# Repetitions in Words—Part I

Narad Rampersad

Department of Mathematics and Statistics
University of Winnipeg

# Repetitions in words

- What kinds of repetitions can/cannot be avoided in words (sequences)?
- e.g., the word

$$ab\color{magenta}{aa}\color{black}{bbaba}\color{magenta}{abab}$$

  contains several repetitions
- but in the word

$$abcbacbcabcba$$

  the same sequence of symbols never repeats twice in succession

# Types of repetitions

- a square is a non-empty word of the form $xx$ (like `tauntaun`)
- a word is squarefree if it contains no square
- a cube is a non-empty word $xxx$
- a $t$-power is a non-empty word $x^t$ ($x$ repeated $t$ times)
- any long word over $2$ symbols contains squares
- Over $3$ symbols?

# Thue's work

## Theorem (Thue 1906)

There is an infinite squarefree word over $3$ symbols.

# Subsequent work

- Thue's result was rediscovered many times
- e.g., by Arshon (1937); Morse and Hedlund (1940)
- a systematic study of avoidable repetitions was begun by Bean, Ehrenfeucht, and McNulty (1979)

# Morphisms

- typical construction of squarefree words: find a map that produces a longer squarefree word from a shorter squarefree word
- e.g., the map (morphism) $f$ that sends $a \to abcab$; $b \to acabcb$; $c \to acbcacb$
- $f(acb) = abcab\,acbcacb\,acabcb$ is squarefree
- if this morphism preserves squarefreeness we can generate an infinite word by iteration

# Preserving squarefreeness

- What conditions on a morphism guarantee that it preserves squarefreeness?
- we say a morphism is infix if no image of a letter appears inside the image of another letter
- $a \to abc$; $b \to ac$; $c \to b$ is not infix

# A sufficient condition for infix morphisms

## Theorem (Thue 1912; Bean et. al. 1979)

Let $f : A^* \to B^*$ be a morphism from words over an alphabet $A$ to words over an alphabet $B$. If $f$ is infix and $f(x)$ is squarefree whenever $x$ is a squarefree word of length at most $3$, then $f$ preserves squarefreeness in general.

# Generating squarefree words

- the map $a \to abcab$; $b \to acabcb$; $c \to acbcacb$ satisfies the conditions of the theorem
- so it preserves squarefreeness
- if we iterate it we get squarefree words:

$$a \to abcab \to abcabacabcbacbcacbabcabacabcb$$

- so there is an infinite squarefree word

# A general criterion

**Theorem (Crochemore 1982)**

Let $f : A^* \to B^*$ be a morphism. Then $f$ preserves squarefreeness if and only if it preserves squarefreeness on words of length at most

$$\max\left\{3, 1 + \left\lceil \frac{M(f) - 3}{m(f)} \right\rceil\right\},$$

where $M(f) = \max_{a \in A} |f(a)|$ and $m(f) = \min_{a \in A} |f(a)|$.

# Consequences

- we have an algorithm to decide if a morphism is squarefree
- simply test if it is squarefree on words of a certain length (the bound in the theorem)
- What about $t$-powers?
- Recall: a square looks like $xx$; a $t$-power looks like $xx \cdots xx$ ($t$-times)

# A criterion for $t$-power-freeness

## Theorem (Richomme and Wlazinski 2007)

Let $t \geq 3$ and let $f : A^* \to B^*$ be a uniform morphism. There exists a finite set $T \subseteq A^*$ such that $f$ preserves $t$-power-freeness if and only if $f(T)$ consists of $t$-power-free words.

(uniform means the lengths of the images, $|f(a)|$, are the same for all $a \in A$)

# The general case

## Open problem

Is there an algorithm to determine if an arbitrary morphism is $t$-power-free?

# Changing the problem slightly

- our initial goal was to generate long $t$-power-free words
- a morphism that preserves $t$-power-freeness can accomplish this
- but some morphisms can generate long $t$-power-free words without preserving $t$-power-freeness in general

# An non-squarefree morphism

- consider $f$ defined by

$$a \to abc \qquad b \to ac \qquad c \to b$$

- iterates are squarefree:

$$a \to abc \to abcacb \to abcacbabcbac \to \cdots$$

- but $f(aba) = abcacabc$ is not

# Fixed points

- suppose $f$ generates an infinite word $\mathbf{x}$ by iteration
- we write $\mathbf{x} = f(\mathbf{x})$ and call $\mathbf{x}$ a fixed point of $f$
- Can we determine if $\mathbf{x}$ is $t$-power-free?

# Deciding if a fixed point is $t$-power-free

**Theorem (Mignosi and Séébold 1993)**

There is an algorithm to decide the following problem:

> *Given* $t \geq 2$ *and a morphism* $f$ *with fixed point* $\mathbf{x}$*, is* $\mathbf{x}$ $t$*-power-free?*

# Investigating a special class of morphisms

- we now restrict our attention to a particular class of morphisms
- primitive morphisms have nice properties that make them easy to analyse

# Primitive morphisms

- a morphism $f : \Sigma^* \to \Sigma^*$ is primitive if there is a constant $d$ such that for all $a, b \in \Sigma$, $a$ appears in $f^d(b)$
- the term “primitive” comes from matrix theory

# A example of a primitive morphism

Suppose $f$ maps

$$a \to ab \qquad b \to bc \qquad c \to a.$$

Then

$$\begin{array}{ccccccc} a & \to & ab & \to & abbc & \to & abbcbca \\ b & \to & bc & \to & bca & \to & bcaab \\ c & \to & a & \to & ab & \to & abbc \end{array}$$

and $a$, $b$, $c$ all appear in the third iterates.

# The matrix of a morphism

- let $f : \Sigma^* \to \Sigma^*$ be a morphism
- $\Sigma = \{a_1, a_2, \ldots, a_k\}$
- define a matrix

$$M = (m_{i,j})_{1 \leq i,j \leq k}$$

where $m_{i,j}$ is the number of occurrences of $a_i$ in $f(a_j)$

# An example

$$\begin{aligned} & a \to ab \\ f : \; & b \to bc \\ & c \to a. \end{aligned}$$

$$M = \begin{array}{c} \\ a \\ b \\ c \end{array} \begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix} \end{array}$$

# Primitive matrices

- a non-negative matrix $M$ is primitive if there is a positive integer $d$ such that $M^d > 0$
- the least such $d$ is the index of primitivity
- if $M$ is $k \times k$ then $d \leq k^2 - 2k + 2$ (Wielandt 1950)
- if a morphism is primitive then its matrix is primitive

# From the previous example

$$M = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix} \qquad M^3 = \begin{pmatrix} 2 & 2 & 1 \\ 3 & 2 & 2 \\ 2 & 1 & 1 \end{pmatrix} > 0$$

# Repetitions and primitive morphisms

## Theorem (Mossé 1992)

Let $\mathbf{x}$ be an infinite fixed point of a primitive morphism $f$.

Then either

- $\mathbf{x}$ is periodic, or
- there exists a positive integer $t$ such that $\mathbf{x}$ is $t$-power-free.

# Linear recurrence

- this result is a consequence of another important property
- an infinite word $\mathbf{x}$ is recurrent if each of its factors occurs infinitely often
- it is linearly recurrent if there exists a constant $C$ such that any factor of $\mathbf{x}$ of length $Cn$ contains all factors of $\mathbf{x}$ of length $n$.
- an infinite word generated by a primitive morphism is linearly recurrent

# The connection with repetitions

- let $\mathbf{x}$ be an aperiodic fixed point of a primitive morphism
- let $C$ be the constant of linear recurrence
- Claim: $\mathbf{x}$ does not contain any repetition of the form $v^C$

# Proving $\mathbf{x}$ avoids $C$-powers

- $\mathbf{x}$ aperiodic implies that for all $n$ the word $\mathbf{x}$ has at least $n+1$ factors of length $n$ (Coven and Hedlund 1973)
- suppose $\mathbf{x}$ contains $v^C$, where $|v| = m$
- $v^C$ contains $\leq m$ factors of length $m$
- but $|v^C| = Cm$ and by linear recurrence $v^C$ contains all factors of $\mathbf{x}$ of length $m$
- $\mathbf{x}$ has $\leq m$ factors of length $m$, contradiction

# Proving linear recurrence

It remains to prove:

## Theorem (Durand 1998)

If $\mathbf{x}$ is a fixed point of a primitive morphism $f$, then there exists a constant $C$ such that for every $n$, every factor of $\mathbf{x}$ of length $Cn$ contains every factor of $\mathbf{x}$ of length $n$.

# The Perron–Frobenius Theory

Let $M$ be the matrix of $f$; so $M$ is primitive. The fundamental result concerning primitive matrices is:

## Theorem (Perron 1907; Frobenius 1912)

A primitive matrix $M$ has a dominant eigenvalue $\theta$; i.e., $\theta$ is a positive, real eigenvalue of $M$ and is strictly greater in absolute value than all other eigenvalues of $M$.

# Asymptotic growth of $M^n$

## Corollary

The limit

$$\lim_{n\to\infty} \frac{M^n}{\theta^n}$$

exists and is positive.

# The length of the iterates of a morphism

- Let $f$ be a primitive morphism, $M$ its matrix, and $\theta$ the dominant eigenvalue of $M$.
- For each letter $a$, there exists a positive constant $C_a$ such that
$$\lim_{n\to\infty} \frac{|f^n(a)|}{\theta^n} = C_a.$$
- There exist positive constants $A, B$ such that for all $n$,
$$A\theta^n \leq \min_{a\in\Sigma} |f^n(a)| \leq \max_{a\in\Sigma} |f^n(a)| \leq B\theta^n.$$

# The constant of linear recurrence

- let $\mathbf{x}$ be a fixed point of $f$
- we want to define a $C$ such that any factor of $\mathbf{x}$ of length $Cn$ contains all factors of length $n$
- it is not hard to show that for $n = 2$ there exists $C_2$ such that every factor of length $C_2$ contains all factors of length $2$
- we focus on $n \geq 3$
- let $A, B, \theta$ be as defined previously
- Claim: we can take $C = (C_2 + 2)(B/A)\theta$.

# Establishing the claim

- write $\mathbf{x} = x_1 x_2 \cdots$
- consider a factor $w = x_i x_{i+1} \cdots x_{i+Cn-1}$ of $\mathbf{x}$
- $|w| = Cn$
- since $\mathbf{x}$ is a fixed point of $f$ we have $\mathbf{x} = f(\mathbf{x})$
- by iteration we have

$$\mathbf{x} = f^p(x_1) f^p(x_2) \cdots$$

for every $p \geq 1$

# Taking the preimage of $w$

- choose $p$ satisfying

$$\min_{a\in\Sigma} |f^{p-1}(a)| < n < \min_{a\in\Sigma} |f^p(a)|$$

- write $w = uf^p(x_r)f^p(x_{r+1})\cdots f^p(x_{r+j-1})v$
- $u$ and $v$ as small as possible
- we get

$$\begin{aligned} |w| = Cn &\leq |u| + |v| + j\max_{a\in\Sigma} |f^p(a)| \\ &\leq 2\max_{a\in\Sigma} |f^p(a)| + j\max_{a\in\Sigma} |f^p(a)| \end{aligned}$$

# Rearranging the last inequality

Rearrange to get

$$\begin{aligned} j &\geq \frac{Cn}{\max_{a\in\Sigma}|f^p(a)|} - 2 \\ &\geq \frac{(C_2+2)(B/A)\theta n}{B\theta^p} - 2. \end{aligned}$$

Recall that $n > \min_{a\in\Sigma} |f^{p-1}(a)| \geq A\theta^{p-1}$.

Using this inequality to replace $n$ gives

$$\begin{aligned} j &\geq \frac{(C_2+2)(B/A)\theta A\theta^{p-1}}{B\theta^p} - 2 \\ &= C_2. \end{aligned}$$

# Concluding the proof

- Recall: $w = uf^p(x_r)f^p(x_{r+1})\cdots f^p(x_{r+j-1})v$
- since $j \geq C_2$ we have $|x_r x_{r+1}\cdots x_{r+j-1}| \geq C_2$
- $x_r x_{r+1}\cdots x_{r+j-1}$ contains all factors of $\mathbf{x}$ of length $2$
- any factor of $\mathbf{x}$ of length $n$ is a factor of some $f^p(z)$, where $z$ is a factor of $\mathbf{x}$ of length at most $2$
- $w$ contains all such $f^p(z)$ and thus all factors of length $n$
- since $w$ was an arbitrary factor of length $Cn$, the proof is complete

# Recapping the argument

- we have shown that a fixed point $\mathbf{x}$ of a primitive morphism $f$ is linearly recurrent
- from this we deduced that $\mathbf{x}$ is either periodic, or avoids $C$-powers, where $C$ is the constant of linear recurrence
- this $C$ may not be optimal
- How can we tell if $\mathbf{x}$ is (ultimately) periodic?
- we address this question (for arbitrary morphisms) in the second part

# Subword complexity

- if $\mathbf{x}$ is an infinite word, its subword complexity function $p(n)$ counts the number of distinct factors of $\mathbf{x}$ of length $n$
- we have seen that $p(n)$ is bounded if $\mathbf{x}$ is ultimately periodic
- and that $p(n) \geq n + 1$ if $\mathbf{x}$ is aperiodic
- if $\mathbf{x}$ is generated by iterating a primitive morphism then $p(n) = O(n)$ (follows from linear recurrence)

# Possible complexity functions

## Theorem (Pansiot 1984)

Let $\mathbf{x}$ be an infinite word generated by iterating a morphism. The subword complexity function $p(n)$ of $\mathbf{x}$ satisfies one of the following: $p(n) = \Theta(1)$, $p(n) = \Theta(n)$, $p(n) = \Theta(n \log \log n)$, $p(n) = \Theta(n \log n)$, or $p(n) = \Theta(n^2)$.

# Complexity functions of repetition-free words

- Ehrenfeucht and Rozenberg (80's) investigated the subword complexities of repetition-free words generated by morphisms
- let $\mathbf{x}$ be an infinite word generated by iterating a morphism
- if $\mathbf{x}$ avoids $t$-powers for some $t \geq 2$, then $p(n) = O(n \log n)$
- if $\mathbf{x}$ is a cubefree binary word, then $p(n) = \Theta(n)$
- there is a cubefree ternary word with $p(n) = \Theta(n \log n)$

# Constructing such a cubefree word

Let $f$ be the morphism that maps

$$a \to ab, \quad b \to ba, \quad c \to cacbc.$$

Then

$$c \to cacbc \to cacbcabcacbcbacacbc \to \cdots$$

is cubefree and has complexity $p(n) = \Theta(n \log n)$.

(Note: $f$ is not primitive.)

# Complexity of squarefree words

- let $\mathbf{x}$ be an infinite word generated by iterating a morphism
- if $\mathbf{x}$ is a squarefree ternary word, then $p(n) = \Theta(n)$
- Ehrenfeucht and Rozenberg (1983) constructed a D0L language with subword complexity $p(n) = \Theta(n \log n)$

# Constructing the D0L language

Let $f$ be the morphism that maps

$$a \to abcab, \quad b \to acabcb, \quad c \to acbcacb$$

$$d \to dcdadbdadcdbdcd$$

The language obtained by repeatedly applying $f$ to the word $dabcd$ is squarefree and has complexity $p(n) = \Theta(n \log n)$

# Non-morphic words

- the previous results all concerned repetition-free words generated by iterating a morphism
- if we consider arbitrary words, then it is not too difficult to construct an infinite ternary squarefree word with exponential subword complexity

# The End