Repetitions in Words—Part I

Narad Rampersad

Department of Mathematics and Statistics
University of Winnipeg

Words

- words are sequences of letters taken from an alphabet
- words can be finite or infinite
- if Σ is an alphabet then Σ^* is the set of all finite words over Σ
- ϵ denotes the empty word

Factors

- w' is a factor of w if we can write $w = uw'v$ for some words u and v
- so `abbcb` is a factor of `bbabbcbabc`
- the notions of prefix/suffix should be clear

Infinite words

- suppose we have infinitely many finite words that have some interesting property
- we want to show that there is an infinite word with the same property
- König's Infinity Lemma: Let Σ be a finite alphabet, and let A be an infinite subset of Σ^*. Then there exists an infinite word $\mathbf{w}$ such that every prefix of $\mathbf{w}$ is a prefix of at least one word in A.

Repetitions

- a square is a non-empty word of the form xx (like `bonbon`)
- a word is squarefree if it contains no square as a factor
- a cube is a non-empty word xxx
- a k-power is a non-empty word x^k (x repeated k times)
- any long word over 2 symbols contains squares
- What if we use 3 symbols?

An infinite word avoiding squares

Theorem (Thue 1906)

There is an infinite squarefree word over 3 symbols.

Generating squarefree words

- iterate the map $0 \to 012$; $1 \to 02$; $2 \to 1$:

$$0 \to 012 \to 012021 \to 012021012102 \to \cdots$$

- these words are squarefree
- there is an infinite squarefree word

Morphisms

- the map $0 \to 012$; $1 \to 02$; $2 \to 1$ is a morphism
- (some say "substitution")
- a morphism $h : \Sigma^* \to \Delta^*$ is a map that satisfies $h(xy) = h(x)h(y)$
- if there is a letter a such that $h(a) = ax$ and h is non-erasing, then we can generate an infinite word by iterating h

The Thue–Morse word

- iterate the morphism $0 \to 01;\ 1 \to 10$:

$$0 \to 01 \to 0110 \to 01101001 \to 0110100110010110 \to \cdots$$

- the limit word is the Thue–Morse word $\mathbf{t}$.

Words avoiding overlaps

- Thue 1912: $\mathbf{t}$ contains no overlap.
- an overlap is a word of the form $axaxa$, where a is a single letter (like `entente`).
- a word is overlap-free if it contains no overlap as a factor
- binary overlap-free words have a lot of structure

Structure of binary overlap-free words

Theorem (Restivo and Salemi 1985)

Let μ be the Thue–Morse morphism: $0 \to 01,\ 1 \to 10$. Let $x \in \{0,1\}^*$ be overlap-free. There exist $u, v \in \{\epsilon, 0, 1, 00, 11\}$ and an overlap-free word y such that $x = u\mu(y)v$.

Example of the structure theorem

- 0011010011 is overlap-free
- $0011010011 = 0\mu(0110)1$
- 0110 is again overlap-free
- factorization can be iterated
- $0011010011 = 0\mu(\,\mu(01)\,)1$

A one-sided version

Theorem

Let $\mathbf{x} \in \{0,1\}^\omega$ be an overlap-free infinite word. Then there exist $u \in \{\epsilon, 0, 1, 00, 11\}$ and an overlap-free $\mathbf{y} \in \{0,1\}^\omega$ such that $\mathbf{x} = u\mu(\mathbf{y})$. Furthermore, u and the first two letters of $\mathbf{y}$ are completely determined by a prefix of length 4 of $\mathbf{x}$, unless $\mathbf{x}$ begins with 0010 or 1101, in which case a prefix of length 5 suffices.

An encoding of infinite overlap-free words

- Define $p_0 = \epsilon$, $p_1 = 0$, $p_2 = 00$, $p_3 = 1$, and $p_4 = 11$.
- Let $P = \{p_0, p_1, p_2, p_3, p_4\}$.
- Every infinite overlap-free word $\mathbf{x}$ can be written in the form

$$\mathbf{x} = p_{i_1}\mu(p_{i_2}\mu(p_{i_3}\mu(\cdots)))$$

- We can encode $\mathbf{x}$ by the sequence $i_1 i_2 i_3 \cdots$.

An encoding of infinite overlap-free words

- Some sequences $(i_j)_{j\geq 1}$ of indices do not correspond to an overlap-free word.
- $21\cdots$ represents $00\mu(0\mu(\ldots))$ which begins with the overlap 000.
- Which sequences give overlapfree words?

Some observations

- Let $\mathcal{O}$ denote the set of (right-) infinite binary overlap-free words.
- let $a \in \{0, 1\}$
- $\mathbf{x} \in \mathcal{O} \iff \mu(\mathbf{x}) \in \mathcal{O}$
- $a\,\mu(\mathbf{x}) \in \mathcal{O} \iff \overline{a}\mathbf{x} \in \mathcal{O}$
- $a\,a\,\mu(\mathbf{x}) \in \mathcal{O} \iff \overline{a}\mathbf{x} \in \mathcal{O}$ and $\mathbf{x}$ begins with $\overline{a}a\overline{a}$

Classifying overlap-free words

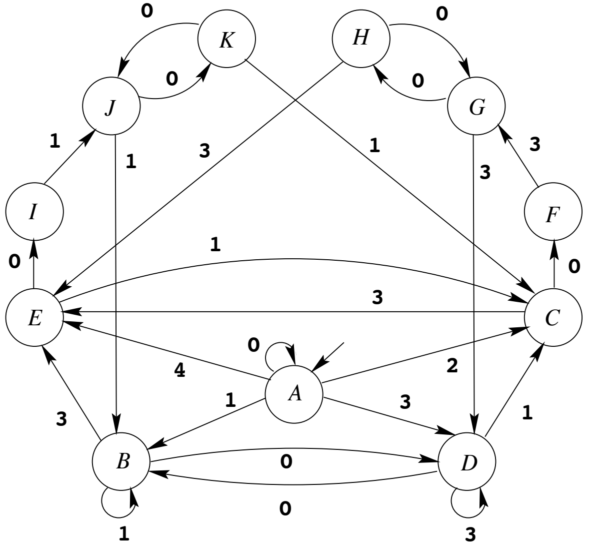
We now define 11 subsets of $\mathcal{O}$:

$$
\begin{aligned}
A &= \mathcal{O} \\
B &= \{\mathbf{x} \in \Sigma^\omega \ : \ 1\mathbf{x} \in \mathcal{O}\} \\
C &= \{\mathbf{x} \in \Sigma^\omega \ : \ 1\mathbf{x} \in \mathcal{O} \text{ and } \mathbf{x} \text{ begins with } 101\} \\
D &= \{\mathbf{x} \in \Sigma^\omega \ : \ 0\mathbf{x} \in \mathcal{O}\} \\
E &= \{\mathbf{x} \in \Sigma^\omega \ : \ 0\mathbf{x} \in \mathcal{O} \text{ and } \mathbf{x} \text{ begins with } 010\} \\
F &= \{\mathbf{x} \in \Sigma^\omega \ : \ 0\mathbf{x} \in \mathcal{O} \text{ and } \mathbf{x} \text{ begins with } 11\}
\end{aligned}
$$

Classifying overlap-free words

$$\begin{aligned}
G &= \{\mathbf{x} \in \Sigma^\omega \;:\; 0\mathbf{x} \in \mathcal{O} \text{ and } \mathbf{x} \text{ begins with } 1\} \\
H &= \{\mathbf{x} \in \Sigma^\omega \;:\; 1\mathbf{x} \in \mathcal{O} \text{ and } \mathbf{x} \text{ begins with } 1\} \\
I &= \{\mathbf{x} \in \Sigma^\omega \;:\; 1\mathbf{x} \in \mathcal{O} \text{ and } \mathbf{x} \text{ begins with } 00\} \\
J &= \{\mathbf{x} \in \Sigma^\omega \;:\; 1\mathbf{x} \in \mathcal{O} \text{ and } \mathbf{x} \text{ begins with } 0\} \\
K &= \{\mathbf{x} \in \Sigma^\omega \;:\; 0\mathbf{x} \in \mathcal{O} \text{ and } \mathbf{x} \text{ begins with } 0\}
\end{aligned}$$

Relationships between the classes



Valid encoding of overlap-free words

The sequence $(i_n)_{n \geq 1}$ of labels of any infinite path through this graph encodes an infinite overlap-free word

$$\mathbf{x} = p_{i_1}\mu(p_{i_2}\mu(p_{i_3}\mu(\cdots)))$$

Furthermore, all infinite overlap-free binary words can be so obtained.

Consequences of this characterization

- The lexicographically least overlap-free binary word is $001001\bar{\mathbf{t}}$, where $\bar{\mathbf{t}}$ is the complement of the Thue–Morse word.
- There are uncountably many infinite overlap-free binary words.
- The infinite overlap-free binary words are almost completely understood.
- not the case for squarefree ternary words

Enumeration of overlap-free words

- $x = x_0$ a nonempty overlap-free binary word
- write $x_0 = u_1\mu(x_1)v_1$ with $|u_1|, |v_1| \leq 2$
- iterate: $x_{i-1} = u_i\mu(x_i)v_i$ for $i = 1, 2, \ldots$
- until $|x_{t+1}| = 0$ for some t

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

Enumeration of overlap-free words

- $1 \le |x_t| \le 4$
- $2|x_i| \le |x_{i-1}| \le 2|x_i| + 4$, $1 \le i \le t$
- an easy induction gives $2^t \le |x| \le 2^{t+3} - 4$
- thus $t \le \log_2 |x| < t + 3$, and so

$$\log_2 |x| - 3 < t \le \log_2 |x|.$$

Enumeration of overlap-free words

- at most 5 possibilities for each u_i and v_i
- a most 22 possibilities for x_t (since $1 \leq |x_t| \leq 4$ and x_t is overlap-free)
- from the inequality $\log_2 |x| - 3 < t \leq \log_2 |x|$ there are at most 3 possibilities for t
- let $n = |x|$
- there are at most $3 \cdot 22 \cdot 5^{2 \log_2 n} = 66 n^{\log_2 25}$ overlap-free words of length n

Enumeration of overlap-free words

Theorem

There are $O(n^{\log_2 25}) = O(n^{4.644})$ binary words of length n that are overlap-free.

Best known results on overlap-free words

Theorem (Jungers, Protasov, and Blondel 2009)

There are constants C_1 and C_2 such that the number u_n of overlap-free words of length n over a binary alphabet satisfies

$$C_1 n^{\alpha} \leq u_n \leq C_2 n^{\beta},$$

where $1.2690 < \alpha < 1.2736$ and $1.3322 < \beta < 1.3326$.

Counting squarefree words

- How many squarefree words of length n do we have over a 3-letter alphabet?
- Consider the substitution h (Ekhad and Zeilberger 1998):

$$\begin{array}{lcl} 0 & \to & \{2102012021201020 12,\ 21020102120210 2012\} \\ 1 & \to & \{02101201020121012 0,\ 0210121020102101 20\} \\ 2 & \to & \{10212012101202120 1,\ 1021202101210212 01\} \end{array}$$

- $h(w)$ consists of squarefree words if w is squarefree.
- At least $2^{n/17} \approx 1.0416^n$ squarefree words of length n.

Best known results on squarefree words

Theorem (Shur)

There number of squarefree words of length n over a 3-letter alphabet grows like ρ^n, where $\rho \in [1.3017579, 1.3017619]$.

Observations regarding growth rates

- there are exponentially many ternary squarefree of words of length n
- there are only polynomially many binary overlapfree words
- note: exponentially many binary cubefree words
- polynomial growth relatively rare for classes of words defined by avoidance properties
- due to the highly structured nature of binary overlapfree words

The End