Periodicity and Repetitions in Automatic Sequences

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

Joint work with Jean-Paul Allouche and Narad Rampersad

Decision problems for infinite words

The goal:

Given an infinite word determine

- if $\mathbf{w}$ is ultimately periodic;
- if $\mathbf{w}$ has squares, cubes, or higher powers;
- if $\mathbf{w}$ has a given fractional power;
- what the lexicographically least (or greatest) word in the orbit closure of $\mathbf{w}$ is;
- if $\mathbf{w}$ is recurrent;
- if $\mathbf{w}$ is uniformly recurrent.

Morphic words

Can't answer these questions for arbitrary infinite words, so we restrict to a subclass: *morphic words*.

A *morphism* is a map h from Σ^* to Δ^* such that

$$h(xy) = h(x)h(y)$$

for all words x, y.

If $\Sigma = \Delta$, we can iterate h, writing h^2 for $h \circ h$, etc.

Morphic words

For example, if $\mu(0) = 01$, $\mu(1) = 10$, then

$$\begin{array}{rcl} \mu^0(0) & = & 0 \\ \mu^1(0) & = & 01 \\ \mu^2(0) & = & 0110 \\ \mu^3(0) & = & 01101001 \\ & \vdots & \end{array}$$

As each word is a prefix of the next, there is a unique infinite word of which each $\mu^i(0)$ is a prefix, which we write as

$$\mu^\omega(0) = 0110100110010110 \cdots ,$$

which is $\mathbf{t}$, the *Thue-Morse* word.

Morphisms and morphic words - terminology

- A morphism is said to be *k-uniform* if every letter is mapped to a word of length k.
- A morphism is *uniform* if it is k-uniform for some k.
- A *coding* is a 1-uniform morphism.
- An infinite word is said to be *pure morphic* if it can be generated by iterating a morphism.
- An infinite word is said to be *morphic* if it is the image (under a *coding*) of a pure morphic word.
- An infinite word is said to be *automatic* if it is the image (under a coding) of a word generated by iterating a k-uniform morphism.

Decision problems for repetitions

- Decision problem: given a morphism h, does the word $h^\omega(a)$ it generates by iteration contain squares?
- If the morphism is over a three-letter alphabet, a decision procedure was given by Berstel in 1979.

Decision problems for repetitions

- Similarly, one can ask, does the word generated by a morphism contains cubes, or higher powers?
- A general decision procedure was given by Mignosi and Séébold in 1993.
- Cassaigne (1994) gave a general decision procedure for certain kinds of HD0L words.
- Krieger (2007, 2008) gave a procedure to compute the critical exponent for binary uniform morphisms and an "almost algorithm" for arbitrary non-erasing morphisms.

Goal of this talk

- Will show how a simple idea leads to a general decision procedure for all kinds of repetitions in the case of *uniform morphisms* and, more generally, *automatic sequences*
- Will show how this idea can be applied to other kinds of questions

Automatic sequences

But first, one more observation:

- A sequence $(a_n)_{n\geq 0}$ is k-automatic if there is an automaton with output such that, after feeding in n expressed in base k, you arrive at a state with an output of a_n.
- The input alphabet is $\Sigma_k = \{0, 1, \ldots, k-1\}$.
- Cobham proved: a sequence is k-automatic iff it is k-uniform morphic.

Example: the Thue-Morse sequence

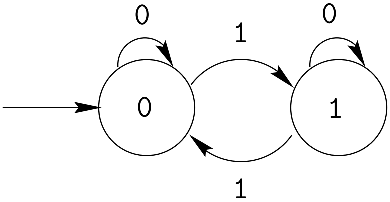

Figure: Automaton generating the Thue-Morse sequence

Testing if an automatic sequence is ultimately periodic

An infinite word $\mathbf{a} = (a_n)_{n \geq 0}$ is *ultimately periodic* if there exist integers $P \geq 1, N \geq 0$ such that $a_I = a_{I+P}$ for all $I \geq N$.

In 1986, Honkala gave a procedure to decide if an automatic sequence is ultimately periodic.

Leroux (LICS 2005) even gave a method to decide this question in polynomial time.

But both methods are rather complicated.

Testing if an automatic sequence is ultimately periodic

Our approach:

- Construct an NFA M_1 that on input (P, N) "guesses" I and accepts iff $I \geq N$ and $a_I \neq a_{I+P}$.
- Convert this NFA M_1 to a DFA M_2 using the usual subset construction.
- Interchange accepting and nonaccepting states, obtaining a DFA M_3 such that M_3 accepts (P, N) iff $a_I = a_{I+P}$ for all $I \geq N$.
- Now $\mathbf{a}$ is ultimately periodic iff M_3 accepts some input with $P \geq 1$, which can be checked using depth-first search to see if there is a path from M_3's initial state to a final state.

Filling in the details

What does it mean to have input (P, N)?

Answer: we actually feed in the base-k representation of P and N *in parallel*, starting with the least significant digit, where one expansion is padded with leading zeroes, if necessary.

Saying that the digits are fed in in parallel means the input alphabet is $\Sigma_k \times \Sigma_k$.

Filling in the details

What does it mean to "guess I"?

Answer: it really means we successively guess the base-k digits of I, starting with the least significant digit.

Filling in the details

How do we verify that $I \geq N$?

Answer: we maintain a flag that keeps track of whether the digits of I we guessed so far represent a number that is $\geq$ the digits of N seen so far, and we update this flag as we see additional digits.

$$
\begin{aligned}
u(<, i', n') &= \begin{cases} <, & \text{if } i' \leq n'; \\ >, & \text{if } i' > n'; \end{cases} \\
u(=, i', n') &= \begin{cases} <, & \text{if } i' < n'; \\ =, & \text{if } i' = n'; \\ >, & \text{if } i' > n'; \end{cases} \\
u(>, i', n') &= \begin{cases} <, & \text{if } i' < n'; \\ >, & \text{if } i' \geq n'. \end{cases}
\end{aligned}
$$

Filling in the details

What if the base-k representation of the appropriate I to guess is much larger than the base-k representations of the input P and N?

Answer: we use the fact that there are infinitely many different representations of P and N, arising from allowing leading zeroes (actually trailing zeroes, since we are inputting representations starting with the least significant digit).

To handle this, we modify the accepting states, allowing a state to be accepting if we could reach it by following a path labeled with $\overbrace{(0,0)\cdots(0,0)}^{j}$.

Applying the same idea

We can use the same idea to solve other problems, provided we can express our decision problem as a predicate involving quantifiers, addition, and inequalities.

Testing for the presence of overlaps

An *overlap* is a word of the form $axaxa$, where a is a single letter and x is a (possibly) empty string, as in `alfalfa` or `entente`.

Testing if a sequence contains an overlap can be phrased as:

$\mathbf{a} = (a_n)_{n \geq 0}$ contains an overlap if and only if there exist integers $I \geq 0$, $T \geq 1$ such that $a_{I+J} = a_{I+T+J}$ for all J with $0 \leq J \leq T$.

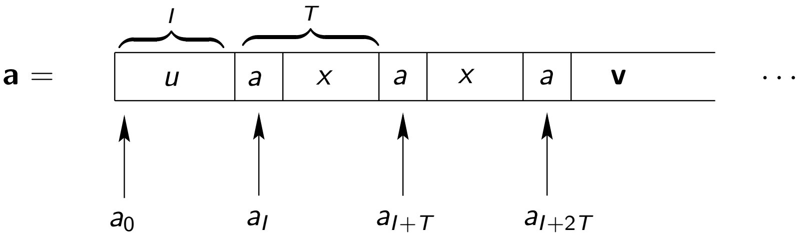

Figure: Hypothesized overlap

Carrying out the construction for the Thue-Morse sequence

- Whether an automatic sequence has overlaps is decidable
- We carried out the overlap-testing construction for the Thue-Morse sequence **t**.
- The original NFA M_1 had 72 states.
- We converted this to a DFA with 801 states.
- We then minimized, obtaining a DFA with 2 states accepting only words with $T = 0$. Thus **t** is overlap-free.

Fractional powers

We say a word $x = a_0 a_1 \cdots a_{n-1}$ is a p/q power (for integers $p \geq 0, q \geq 1$) if

- p divides n; and
- x is periodic with period length nq/p; in other words, $a_i = a_{i+nq/p}$ for all suitable i.

For example:

- `entanglement` is a $4/3$-power;
- `alfalfa` is a $7/3$-power.

Applying our idea to fractional powers

Expressed as predicate:

$\mathbf{a} = (a_n)_{n \geq 0}$ contains a $\geq p/q$-power iff there exist $I \geq 0,\ T \geq 1$ such that $a_{I+J} = a_{I+T+J}$ for all J, $0 \leq qJ < (p-q)T$.

Thus whether a k-automatic sequence has a particular fractional power is decidable.

Decidability of similar predicates

In a similar way, the following problems are also decidable:

Whether a k-automatic sequence:

- Has infinitely many occurrences of α-powers;
- Has infinitely many distinct α-powers;
- Avoids palindromes of length $\geq L$;
- Satisfies the property that if x is a factor, then x^R is not, for all x of length $\geq L$.

Testing recurrence

An infinite word $\mathbf{a} = (a_n)_{n \geq 0}$ is said to be *recurrent* if every factor that occurs at least once in **a** occurs infinitely often.

Given an automatic sequence, can we decide if it is recurrent?

Using our technique, the answer is yes. To see this, rewrite the definition of "recurrent" as follows: a word is recurrent if and only if for each occurrence of a factor of **a**, there exists a later occurrence of that factor in **a**.

Equivalently, for every $N \geq 0$, $K \geq 1$, there exists $M > N$ such that $a_{N+I} = a_{M+I}$ for $0 \leq I < K$.

Testing uniform recurrence

An infinite word $\mathbf{a} = (a_n)_{n \geq 0}$ is said to be *uniformly recurrent* if every factor that occurs at least once in $\mathbf{a}$ occurs infinitely often, with bounded gaps between consecutive occurrences.

Given an automatic sequence, can we decide if it is uniformly recurrent?

Using our technique, the answer is yes. To see this, rewrite the definition of "uniformly recurrent" as follows: a word $\mathbf{a} = (a_n)_{n \geq 0}$ is uniformly recurrent iff for every $K \geq 1$ there exists $A > 0$ such that for every $N \geq 0$ there exists $M \geq 0$ with $N < M < N + A$ such that $a_{N+I} = a_{M+I}$ for $0 \leq I < K$.

Better results have recently been obtained by Nicolas and Pritykin.

Orbit closure

- We can associate a dynamical system with any infinite word: the topological closure of all shifts of that word.
- The *orbit closure* of a sequence $\mathbf{a} = (a_n)_{n \geq 0}$ is the set of all sequences $\mathbf{b} = (b_n)_{n \geq 0}$ such that every finite prefix of $\mathbf{b}$ is a factor of $\mathbf{a}$.
- If $\mathbf{a}$ is recurrent but not periodic, then the orbit closure is uncountable, but this is not necessarily true if $\mathbf{a}$ is not recurrent.

Orbit closure

Two sequences of particular interest in the orbit closure are the *lexicographically least* and *lexicographically greatest*.

For example, the lexicographically least sequence in the orbit closure of $\mathbf{t}$ is the sequence obtained from the complement of $\mathbf{t}$ by dropping the first letter:

$$001011001101001\cdots$$

Orbit closure

We can define the lexicographically least sequence in the orbit closure as follows:

Let $\mathbf{a} = (a_n)_{n \geq 0}$ be a sequence, and let $\mathbf{b} = (b_n)_{n \geq 0}$ be the lexicographically least sequence in the orbit closure of $\mathbf{a}$. Then $b_I = c$ if and only if there exists $J \geq 0$ such that $a_{J+I} = c$ and $a_L a_{L+1} \cdots a_{L+I} \geq a_J a_{J+1} \cdots a_{J+I}$ for all $L \geq 0$.

Theorem If $\mathbf{a} = (a_n)_{n \geq 0}$ is a k-automatic sequence then so is the lexicographically least sequence in the orbit closure of $\mathbf{a}$.

Continued fractions

Every real number α can be expressed essentially uniquely as a *continued fraction*

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \ddots + \cfrac{1}{a_n}}}}$$

which is usually abbreviated

$$[a_0, a_1, \ldots, a_n].$$

Here the a_i are positive integers except possibly for a_0, which can be any integer, and the expansion terminates iff α is irrational. The a_i are called *partial quotients*.

Continued fractions

We say a real number is k-automatic if its continued fraction has bounded partial quotients and the sequence of partial quotients is k-automatic.

If we truncate a continued fraction $[a_0, a_1, \ldots, a_n]$ after the $(n+1)$th term, we get a rational number p_n/q_n, called the n'th *convergent*.

Galois proved that

$$q_n/q_{n-1} = [a_n, a_{n-1}, \ldots, a_1]$$

Our original motivation

We can use our technique to prove

Theorem. Suppose $\alpha = [a_0, a_1, \ldots,]$ is an irrational real number with a k-automatic continued fraction expansion. Then $\limsup_{n\to\infty} q_n/q_{n-1}$ does too.

The idea of the proof is similar to what we have seen, but there are some small complications arising from the fact that the ordering on continued fraction expansions is not quite the lexicographic order (it reverses alternately at odd and even positions).

The quantity $\limsup_{n\to\infty} q_n/q_{n-1}$

The quantity $\limsup_{n\to\infty} q_n/q_{n-1}$ comes up in

- The value of the recurrence quotient of a Sturmian word with slope α (Cassaigne);
- irrationality measure of numbers of the form $(b-1)\sum_{n\geq 1} b^{-\lfloor n\alpha \rfloor}$; (Adamczewski & Allouche);
- critical exponent of Sturmian words (Damanik & Lenz; Cao & Wen)

An example

For integers $k \geq 3$ the real number

$$\alpha_k = \sum_{i \geq 0} k^{-2^i} = [0, k-1, k+2, k, k, k-2, k, k+2, k, k-2, k+2, k, \ldots]$$

has a 2-automatic continued fraction expansion, as given below:

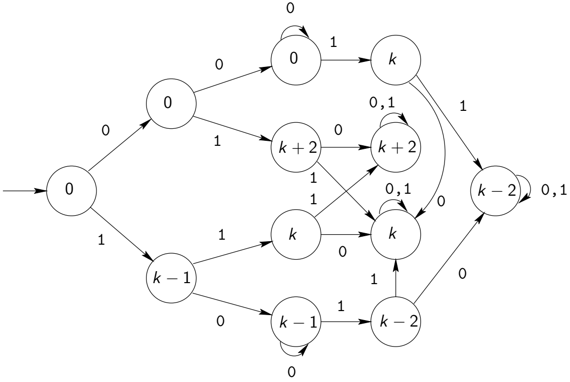

Figure: Automaton generating the continued fraction for α_k

Then
$\zeta = \limsup_{n\geq 0} q_n/q_{n-1} = [k+2, k-2, k, k+2, k, k-2, k, k, \ldots]$
is 2-automatic.

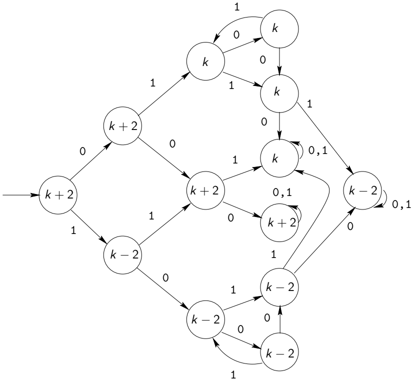

Figure: Automaton generating the continued fraction for ζ

An open problem

Let $\mathbf{r} = (r_n)_{n \geq 0}$ be the Rudin-Shapiro sequence, defined as follows: r_n is the parity of the number of occurrences of 11 in the binary expansion of n.

Then the lexicographically least sequence in the orbit closure of $\mathbf{r}$ seems to be the sequence obtained by concatenating 0 on the front of $\mathbf{r}$.

But we have not succeeded in proving this rigorously.

Another open problem

Extend all these ideas to arbitrary morphic sequences, not just uniform ones.

For Further Reading

- F. Nicolas and Yu. Pritykin, On uniformly recurrent morphic sequences, *Int. J. Found. Comput. Sci.*, to appear.
- J.-P. Allouche, N. Rampersad, and J. Shallit, Periodicity, repetitions, and orbits of an automatic sequence. *Theoret. Comput. Sci.*, to appear.