`.

6. J. Shallit. Simultaneous avoidance of large squares and fractional powers in infinite binary words. *Int. J. Found. Comput. Sci.* **15** (2004), 317–327.