Combinatorics on Words

Gabriele Fici

CWI, Amsterdam — April 2024

Part 2: Infinite Words

Infinite Words

We start by giving the fundamental definitions about infinite words.

Definition 1

An infinite word $x = x_0 x_1 x_2 \cdots$ over Σ is a non-ending sequence of elements of Σ, that is, a map from the set $\mathbb{N}$ of natural numbers to Σ.

We let $\Sigma^{\mathbb{N}}$ denote the set of all infinite words over Σ, that is, the set of all maps from $\mathbb{N}$ to Σ.

Infinite Words

Definition 2

An infinite word x is purely periodic if $x = w^\omega$ for some nonempty word w, where w^ω stands for the infinite word $www\cdots$ obtained by concatenating an infinite number of copies of w.

An infinite word x is ultimately periodic if x is not purely periodic but can be written as $x = ux'$ for a finite nonempty word u (that we suppose of minimal length, to make the expression $x = ux'$ unique) and a purely periodic word x'.

An infinite word is aperiodic if it is not purely periodic nor ultimately periodic.

For example, $x = 01(001)^\omega$ is purely periodic, since we can write it as $x = (010)^\omega$; while $x = 10(100)^\omega$ is ultimately periodic. Notice that the factor 101 occurs only once in it.

Infinite Words

The property of being periodic/aperiodic can be related to the unbordered factors of an infinite word.

Theorem 3

Every aperiodic infinite word contains arbitrarily long unbordered factors, whereas in a purely periodic infinite word the maximum length of an unbordered factor is bounded.

Note that there exist ultimately periodic words containing arbitrarily long unbordered factors (e.g., the word $x = 01^\omega = 011111\cdots$).

Infinite Words

The following theorem extends the Lyndon–Schützenberger theorem to purely periodic words.

Theorem 4

Let $x, y \in \Sigma^+$. The following conditions are equivalent:

1. $xy = yx$;
2. $x^\omega = y^\omega$;
3. $(xy)^\omega = (yx)^\omega$;
4. $(xy)^\omega = x^\omega$;
5. $(xy)^\omega = y^\omega$.

Infinite Words

Another aspect of infinite words often taken into account (which comes from symbolic dynamics) is related to the occurrences of finite factors.

Definition 5

An infinite word x is recurrent if every finite factor of x occurs in x infinitely often. Equivalently, x is recurrent if and only if every finite prefix of x has a second occurrence as a factor.

An infinite word x is uniformly recurrent if every finite factor of x occurs syndetically (that is, it occurs infinitely often and with bounded gaps). Equivalently, x is uniformly recurrent if and only if for every finite factor u of x there exists an integer m (that depends on u) such that u occurs in every factor of x of length $m|u|$.

An infinite word x is linearly recurrent if there exists an integer m such that for every finite factor u of x, u occurs in every factor of x of length $m|u|$.

Infinite Words

Remark 6

An ultimately periodic word is not recurrent. A purely periodic word is (linearly) recurrent. Therefore, a recurrent word is either aperiodic or purely periodic.

Remark 7

Let x be an aperiodic infinite word. If there exists a nonempty word u such that u^n is a factor of x for every $n \geq 0$, then x cannot be uniformly recurrent.

Infinite Words

Definition 8

Given an infinite word x, the recurrence function $R_x(n)$ is defined as the minimum m such that every factor of length m of x contains at least one occurrence of every factor of length n of x.

For a uniformly recurrent word x, the recurrence function is defined for every n. For a linearly recurrent word, the recurrence function is a linear function (whence the name).

Equivalently, a recurrent word x is linearly recurrent (with constant m, i.e., $R_x(n) \leq mn$ for every $n \geq 0$) if and only if for every factor u of x, the distance between two consecutive occurrences of u in x is at most $(m-1)|u| + 1$.

Infinite Words

Example 9

The word $x = \prod_{n>0} 01^n = 010110111\cdots$ is aperiodic and not recurrent. For example, 010 occurs only once in it.

The infinite word obtained by concatenating the binary representation of n for every $n \geq 0$, called the Champernowne word

$$C_2 = 01101110010111011110001001101010111100110111101111\cdots$$

is aperiodic and recurrent but not uniformly recurrent.

Infinite Words

The Thue–Morse word

$$t = 0110100110010110100101100110100110010110011010010110100 \cdots$$

is the word such that the letter in position $n \geq 0$ is equal to the number of occurrences of 1, modulo 2, in the binary representation of n. So, for example, the binary representation of 5 is 101, which has an even number of 1s, hence $t(5) = 0$, while the binary representation of 2 is 10, which has an odd number of 1s, hence $t(2) = 1$.

The Thue–Morse word can be generalized in several ways. For example, by considering the infinite word whose nth letter is equal to the number of occurrences of 1, modulo k, in the binary representation of n. For $k = 3$ one obtains the word

$$\hat{t} = 011212201220200112202001200101121220200120010112200101120 \cdots$$

Infinite Words

Another possible generalization of the definition of the Thue–Morse word gives the Rudin–Shapiro word

$$r = 0001001000011101000100101110001000010010000111011110110\cdots$$

whose nth letter is equal to the number of occurrences (possibly overlapping) of 11, modulo 2, in the binary representation of $n \geq 0$.

Another possible generalization is to consider the number of occurrences of $k-1$, modulo 2, in the k-ary representation of $n \geq 0$. For $k = 3$, this gives the Mephisto–Waltz word

$$mw = 00100111000100111011011000100100111000100111011011000\cdots$$

Infinite Words

Since the Thue–Morse word can also be defined as the word whose nth letter is the sum of digits modulo 2 of the binary representation of n, we can define the generalized Thue–Morse word t_k, $k \geq 2$, as the word whose nth letter is the sum of digits, modulo k, of the k-ary representation of $n \geq 0$. For $k = 3$ we get

$$t_3 = 012120201120201012201012120120201012201012120012120201\cdots$$

A less studied word, called twisted Thue–Morse word, is the word such that the letter in position $n \geq 1$ is equal to the number of occurrences of 0, modulo 2, in the binary representation of n:

$$tt = 01001101001011001101001100101101001011001101001011010011\cdots$$

Infinite Words

n	binary	occ. of 1	mod. 2	mod. 3	occ. of 11	mod. 2
0	0 or ε	0	0	0	0	0
1	1	1	1	1	0	0
2	10	1	1	1	0	0
3	11	2	0	2	1	1
4	100	1	1	1	0	0
5	101	2	0	2	0	0
6	110	2	0	2	1	1
7	111	3	1	0	2	0
8	1000	1	1	1	0	0
9	1001	2	0	2	0	0
10	1010	2	0	2	0	0
11	1011	3	1	0	1	1
12	1100	2	0	2	1	1
13	1101	3	1	0	1	1
14	1110	3	1	0	2	0
15	1111	4	0	1	3	1
16	10000	1	1	1	0	0
17	10001	2	0	2	0	0

Infinite Words

The Fibonacci word

$$f = 0100101001001010010100100101001001010010100100101001010\cdots$$

is the word such that, for every $n > 0$, the distance between the nth 0 and the nth 1 is n.

The Fibonacci word is an example of a Sturmian word, an aperiodic word having exactly $n + 1$ distinct factors of length n, for every $n \geq 0$. We will have a lecture devoted to Sturmian words.

Infinite Words

The period-doubling word

$d = 0100010101000100010001010100010101000101010001000100010101\cdots$

is the word whose nth letter is the parity of the number of trailing 0s in the binary representation of $n > 0$.

Infinite Words

Another example is the (regular) paperfolding word

$$p = 0010011000110110001001110011011000100110001101110010 01\cdots$$

which is the sequence of ridges and valleys obtained by unfolding a sheet of paper that has been folded in half infinitely many times in the same direction.

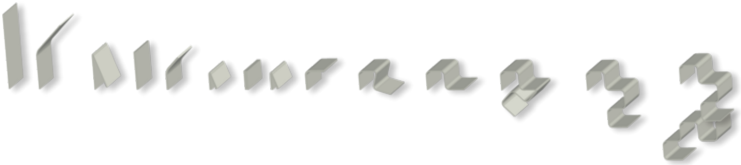


Figure: The paperfolding word $p = 001001100011011\cdots$ can be obtained by unfolding a sheet of paper which has been folded in half infinitely many times along the same direction.

Infinite Words

The n-th letter of the paperfolding word is actually the digit on the left of the rightmost 1 in the binary representation of n (writing the binary representation with leading zeroes).

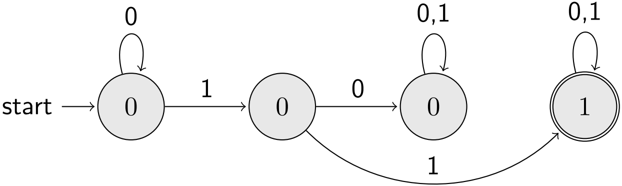
n	binary	p
1	$0000\underline{0}1$	0
2	$000\underline{0}10$	0
3	$0000\underline{1}1$	1
4	$00\underline{0}100$	0
5	$0001\underline{0}1$	0
6	$000\underline{1}10$	1
7	$0001\underline{1}1$	1
8	$0\underline{0}1000$	0
9	$0010\underline{0}1$	0
10	$001\underline{0}10$	0
11	$0010\underline{1}1$	1
12	$00\underline{1}100$	1
13	$0011\underline{0}1$	0
14	$001\underline{1}10$	1
15	$0011\underline{1}1$	1

Infinite Words

The paperfolding word can be therefore defined in terms of recognizability of words.

Let w be the binary representation of n (with leading zeroes). Then n-th letter of p is 1 if and only if $\tilde{w}$ belongs to the language $0^*(11)(0+1)^*$.

So, the n-th letter of p is determined by reading the (reverse of the) binary representation of n on a deterministic finite automaton (with output) recognizing the language $0^*(11)(0+1)^*$.



For this reason, the regular paperfolding word is a 2-automatic word, i.e., its digits are produced by a finite state automaton with output taking as inputs the binary representations of integers.

Infinite Words

A word defined by means of the ternary representation of integers is

$$lnd_3 = 1211221211211221221211221211211221211211221221211\cdots$$

whose nth letter is the last nonzero digit in the ternary representation of $n \geq 1$. (This can be generalized to every base $k \geq 3$.)

The Sierpiński word (also known as Cantor word)

$$s = 0101110101111111110101110101111111111111111111111111110\cdots$$

is the word such that the letter in position $n \geq 0$ is 1 if the ternary representation of n contains at least a 1.

The words lnd_3 and s are 3-automatic (notice that s is not uniformly recurrent as it contains arbitrarily large powers of 1).

Infinite Words

A kind of different example of an infinite word is the Oldenburger–Kolakoski word. It is the fixed point over the alphabet $\{1, 2\}$ of the operator Δ, called runlength encoding, which counts the lengths of the maximal consecutive blocks (runs) of the same letter.

For example, the runlength encoding of the word $w = 001110$ is $\Delta(w) = 231$, since in w there is a run of two 0s followed by a run of three 1s followed by a run of one 0.

For a word over the alphabet $\{1, 2\}$ that has the property that neither of 111 and 222 occurs as a factor, the runlength encoding operator produces a word over the same alphabet $\{1, 2\}$.

The Oldenburger–Kolakoski word

$$k = 2211212212211211221211212211211212212211212212112112212\cdots$$

is one of the two the fixed points of the runlength encoding operator (the other is the word $1k = 1221121221\cdots$).

Infinite Words

In fact, the Oldenburger–Kolakoski word is an example of self-generating word. Starting from the first letter, 2, by definition the first run must have length 2, so the word begins with 22; but this implies that also the second run must have length 2, hence k begins with 2211; now the third and the fourth run have length 1, so k begins with 221121, and so on.

It is an open problem whether the Oldenburger–Kolakoski word is recurrent.

Infinite Words

Another definition often taken into account when studying infinite words is the following:

Definition 10

Let x be an infinite word over Σ. The frequency of the letter $a \in \Sigma$ in x, when it exists, is the limit of the ratio between the number of occurrences of the letter a in the prefix of x of length n, and n.

For example, in the Thue–Morse word t one has that the frequency of 0 and the frequency of 1 are both equal to $1/2$.

In the Fibonacci word, the frequency of 0 is $\lim_{n\to\infty} F_n/F_{n+1} = 1/\varphi = \varphi - 1 \approx 0.618$ and the frequency of 1 is $\lim_{n\to\infty} F_n/F_{n+2} = 1/\varphi^2 = 2 - \varphi \approx 0.382$.

It is conjectured that the letter frequencies of the Oldenburger–Kolakoski are equal to $1/2$.

Infinite Words

Definition 11

An infinite word is rich if every its finite factor is rich, that is, contains as many nonempty palindromes as its length.

For example, the Fibonacci word (actually, every Sturmian word) is rich.

The period-doubling word is rich.

On the opposite, the Thue–Morse word, although it contains infinitely many palindromes, is not rich since, for example, its prefix 011010011 is not rich (the complete return 11010011 to the palindrome 11 is not a palindrome).

Infinite Words

Proposition 12

A rich infinite word is recurrent if and only if its set of factors is closed under reversal.

However, there exist aperiodic recurrent rich infinite words that are not uniformly recurrent. An example is the Sierpiński word $s = 0101110101111111110\cdots$

Proposition 13

A recurrent rich infinite word has infinitely many palindromic prefixes.

Limits of Sequences

One way to define an infinite aperiodic word consists in considering a sequence of finite words of increasing length that are one prefix of another and taking the limit of the sequence.

This limit is well defined in the (ultra)-metric space of infinite words where the distance between two words x and y is defined as $2^{-\delta}$, where δ is the length of the longest common prefix of x and y, provided that one defines, for a finite word w and an infinite word x, $d(w, x) = d(w\#^{\omega}, x)$, where $\#$ is a symbol not belonging to the alphabet of x.

Limits of Sequences

For example, over Σ_2, define the sequence of words $t_0 = 0$ and $t_n = t_{n-1}\bar{t}_{n-1}$ for every $n > 0$, where $\bar{t}_k$ is the binary complement of t_k, that is, the word obtained from t_k by applying the automorphism of Σ_2 that exchanges 0 and 1.

The limit for n that goes to infinity of the sequence t_n is the Thue–Morse word t.

$$\begin{aligned}
&t_0 = 0\\
&t_1 = 01\\
&t_2 = 0110\\
&t_3 = 01101001\\
&t_4 = 0110100110010110\\
&t_5 = 01101001100101101001011001101001\\
&\ldots
\end{aligned}$$

Limits of Sequences

$$\begin{aligned} t_0 &= 0 \\ t_1 &= 01 \\ t_2 &= 0110 \\ t_3 &= 01101001 \\ t_4 &= 0110100110010110 \end{aligned}$$

Remark 14

For every even n, the word t_n is a palindrome, while for every odd n, the word t_n is an antipalindrome, that is, its reversal $\tilde{t}_n$ is equal to its complement $\bar{t}_n$.

Remark 15

One has

$$t = 0 \prod_{i=0} \bar{t}_i = 0 \cdot 1 \cdot 10 \cdot 1001 \cdot 10010110 \cdots$$

Limits of Sequences

The Mephisto–Waltz word mw generalizes this construction, since it can be defined as the limit of the sequence of words $mw_0 = 0$ and $mw_n = mw_{n-1}mw_{n-1}\overline{mw}_{n-1}$ for every $n > 0$. So, $mw_1 = 001$, $mw_2 = 001001110$, $mw_3 = 001001110001001110110110001$, etc.

Another possible generalization consists in considering the sequence of words defined by $tmm_0 = 0$ and $tmm_n = tmm_{n-1}\overline{tmm}_{n-1}\overline{tmm}_{n-1}$ for every $n > 0$. So, $tmm_1 = 011$, $tmm_2 = 011100100$, etc. The limit of this sequence is the Thue–Morse–Morse word

$$tmm = 0111001001000110111000110111000110110111001\cdots$$

Using the reversal instead of the binary complement, we can define the sequence of words $x_0 = 01$ and $x_n = x_{n-1}x_{n-1}\widetilde{x_{n-1}}$ for every $n > 0$, whose limit is the Stewart–Thue–Morse word

$$stm = 0101100101100110100101100101100110100101100\cdots$$

Limits of Sequences

The paperfolding word p is the limit of the sequence of words p_n defined by $p_0 = 0$ and $p_n = p_{n-1}0\widetilde{\overline{p_{n-1}}}$ for every $n > 0$, where $\widetilde{\overline{p_k}}$ is the reversal of the binary complement of p_k.

Hence, the first few values of the sequence are: $p_1 = 0 \cdot 0 \cdot 1$, $p_2 = 001 \cdot 0 \cdot 011$, $p_3 = 0010011 \cdot 0 \cdot 0011011$, etc.

Indeed, passing from p_n to p_{n+1} describes what happens by folding one more time the sheet of paper along the same direction and then unfolding it.

Limits of Sequences

Consider now the sequence of words defined by $f_1 = 1$, $f_2 = 0$ and $f_n = f_{n-1}f_{n-2}$ for $n > 2$.

The limit of the sequence f_n is the Fibonacci word f.

The words of the sequence (f_n) are called Fibonacci finite words.

Indeed, the sequence of the lengths of the Fibonacci finite words is the sequence F_n of the Fibonacci numbers $1, 1, 2, 3, 5, 8, 13, \ldots$, defined by $F_1 = F_2 = 1$ and, for every $n > 2$, $F_n = F_{n-1} + F_{n-2}$.

$$\begin{aligned}
f_1 &= 1\\
f_2 &= 0\\
f_3 &= 01\\
f_4 &= 010\\
f_5 &= 01001\\
f_6 &= 01001010\\
f_7 &= 0100101001001
\end{aligned}$$

Limits of Sequences

A generalization of the Fibonacci word is the Tribonacci word.

Recall that the sequence of Tribonacci numbers T_n is defined by $T_1 = 1$, $T_2 = 2$, $T_3 = 4$ and for every $n > 3$, $T_n = T_{n-1} + T_{n-2} + T_{n-3}$. The first few values of the sequence T_n are $1, 2, 4, 7, 13, 24, 44, \ldots$

Consider now the sequence of words defined by $tr_1 = 0$, $tr_2 = 01$, $tr_3 = 0102$ and for $n > 3$, $tr_n = t_{n-1}tr_{n-2}tr_{n-3}$. The limit for n that goes to infinity of the sequence tr_n is the Tribonacci word

$$tr = 0102010010201010201001020102010010201010201001020100102\cdots$$

As it is easy to guess, the sequence of lengths of the words tr_n is the sequence of Tribonacci numbers.

Limits of Sequences

The Pell word

$$pl = 001001000100100010010010001001000100100100010010001001000\cdots$$

is defined as the limit of the sequence of words $pl_0 = 0$, $pl_1 = 001$ and $pl_n = pl_{n-1}pl_{n-1}pl_{n-2}$ for every $n > 1$.

So, $pl_2 = 001 \cdot 001 \cdot 0$, $pl_3 = 0010010 \cdot 0010010 \cdot 001$, etc.

It is a word analogue of Pell numbers, defined by $P_0 = 0$, $P_1 = 1$ and $P_n = 2P_{n-1} + P_{n-2}$ for each $n > 1$.

The first few Pell numbers are: $0, 1, 2, 5, 12, 29, 70, 169, 408, 985, \ldots$

The sequence of lengths of the words pl_n is the sequence of consecutive sums of Pell numbers.

Limits of Sequences

The period-doubling word

$$d = 0100010101000100010001010100010101000101010001000100010 \cdots$$

is the limit of the sequence of words d_n defined by $d_0 = 0$ and $d_{n+1} = d_n \, d'_n$ for every $n > 0$, where d'_n is the word obtained from d_n by changing the last letter.

The first few values of the sequence d_n are $d_1 = 01$, $d_2 = 0100$, $d_3 = 01000101$, etc.

Limits of Sequences

The Sierpiński word

$s = 0101110101111111110101110101111111111111111111111111110\cdots$

is the limit of the sequence of words defined by $s_0 = 0$ and $s_{n+1} = s_n 1^{3^n} s_n$ for $n \geq 1$.

The von Neumann word

$v = 0010011001001110010011001001111001001100100111001001100\cdots$

is the limit of the sequence of words defined by $v_0 = 0$ and $v_{n+1} = v_n 1^{n-1} v_n$ for $n \geq 1$.

Limits of Sequences

The Chacon word

$$c = 0010001010010001000101001010010001010010001000101001000 \cdots$$

is the limit of the sequence of words defined by $c_0 = 0$ and for every $n > 0$, $c_n = c_{n-1}c_{n-1}1c_{n-1}$.

$$\begin{aligned} c_0 &= 0 \\ c_1 &= 0010 \\ c_2 &= 0010001010010 \\ c_3 &= 0010001010010001000101001010010001010010 \end{aligned}$$

Morphisms

The rules used for producing an infinite word as the limit of a sequence of finite words that we described in the previous section are of different types and do not follow a well-defined scheme.

For example, it is not clear what kind of transformations can be used for generating the next word in the sequence.

From an algebraic point of view, a more natural approach consists in defining infinite words using morphisms.

Morphisms

Definition 16

Given two alphabets Σ and Δ, a morphism is a map μ from Σ^* to $\Delta*$ such that $\mu(uv) = \mu(u)\mu(v)$ for any words u and v. When $\Delta = \Sigma$, we say that μ is an endomorphism of Σ^*.

We restrict our attention to non-erasing morphisms (i.e., such that $\mu(a) \neq \varepsilon$, for every $a \in \Sigma$).

By definition, a morphism can be described by just specifying the images of the letters of Σ. The domain of a morphism can be easily extended to infinite words.

Remark 17

A morphism injective on Σ^* is a (variable length) code.

Morphisms

A morphism is called uniform if the images of the letters have the same length k, also called the length of the uniform morphism.

A coding is a morphism of length 1, i.e., a mapping of letters (not necessarily injective), i.e., a partition of the alphabet.

Morphisms

Definition 18

A morphism μ such that there exists a letter a for which $\mu(a)$ is a word starting with a and $\lim_{n\to\infty} |\mu^n(a)| = +\infty$ is called prolongable on a. If μ is non-erasing and prolongable on a, we can iterate it to obtain a fixed point of μ, that is the infinite word $x = \lim_{n\to\infty} \mu^n(a)$ such that $x = \mu(x)$. Such an infinite word is called a pure morphic word.

Definition 19

A morphism μ is irreducible if for every pair of letters a, b in Σ, there exists a positive integer k such that the letter a occurs in $\mu^k(b)$.

A morphism μ is primitive if it is irreducible and aperiodic, i.e., there exists a positive integer k such that, for every pair of letters a, b in Σ, the letter a occurs in $\mu^k(b)$[a].

[a] An example of non-primitive irreducible morphism is $0 \mapsto 1, 1 \mapsto 0$.

Morphisms

For example, the (primitive uniform) morphism

$$\tau : 0 \mapsto 01,\ 1 \mapsto 10,$$

called the Thue–Morse morphism, is prolongable both on 0 and on 1. The fixed point starting with 0 of τ is the Thue–Morse word t.

Indeed, $\tau(0) = 01$, $\tau^2(0) = \tau(\tau(0)) = 0110$, $\tau^3(0) = 01101001$, etc. so that the sequence 0, $\tau(0)$, $\tau^2(0)$, etc. is the sequence of Thue–Morse finite words t_i.

The fixed point starting with 1 of τ is the binary complement $\bar{t} = 1001011001101001 \cdots$ of the Thue–Morse word.

Morphisms

Remark 20

Let μ be a morphism prolongable on letter a and let $\mu(a) = aw$. Then its fixed point x starting with a will be equal to

$$x = a \cdot w \cdot \mu(w) \cdot \mu^2(w) \cdot \mu^3(w) \cdots$$

For example, the Thue–Morse word t can be obtained by

$$t = 0 \cdot 1 \cdot \tau(1) \cdot \tau^2(1) \cdot \tau^3(1) \cdots = 0 \cdot 1 \cdot 10 \cdot 1001 \cdot 10010110 \cdots$$

which is the factorization presented in Remark 15.

We call this factorization the natural factorization of the fixed point of μ starting with a.

Morphisms

The Mephisto–Waltz word mw is the fixed point starting with 0 of the (primitive uniform) morphism $0 \mapsto 001,\ 1 \mapsto 110$.

The Thue–Morse–Morse word tmm is the fixed point starting with 0 of the (primitive uniform) morphism $0 \mapsto 011,\ 1 \mapsto 100$.

The ternary Thue–Morse word $\hat{t}$ is the fixed point starting with 0 of the (primitive uniform) morphism

$$\hat{\tau} : 0 \mapsto 01,\ 1 \mapsto 12,\ 2 \mapsto 20$$

The generalized Thue–Morse word t_k is the fixed point starting with 0 of the (primitive uniform) morphism

$$\tau_k : 0 \mapsto 01 \cdots (k{-}1),\ 1 \mapsto 12 \cdots (k{-}1)0,\ \ldots,\ (k{-}1) \mapsto (k{-}1)0 \cdots (k{-}2)$$

Morphisms

The fixed point of the (primitive) morphism

$$\varphi : 0 \mapsto 01,\ 1 \mapsto 0$$

is the Fibonacci word f.

Indeed, $\varphi(0) = 01$, $\varphi^2(0) = \varphi(\varphi(0)) = 010$, $\varphi^3(0) = 01001$, etc. so that the sequence 1, 0, $\varphi(0)$, $\varphi^2(0)$, etc. is the sequence of Fibonacci finite words f_i.

The Tribonacci word tr is the fixed point of the (primitive) morphism

$$0 \mapsto 01,\ 1 \mapsto 02,\ 2 \mapsto 0$$

More generally, for every $m > 1$, the (primitive) morphism

$$0 \mapsto 01,\ 1 \mapsto 02,\ \ldots, (m-2) \mapsto 0(m-1),\ (m-1) \mapsto 0$$

generates the so-called m-bonacci word.

Morphisms

The Pell word

$$pl = 00100100010010001001001000100100010010010001001000100\cdots$$

is the fixed point of the (primitive) morphism

$$\pi = 0 \mapsto 001,\ \ 1 \mapsto 0$$

The period-doubling word

$$d = 01000101010001000100010101000101010001010100010001000\cdots$$

is the fixed point of the (primitive uniform) morphism

$$\delta : 0 \mapsto 01,\ \ 1 \mapsto 00$$

Morphisms

The word

$$lnd_3 = 12112212112112212212112212112112212112\cdots$$

(sequence of the last nonzero digits in the ternary representation of n) is the fixed point of the (primitive uniform) morphism

$$1 \mapsto 121,\ 2 \mapsto 122$$

More generally, for any $k \geq 3$, lnd_k is the fixed point of the (primitive uniform) morphism

$$1 \mapsto 12\cdots k1,\ 2 \mapsto 12\cdots k2,\ \ldots,\ k \mapsto 12\cdots kk$$

Remark 21

Taking the word $lnd_4 = 12311232123312311231123212331232123\cdots$ *modulo* 2*, one obtains the word* $\overline{d}$*, the binary complement of the period-doubling word.*

Morphisms

Remark 22

Every periodic infinite word is pure morphic.

Every ultimately periodic infinite word is morphic, but may not be pure morphic, as for example the word 001^ω, which cannot be generated by a morphsim.

On the other hand, it is not always easy to decide if a pure morphic word is (ultimately) periodic. An example of ultimately periodic pure morphic word is the fixed point of the morphism $0 \mapsto 012, 1 \mapsto 2, 2 \mapsto 1$.

The same pure morphic word can be generated by different morphisms. For example, composing a morphism with itself gives a morphism generating the same word.

As another example, the periodic word 0^ω can be generated by $0 \mapsto 00$, $1 \mapsto w$, for any word w, starting from 0, or by $0 \mapsto 0$, $1 \mapsto 01$, starting from 1.

Morphisms

An example of an aperiodic word that is not pure morphic is the twisted Thue–Morse word $tt = 0100110100101 10\cdots$.

Nevertheless, it is a morphic word, meaning that it is the image of a pure morphic word under a coding.

Indeed, it is obtained by applying to the word

$$012314212312142314212314231214212312142314212312142123142312\cdots$$

fixed point of the uniform morphism

$$0 \mapsto 01,\ 1 \mapsto 23,\ 2 \mapsto 14,\ 3 \mapsto 21,\ 4 \mapsto 12$$

the coding $0, 2, 3 \mapsto 0$; $1, 4 \mapsto 1$.

Morphisms

The twisted Thue–Morse word verifies the equation $x = 010\tau^2(x)$ (actually, it is the only infinite word verifying this equation), and also the equation $x = 0\tau(\overline{x}) = 0\tau'(x)$, where τ' is the twisted Thue–Morse morphism $0 \mapsto 10,\ 1 \mapsto 01$; moreover, we have

$$tt = 0\tau(1)\tau^2(0)\tau^3(1)\tau^4(0)\cdots = 0 \cdot 10 \cdot 0110 \cdot 10010110 \cdots$$

Equivalently,

$$tt = \prod_{i=0} t_i\bar{t}_{i+1} = (0 \cdot 10) \cdot (0110 \cdot 10010110) \cdots$$

Morphisms

Another example of an aperiodic word that is morphic but not pure morphic is the paperfolding word p. Indeed, it is obtained by applying to the word

$$p' = 0121032101230321012103230123032101210321012303230121 03\cdots$$

fixed point of the uniform morphism $0 \mapsto 01,\ 1 \mapsto 21,\ 2 \mapsto 03,\ 3 \mapsto 23$, the coding $0, 1 \mapsto 0$; $2, 3 \mapsto 1$.

The paperfolding word can also be obtained starting from 00 and applying the 2-letter substitution

$$\begin{aligned} 00 &\mapsto 0010 \\ 10 &\mapsto 0110 \\ 01 &\mapsto 0011 \\ 11 &\mapsto 0111 \end{aligned}$$

that is, $ab \mapsto 0a1b$, for every $a, b \in \Sigma_2$.

Morphisms

The Rudin–Shapiro word

$$r = 000100100001110100010010111000100001001000011101111011\cdots$$

is also morphic but not pure morphic. It can be obtained by applying to the word

$$r' = 0102013101023202010201313231013101020131010232023231320\cdots$$

fixed point of the uniform morphism $0 \mapsto 01,\ 1 \mapsto 02,\ 2 \mapsto 31,\ 3 \mapsto 32$, the coding $0, 1 \mapsto 0$; $2, 3 \mapsto 1$.

The Rudin–Shapiro word can also be obtained starting from 00 and applying the 2-letter substitution

$$\begin{aligned} 00 &\mapsto 0001 \\ 01 &\mapsto 0010 \\ 10 &\mapsto 1101 \\ 11 &\mapsto 1110 \end{aligned}$$

Morphisms

Remark 23

A celebrated theorem of Cobham states that a word is q-automatic if and only if it is obtained by applying a coding to a fixed point of a q-uniform morphism.

However, it must be noticed that a q-automatic word may also be obtained as a fixed point of a non-uniform morphism (and without coding). This is the case, for example, of the word

$$vtm = 210201210120210201202101210201210120210\cdots$$

(sometimes called Variant of Thue–Morse) fixed point of the morphism $0 \mapsto 1,\ 1 \mapsto 20,\ 2 \mapsto 210$. It is 2-automatic, since it can be obtained from the fixed point of $0 \mapsto 12,\ 1 \mapsto 13,\ 2 \mapsto 20,\ 3 \mapsto 21$ by applying the coding $0, 3 \mapsto 1;\ 1 \mapsto 2,\ 2 \mapsto 0$.

Morphisms

A pure morphic word may not be recurrent, for example the fixed point of the morphism $0 \mapsto 01,\ 1 \mapsto 1$ is 01^ω. But even if it is recurrent, it may not be uniformly recurrent.

An example is the Sierpiński word

$$s = 010111010111111111010111010\cdots$$

which is the fixed point of the (non-primitive) uniform morphism: $0 \mapsto 010,\ 1 \mapsto 111.$

The same holds for the von Neumann word

$$v = 0010011001001110010011001001111\cdots$$

fixed point of the (non-primitive) morphism $0 \mapsto 001, 1 \mapsto 1.$

Morphisms

However, if the morphism is primitive, then each of its fixed points is uniformly recurrent.

Indeed, let x be the fixed point starting with 0 of a primitive morphism μ. Let us prove that x is uniformly recurrent. Let w be a factor of x. Then w is a factor of some $\mu^n(0)$. Since μ is primitive, 0 occurs in $\mu^k(a)$ for every letter a. Thus w appears in every $\mu^{nk}(a)$, hence it appears infinitely often with bounded gaps.

Actually, Durand proved a stronger result:

Theorem 24 (Durand, 1998)

Every fixed point of a primitive morphism is linearly recurrent.

Morphisms

On the other hand, even a fixed point of a non-primitive morphism can be uniformly recurrent. An example is the Chacon word

$$c = 0010001010010001000101001010010001010010001000101001000\cdots$$

which is the fixed point of the non-primitive morphism $0 \mapsto 0010,\ 1 \mapsto 1$.

Finally, the Oldenburger–Kolakoski word k is not pure morphic, but can be obtained starting from 22 and applying the 2-letter substitution

$$\begin{aligned} 22 &\mapsto 2211 \\ 21 &\mapsto 221 \\ 12 &\mapsto 211 \\ 11 &\mapsto 21 \end{aligned}$$

It is an open question whether the Oldenburger–Kolakoski word is morphic.

Morphisms

word	symbol	morphism	primitive	uniform
Thue–Morse	t	$0 \mapsto 01, 1 \mapsto 10$	yes	yes
Mephisto–Waltz	mw	$0 \mapsto 001, 1 \mapsto 110$	yes	yes
Thue–Morse–Morse	tmm	$0 \mapsto 011, 1 \mapsto 100$	yes	yes
ternary Thue–Morse	$\hat{t}$	$0 \mapsto 01, 1 \mapsto 12, 2 \mapsto 20$	yes	yes
period-doubling	d	$0 \mapsto 01, 1 \mapsto 00$	yes	yes
last nonzero digit	lnd_3	$1 \mapsto 121, 2 \mapsto 122$	yes	yes
Variant of Thue–Morse	vtm	$0 \mapsto 1, 1 \mapsto 20, 2 \mapsto 210$	yes	no
Fibonacci	f	$0 \mapsto 01, 1 \mapsto 0$	yes	no
Tribonacci	tr	$0 \mapsto 01, 1 \mapsto 02, 2 \mapsto 0$	yes	no
Pell	pl	$0 \mapsto 001, 1 \mapsto 0$	yes	no
Sierpiński	s	$0 \mapsto 010,\ 1 \mapsto 111$	no	yes
Chacon	c	$0 \mapsto 0010, 1 \mapsto 1$	no	no
von Neumann	v	$0 \mapsto 001, 1 \mapsto 1$	no	no

Perron–Frobenius Theory

The Parikh vector (or composition vector, or abelianization map) of a word w over Σ_k is the vector $P(w) = (|w|_0, |w|_1, \ldots, |w|_{k-1})$, whose ith entry is the frequency of the letter i in w.

With each morphism one can associate the matrix whose columns are the Parikh vectors of the images of the letters, called the incidence matrix of the morphism.

For example, the incidence matrix of the Fibonacci morphism $\varphi : 0 \mapsto 01,\ 1 \mapsto 0$ is the matrix $M_\varphi = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$.

Perron–Frobenius Theory

Remark 25

For every word w and every $n \geq 1$, one has

$$P(\mu^n(w)) = M_\mu^n P(w).$$

As a consequence, $M_{\mu^n} = M_\mu^n$.

For example, the word $w = 01001$ has Parikh vector $(3, 2)$. If we apply to w the Fibonacci morphism φ, we get $\varphi(w) = 01001010$, whose Parikh vector is $\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} (3, 2) = (5, 3)$.

Perron–Frobenius Theory

Definition 26

A morphism μ is primitive if and only if its incidence matrix M_μ is primitive, i.e., there exists a positive integer d such that all the entries of M_μ^d are greater than 0.

Such a d, called index of primitivity, is at most $k^2 - 2k + 2$, where k is the size of the alphanet.

Perron–Frobenius Theory

The Perron–Frobenius theorem says that for any irreducible morphism (hence a fortiori for any primitive morphism) μ with incidence matrix M_μ there always exist an eigenvalue $\lambda > 0$ (called expansion number) and an associated eigenvector **u**,

$$M_\mu \mathbf{u} = \lambda \mathbf{u},$$

such that λ is a Perron–Frobenius number, i.e., all other eigenvalues of M_μ have modulus less than λ or, equivalently, λ is the radius of the spectrum of M_μ and is a simple root of the characteristic polynomial of M_μ.

For example, the characteristic polynomial of $M_\varphi = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$ is

$$|M_\varphi - \lambda I| = \begin{vmatrix} 1-\lambda & 1 \\ 1 & -\lambda \end{vmatrix} = \lambda^2 - \lambda - 1,$$

whose Perron–Frobenius eigenvalue is the golden ratio φ.

Perron–Frobenius Theory

Moreover, the normalized eigenvector associated with the Perron–Frobenius eigenvalue (normalized in such a way that the sum of the components is equal to 1) gives the frequencies of the letters in the fixed point of the morphism.

Indeed, the frequency of the generic letter a in the fixed point of μ starting with 0 is given by

$$\lim_{n\to\infty} \frac{|\mu^n(0)|_a}{|\mu^n(0)|}.$$

For example, the normalized eigenvector of the Perron–Frobenius eigenvalue φ for the incidence matrix M_φ is $(1/\varphi, 1/\varphi^2) = (\varphi - 1, 2 - \varphi) \approx (0.618, 0.382)$.

As a consequence, for fixed points of primitive morphisms the frequencies of the letters always exist.

Perron–Frobenius Theory

Proposition 27

Let x be a morphic (resp. an automatic) sequence. If the frequency of a letter exists, then it is an algebraic[a] *(resp. a rational) number.*

[a] An algebraic number is a number that is a root of a non-zero polynomial in one variable with integer (or, equivalently, rational) coefficients.

Perron–Frobenius Theory

Remark 28

A useful method to obtain a good approximation of the normalized Perron–Frobenius eigenvector is using the so-called power method. Let M_μ be the $k \times k$ incidence matrix of the primitive morphism μ. Let $v = (\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k})$. Then the normalized Perron–Frobenius eigenvector is equal to the normalization of the vector $\lim_{n\to\infty} M_\mu^n \cdot v$.

So, a good approximation of the frequencies of letters can be obtained by choosing a sufficiently large value of n.

For example, $M_\varphi^{10} = \begin{pmatrix} 89 & 55 \\ 55 & 34 \end{pmatrix}$ and $M_\varphi^{10} \cdot (0.5, 0.5) = (72, 44.5)$, which, divided by $72 + 44.5 = 116.5$, gives $(0.618026, 0.381974)$.

Notice that in general, for $n \geq 1$, $M_\varphi^n = \begin{pmatrix} F_{n+1} & F_n \\ F_n & F_{n-1} \end{pmatrix}$.

Perron–Frobenius Theory

Exercise 29

Do the same calculations for the Pell morphism π.

Show that for every $n \geq 1$, $M_\pi^n = \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix}^n = \begin{pmatrix} P_{n+1} & P_n \\ P_n & P_{n-1} \end{pmatrix}$.

Perron–Frobenius Theory

An algebraic integer $\lambda > 1$ is a Pisot number (or Pisot–Vijayaraghavan number) if all its algebraic conjugates α other than λ itself satisfy $|\alpha| < 1$.

The key property of a Pisot number λ is that the distance from λ^n to the nearest integer tends to zero as n tends to infinity.

Conversely, if λ is any algebraic number bigger than 1 with this property, then λ must be a Pisot number; it is a conjecture of Pisot that no transcendental number has this property.

Examples of Pisot numbers are the golden ratio $(1+\sqrt{5})/2$ and the silver ratio $1+\sqrt{2}$.

Definition 30

A primitive morphism is called a Pisot morphism if its Perron–Frobenius eigenvalue is a Pisot number.

Abstract Numeration Systems

Let $x = \lim_{n\to\infty} \mu^n(a)$ be a fixed point of a morphism μ of $\Sigma = \{a_1, a_2, \ldots, a_\sigma\}$ prolongable on the letter a, and let $n > 0$ be the maximum length of an image of a letter under μ.

We can associate with μ an automaton $\mathcal{A}_\mu$ in the following way: The set of states of $\mathcal{A}_\mu$ is Σ, and for each letter $a_i \in \Sigma$, letting $\mu(a_i) = a_{i_0} a_{i_1} \cdots a_{i_n}$, with $a_{i_j} \in \Sigma$, we add the transitions (a_i, j, a_{i_j}) for every j (where a transition (q_r, c, q_s) is an edge from state q_r to state q_s labeled by the letter c).

Remark 31

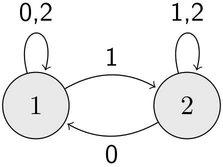
Notice that in this way the adjacency matrix of $\mathcal{A}_\mu$ is equal to the incidence matrix of μ.

Abstract Numeration Systems

For example, let $\mu : 1 \mapsto 121,\ 2 \mapsto 122$ be the morphism generating the word $lnd_3 = 121122121\cdots$.

Then $\mathcal{A}_\mu$ has two states, labeled 1 and 2.

Since $\mu(1) = 121$, we add the transitions $(1,0,1)$, $(1,1,2)$ and $(1,2,1)$; since $\mu(2) = 122$, we add the transitions $(2,0,1)$, $(2,1,2)$ and $(2,2,2)$.



Abstract Numeration Systems

If we make the state associated with the letter a the initial state and all states terminal, we get a DFA over the alphabet $\{0, 1, \ldots, n-1\}$.

This DFA is complete if and only if the morphism μ is uniform, otherwise some transitions are not defined.

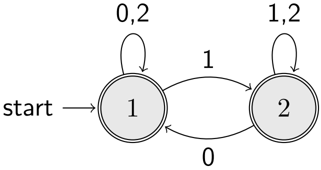


Figure: The DFA associated with the morphism $\mu : 1 \mapsto 121,\ 2 \mapsto 122$ starting from letter 1.

Abstract Numeration Systems

Let L be the language recognized by the DFA. Let L^0 be the language obtained from L after removing the words that start with 0.

Now, provided that we fix an order $<$ on $\Sigma = \{a_1, a_2, \ldots, a_\sigma\}$, if we feed $\mathcal{A}_\mu$ with the words of L^0 in genealogical order (i.e., first by length, then lexicographically on equal-length words), we can build an infinite word by writing, for the n-th word in L^0, call it w_n, the label of the state we reach by reading w_n on $\mathcal{A}_\mu$ from the initial state, called the output of n.

Definition 32

The triple $S = (L, \Sigma, <)$ is called the Abstract Numeration System generated by μ and $<$ w.r.t. the letter a. The word w_n is the S-representation of the natural number n.

Abstract Numeration Systems

For example, the first few words in the ANS generated by $\mu : 1 \mapsto 121,\ 2 \mapsto 122$ with the order $1 < 2$ w.r.t. the letter 1 are: $\varepsilon, 1, 2, 10, 11, 12, 20, 21, 22, 100, 101, 102, \ldots$

So that, for example, 100 is the S-representation of 9 and 102 is the S-representation of 11.

If we feed the automaton with these words, in this order, we generate the sequence $1, 2, 1, 1, 2, 2, 1, 2, 1, 1, \ldots$ which is the sequence of letters of the fixed point of μ, i.e., the word lnd_3.

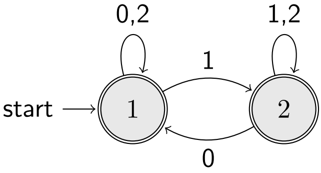
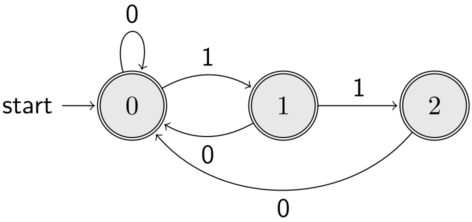


Figure: The DFA associated with the morphism $\mu : 1 \mapsto 121,\ 2 \mapsto 122$ starting from letter 1.

Abstract Numeration Systems

Actually, if μ is prolongable on a, letting $\mu(a) = aw$, we have that feeding the automaton $\mathcal{A}_\mu$ with all the words of L^0 up to length ℓ (in genealogical order), one gets precisely the sequence of letters of the word $a \cdot w \cdot \mu(w) \cdots \mu^{\ell-1}(w)$, i.e., the natural factorization of $\lim_{n\to\infty} \mu^n(a)$.

For example, let $\mu : 0 \mapsto 01, 1 \mapsto 02, 2 \mapsto 0$ be the morphism generating the Tribonacci word tr. The words of length up to 4 of the associated ANS (which are the S-representations of the first 13 natural numbers) are: $\varepsilon, 1, 10, 11, 100, 101, 110, 1000, 1001, 1010, 1011, 1100, 1101$; the associated outputs are: $0, 1, 0, 2, 0, 1, 0, 0, 1, 0, 2, 0, 1$, and we have $0 \cdot 1 \cdot 02 \cdot 010 \cdot 010201 = 0 \cdot 1 \cdot \mu(1) \cdot \mu^2(1) \cdot \mu^3(1)$, which is the prefix of length 13 of the Tribonacci word.



Abstract Numeration Systems

Finally, if we want to generate a generic morphic word, i.e., in addition to the morphism μ we have a coding, we just rename the states of $\mathcal{A}_\mu$ accordingly with the coding.

For example, adding the coding $\chi : 0, 2 \mapsto a, 1 \mapsto b$ to the Tribonacci morphism, we get the morphic word $\chi(\mu^\omega(0)) = abaaabaabaaab \cdots$ by just renaming the labels of the automaton accordingly with the coding.

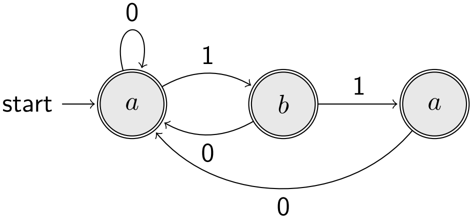


Figure: The DFA associated with the Tribonacci morphism $0 \mapsto 01, 1 \mapsto 02, 2 \mapsto 0$ and the coding $0, 2 \mapsto a, 1 \mapsto b$.

Abstract Numeration Systems

Remark 33

If the morphism μ is uniform of length k, the associated ANS is the standard base-k representation of natural numbers.

So we have an algorithm to construct the automaton associated with a k-automatic word.

Remark 34

Notice that the same morphic word can be obtained from different ANS.

Exercise 35

Build the automaton for the other morphisms shown in this lecture.

Rich Morphisms

Definition 36

A morphism is rich if it maps finite rich words to rich words.

Clearly, since letters, i.e., factors of length 1, are rich, every fixed point of a rich morphism is an infinite rich word.

An example of rich morphism is the Fibonacci morphism $0 \mapsto 01, 1 \mapsto 0$.

Another example of a rich morphism is the von Neumann morphism $0 \mapsto 001,\ 1 \mapsto 1$.

Lyndon Morphisms

Definition 37

Let Σ be a totally ordered alphabet. A morphism μ is order-preserving (or, more precisely, order-preserving on finite words) if for every $u, v \in \Sigma^*$ such that $u \leq v$, $\mu(u) \leq \mu(v)$.

An example of order-preserving morphism is the Thue–Morse morphism $0 \mapsto 01$, $1 \mapsto 10$.

Lyndon Morphisms

In the binary case, we have the following characterization.

Theorem 38

Let μ be an endomorphism of Σ_2. Then μ is order-preserving if and only if $\mu(01) \leq \mu(1)$.

So for example the Fibonacci morphism $\varphi : 0 \mapsto 01,\ 1 \mapsto 0$ is not order-preserving, since $\varphi(01) = 010$ is not smaller than $\varphi(1) = 0$.

Lyndon Morphisms

Definition 39

A (finite or infinite) word is called a Lyndon word if it is lexicographically smaller than all its proper suffixes.

In particular, finite Lyndon words are primitive words.

Definition 40

A morphism μ is a Lyndon morphism if it maps Lyndon words to Lyndon words.

The Thue–Morse morphism is not a Lyndon morphism. For example, it maps the Lyndon word 1 to the word 10, which is not Lyndon (or also 01 to 0110).

Lyndon Morphisms

Theorem 41 (Richomme, 2003)

Let Σ be a totally ordered alphabet of size at least 2. An endomorphism μ of Σ^* is Lyndon if and only if μ is order-preserving and $\mu(a)$ is a Lyndon word for every $a \in \Sigma$.

In the binary case, we have the following characterization.

Proposition 42

Let μ be an endomorphism of Σ_2. Then μ is a Lyndon morphism if and only if $\mu(0)$ and $\mu(1)$ are Lyndon words and $\mu(0) \leq \mu(1)$.

Lyndon Morphisms

We now deal with the question whether a morphism generates an infinite Lyndon word.

If μ is a Lyndon morphism prolongable on the letter a, where a is the smallest letter of Σ, then its fixed point $\lim_{n\to\infty} \mu^n(a)$ is an infinite Lyndon word, since $\mu^n(a)$ is a Lyndon word for every n.

This is, however, not a necessary condition: the morphism $0 \mapsto 010$, $1 \mapsto 11$ is not Lyndon (since 010 is not a Lyndon word) but generates an infinite Lyndon word.

Lemma 43

Let μ be an endomorphism of Σ_2 prolongable on 0. If μ generates an infinite Lyndon word, then μ is order-preserving.

The converse of the previous statement does not hold true, in general. For example, the Thue–Morse–Morse morphism is order-preserving, but its fixed point is not an infinite Lyndon word.

Lyndon Morphisms

Theorem 44 (Richomme, Séébold, 2021)

Let μ be an endomorphism of Σ_2 prolongable on 0. Then μ generates an infinite Lyndon word if and only if all the following conditions hold:

1. *μ is order-preserving;*
2. *the infinite word generated by μ is aperiodic;*
3. *$\mu^3(0)$ is a prefix of a Lyndon word.*

The necessity of the second condition is witnessed by $0 \mapsto 010$, $1 \mapsto 101$, which generates the periodic word $(01)^\omega$.

The necessity of the third condition is witnessed by the Fibonacci morphism $\varphi : 0 \mapsto 01$, $1 \mapsto 0$. Indeed, $\varphi^2(0) = 010$ is a prefix of a Lyndon word, but $\varphi^3(0) = 01001$ is not. In fact, the Fibonacci word f is not an infinite Lyndon word.

Toeplitz Words

The paperfolding word p is actually a regular instance of a more general construction, called Toeplitz construction, which is another way to construct infinite words.

A (general) paperfolding word can be seen as the sequence of ridges and valleys obtained by unfolding a sheet of paper which has been folded infinitely many times.

At each step, one can fold the paper in two different ways, thus generating uncountably many sequences.

The regular paperfolding word $p = 001001100011011\cdots$ is obtained by folding always along the same direction.

Toeplitz Words

Even if paperfolding words are not fixed points of primitive morphisms, so that we cannot apply Theorem 24, they are all linearly recurrent:

Theorem 45

All paperfolding words are linearly recurrent. More specifically, for any $k \geq 0$, any factor of length at least $44k$ contains all the factors of length k.

Toeplitz Words

Let us define $P_1 = 0?1?$ and $P_2 = 1?0?$.

A paperfolding word can be obtained starting from $T^0 = ?^\omega$ and defining for $n > 0$, the word T^n as the word obtained from T^{n-1} by replacing sequentially all occurrences of $?$ by the letters of P_1^ω or P_2^ω, depending on the sequence of instructions $b = b_0 b_1 \cdots$, where $b_i \in \{P_1, P_2\}$.

At the limit for n that goes to infinity, one obtains the paperfolding word associated with the sequence b.

Toeplitz Words

For example, if $b = P_1^\omega$ one obtains the regular paperfolding word p. The first few iterations of the Toeplitz construction of the regular paperfolding word p are:

$$\begin{array}{ccccccccccccccccc}
0 & ? & 1 & ? & 0 & ? & 1 & ? & 0 & ? & 1 & ? & 0 & ? & 1 & ? & \cdots \\
0 & 0 & 1 & ? & 0 & 1 & 1 & ? & 0 & 0 & 1 & ? & 0 & 1 & 1 & ? & \cdots \\
0 & 0 & 1 & 0 & 0 & 1 & 1 & ? & 0 & 0 & 1 & 1 & 0 & 1 & 1 & ? & \cdots \\
0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & ? & \cdots \\
0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & \cdots
\end{array}$$

Toeplitz Words

As another example, if $b = (P_1 P_2)^\omega$ one obtains the alternate paperfolding word

$$a = 011000110111001001100010011100110 \cdots$$

$$\begin{array}{ccccccccccccccccc}
0 & ? & 1 & ? & 0 & ? & 1 & ? & 0 & ? & 1 & ? & 0 & ? & 1 & ? & \cdots \\
0 & 1 & 1 & ? & 0 & 0 & 1 & ? & 0 & 1 & 1 & ? & 0 & 0 & 1 & ? & \cdots \\
0 & 1 & 1 & 0 & 0 & 0 & 1 & ? & 0 & 1 & 1 & 1 & 0 & 0 & 1 & ? & \cdots \\
0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & ? & \cdots \\
0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & \cdots
\end{array}$$

Toeplitz Words

Remark 46

There are uncountably many paperfolding words, only a countable subset of them is q-automatic.

As a consequence of Cobham's theorem (see Remark 23), a paperfolding word is q-automatic if and only if its sequence of instructions is ultimately periodic.

Toeplitz Words

More generally, a Toeplitz word is defined in the same way but taking as sequence of instructions any sequence (P_n) of words over $\Sigma \cup \{?\}$, called partial words.

Starting from $T^0 = ?^\omega$, one defines, for every $i > 0$, T^i as the word obtained from T^{i-1} by replacing sequentially all occurrences of $?$ by the letters of the infinite periodic word $(P_i)^\omega$.

If there are infinitely many values of n such that P_i does not begin with $?$, the associated Toeplitz word $T = \lim_{i\to\infty} T^i$ is a word over Σ.

We suppose the partial words P_i to be of minimal length (i.e., primitive).

Toeplitz Words

A special case is when the sequence (P_i) is constant, that is, always equal to a partial word P. We call these Toeplitz words simple and say that they are generated by the partial word P.

Note that this is also the case of a periodic sequence of partial words, which can be reduced to a single partial word.

For example, the regular paperfolding word is generated by $P = 0?1?$, so it is simple.

The alternate paperfolding word is also simple, since it can be generated by the partial word $P = 011?001?$.

Notice that every purely periodic word is a simple Toeplitz word.

Toeplitz Words

The period-doubling word d is generated by $P = 010?$:

$$\begin{array}{ccccccccccccccccc}
0 & 1 & 0 & ? & 0 & 1 & 0 & ? & 0 & 1 & 0 & ? & 0 & 1 & 0 & ? & \cdots \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & ? & \cdots
\end{array}$$

The word

$$lnd_3 = 12112212112112212212112212112112212112\cdots$$

fixed point of the morphism $1 \mapsto 121, 2 \mapsto 122$, is also a simple Toeplitz word, generated by $P = 12?$:

$$\begin{array}{ccccccccccccccccc}
1 & 2 & ? & 1 & 2 & ? & 1 & 2 & ? & 1 & 2 & ? & 1 & 2 & ? & 1 & \cdots \\
1 & 2 & 1 & 1 & 2 & 2 & 1 & 2 & ? & 1 & 2 & 1 & 1 & 2 & 2 & 1 & \cdots \\
1 & 2 & 1 & 1 & 2 & 2 & 1 & 2 & 1 & 1 & 2 & 1 & 1 & 2 & 2 & 1 & \cdots
\end{array}$$

Toeplitz Words

The following result is a direct consequence of the definition:

Theorem 47

Every simple Toeplitz word is uniformly recurrent.

We have the following immediate characterization of simple Toeplitz words:

Theorem 48

An infinite word x is a simple Toeplitz word if and only if for every n there exists p such that x is constant over the indices equal to n modulo p.

As a consequence, the Fibonacci word f, the Thue–Morse word t and the Rudin–Shapiro word r are not simple Toeplitz words, by known properties of the arithmetic progressions in these words.

Toeplitz Words

Simple Toeplitz words are all automatic words by the aforementioned result of Cobham. More specifically, we have:

Theorem 49 (Cassaigne, Karhumäki, 1997)

Let w be a simple Toeplitz word, generated by the partial word P, and let q be the number of $?$ occurring in P. Then:

1. *If $q = 1$, w is a fixed point of a uniform morphism of length q;*
2. *If q divides $|P|$, w is obtained from a fixed point of a uniform morphism of length q by applying a coding.*

Toeplitz Words

An interesting class of Toeplitz words is that of Stewart words. They are Toeplitz words generated by any sequence of patterns in $\mathcal{P}$, where $\mathcal{P}$ is the set of patterns of length 3 that are permutations of $\{0, 1, ?\}$, called *Stewart patterns*:

$$\begin{aligned} \mathtt{a} &= 01?; & \mathtt{b} &= 10?; \\ \mathtt{c} &= 0?1; & \mathtt{d} &= 1?0; \\ \mathtt{e} &= ?01; & \mathtt{f} &= ?10. \end{aligned}$$

For example:

- The pattern sequence $\mathtt{a}^\omega = \mathtt{aaa}\cdots$ specifies the word

  $$010011010010011011010011010010011010\cdots,$$

  the fixed point of the morphism $0 \mapsto 010, \ 1 \mapsto 011$, which is a recoding over $\{0, 1\}$ of the word lnd_3.
- The pattern sequence $\mathtt{c}^\omega = \mathtt{ccc}\cdots$ specifies the so-called Stewart choral word

  $$00100101100100101100101101100100101\cdots,$$

  the fixed point of the morphism $0 \mapsto 001, \ 1 \mapsto 011$.

Toeplitz Words

- The pattern sequence $(\mathtt{ab})^\omega = \mathtt{ababab}\cdots$ specifies the so-called Sierpiński gasket word

  $$011010010011010011011010011 01\cdots,$$

  the fixed point of the morphism $0 \mapsto 011,\ 1 \mapsto 010$.

- The word generated by $\mathtt{e}^\omega$ and in which the first letter is set to 0[1]

  $$0010011010010011011010011010\cdots$$

  has been considered by Ferenczi (so we call it Ferenczi word), who showed that applying to it the morphism $0 \mapsto 0, 1 \mapsto 10$, one obtains the Chacon word.

[1] In the case that a Stewart word is specified by an infinite word with a suffix in $\{\mathtt{e},\mathtt{f}\}^\omega$, and only in this case, the limit of the Toeplitz construction is an infinite word containing a single occurrence of ?. In this special case, there are two distinct Stewart words, obtained by replacing this single occurrence of ? with 0 and 1, respectively.

Toeplitz Words

Remark 50

The Stewart–Thue–Morse word

$$stm = 0101100101100110100101100101100110100101100\cdots$$

is the image, under the Thue–Morse morphism $\tau : 0 \mapsto 01,\ 1 \mapsto 10$, *of the Stewart choral word*

$$00100101100100101100101101100100101\cdots$$

Toeplitz Words

Theorem 51

The only possible palindromes occurring in the Stewart words are:

$$\{\varepsilon, 0, 1, ?, 00, 11, 010, 101, 0110, 1001, 00100,$$

$$11011, 010010, 101101, 0110110, 1001001\}.$$

Furthermore, each such palindrome occurs in the length-81 word specified by any Stewart pattern sequence of length 4.

Theorem 52

A Stewart word is 3-automatic if and only if its sequence of patterns is ultimately periodic.