New Results on Avoidability in Words

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

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

Squares

- This talk is about words (strings of symbols) over a finite alphabet.
- A nonempty word is called a **square** 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 **squarefree** 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

  $$01101001100101101001011001101001100\cdots$$

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

Fractional Powers

- 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 **avoids e'th powers** if it contains no subwords that are e' powers for $e' \geq e$.
- A word **avoids e^+'th powers** if it contains no subwords that are e' powers for $e' > e$.

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

  $$2102012101202102012021012102012101 20 \cdots$$

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

  $$0 \rightarrow 1010$$
  $$1 \rightarrow 1100$$
  $$2 \rightarrow 0111$$

  The resulting word

  $$\mathbf{w} = 0111110010100111101011 00 \cdots$$

  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 use the following lemma:

  Lemma.

  (a) (inclusion property) Suppose $h(ab) = th(c)u$ for some letters a, b, c and strings t, u. Then this inclusion is trivial (that is, $t = \epsilon$ or $u = \epsilon$).

  (b) (interchange property) Suppose there exist letters a, b, c and strings s, t, u, v such that $h(a) = st$, $h(b) = uv$, and $h(c) = sv$. Then either $a = c$ or $b = c$.

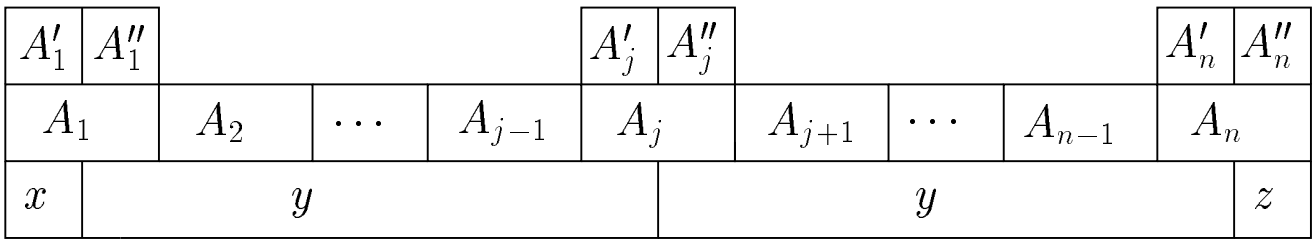
Avoiding Large Squares

- For $i = 1, 2, \ldots, n$ define $A_i = h(a_i)$.
- Assume $w = a_1 a_2 \cdots a_n$ is squarefree, but $h(w) = xyyz$.
- Write

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

where $|A''_1|, |A''_j| > 0$ and

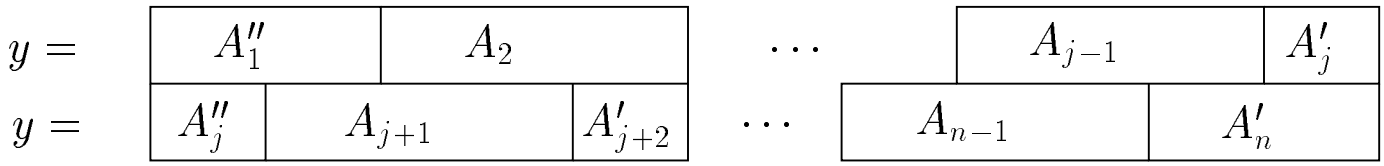
$$\begin{aligned} A_1 &= A'_1 A''_1 \\ A_j &= A'_j A''_j \\ A_n &= A'_n A''_n \\ x &= A'_1 \\ y &= A''_1 A_2 \cdots A_{j-1} A'_j = A''_j A_{j+1} \cdots A_{n-1} A'_n \\ z &= A''_n. \end{aligned}$$



Avoiding Large Squares

- If $|A_1''| > |A_j''|$, then $A_{j+1} = h(a_{j+1})$ is a subword of $A_1''A_2$, hence a subword of $A_1A_2 = h(a_1a_2)$.
- Thus we can write $A_{j+2} = A'_{j+2}A''_{j+2}$ with

$$A_1''A_2 = A_j''A_{j+1}A'_{j+2}.$$



- By the Lemma about inclusions, either $|A_j''| = 0$, or $|A_1''| = |A_j''|$. Both conclusions are impossible.
- Similar reasoning applies if $|A_1''| < |A_j''|$.

Avoiding Large Squares

- Therefore $|A_1''| = |A_j''|$.
- Hence $A_1'' = A_j''$, $A_2 = A_{j+1}$, ..., $A_{j-1} = A_{n-1}$, and $A_j' = A_n'$.
- Since h is injective, we have $a_2 = a_{j+1}, \ldots, a_{j-1} = a_{n-1}$.
- It also follows that $|y|$ is divisible by 4 and $A_j = A_j' A_j'' = A_n' A_1''$.
- But by the Lemma, either (1) $a_j = a_n$ or (2) $a_j = a_1$.
- In the first case, $a_2 \cdots a_{j-1} a_j = a_{j+1} \cdots a_{n-1} a_n$, so w contains the square $(a_2 \cdots a_{j-1} a_j)^2$, a contradiction.
- In the second case, $a_1 \cdots a_{j-1} = a_j a_{j+1} \cdots a_{n-1}$, so w contains the square $(a_1 \cdots a_{j-1})^2$, a contradiction.
- It now follows that if w is squarefree then $h(w)$ avoids squares yy with $|y| \geq 3$.

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 ≥ 2.
- The resulting tree is finite. It has height 19, and contains 478 leaves. The longest label is $0100110001110011 01$ 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

$$0 \rightarrow 0010111010$$
$$1 \rightarrow 0010101110$$
$$2 \rightarrow 0011101010$$

- 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

$0 \rightarrow$ 00110100101100100110110010110100110010110010011011001011001101
01011001001101100101101001100101100110100110110010011010011001
11010011011001001101001011001001101100101101001100101100110100
10110010011010010110010011011001011001101001100101101001101100

$1 \rightarrow$ 00110100101100100110110010110100110010110010011011001011001101
01100101101001101100100110100101100110100110110010011010011001
11010011011001001101001011001001101100101101001100101100110100
10110010011010010110010011011001011001101001100101101001101100

$2 \rightarrow$ 00110100101100100110110010110100110010110011010011011001001101
01100101101001101100100110100101100100110110010110100110010110
10011011001011001101001011001001101100101101001100101100110100
10110010011010010110010011011001011001101001100101101001101100

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 square-free word over $\{0, 1, 2\}$.
- 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 satisfied the requirements of the Lemma. This gave our candidate morphism.

Enumerating Words Avoiding Patterns

- We can also count the number of words of length n avoiding squares, overlaps, cubes, etc.
- Restivo and Salemi proved that there are only polynomially many overlap-free binary words of length n.
- The current best bound is $O(n^{1.37})$, due to Lepistö
- Brandenburg proved there are exponentially many cubefree binary words of length n.
- This raises the natural 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

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

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 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 $\mathbf{w} = a_0a_1a_2\cdots$ and $\mathbf{x} = b_0b_1b_2\cdots$ we mean the word

$$a_0b_0a_1b_1a_2b_2\cdots .$$

The Shuffle Problem

The answer is yes. Consider the morphism f defined by

$$0 \rightarrow 001$$
$$1 \rightarrow 110.$$

The fixed point

$$f^{\omega}(0) = 0010011100010011101101 10 \cdots$$

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 $|x| \geq 4$.

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. Preprint available at `http://www.arxiv.org/abs/math.CO/0304095`.

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. Preprint available at `http://www.arxiv.org/abs/math.CO/0304476`.