Combinatorics on Words

Gabriele Fici

CWI, Amsterdam — April 2024

Part 3: Repetitions and Avoidability

Integer Powers

A word w that can be written as $w = v^k$ for some primitive word v and an integer $k \geq 2$ is called an integer repetition or integer power (of order k), and v is called the primitive root of the integer repetition.

A repetition of order 2 is simply called a square, while a repetition of order 3 is simply called a cube. A repetition of order k is called a k-power.

Fractional Powers

We can extend this definition to the non-integer case.

Definition

Let w be a word and suppose that v is a word-period of w. Then we can write $w = v^\beta$ for $\beta = |w|/|v| \geq 1$. We say that w is a fractional repetition (or a sesquipower), or more specifically a β-power.

If $|v|$ is the smallest positive period of w, the word v is called the fractional root of the fractional repetition.

For example, the word $w = 0010010$ is a $7/3$-power, and $v = 001$ is its fractional root.

The word $w = 01001010010$ has minimum positive period 5 and can be written as $(01001)^{11/5}$.

Remark

Every bordered word is a fractional repetition.

Exponent

Remark

A word like 0000 is both a square and a 4th-power, so the order of an integer repetition is not uniquely determined, unless we assume that the root is primitive.

For fractional repetitions the situation is similar: the exponent is not uniquely determined, unless we assume that the length of the root is the smallest positive period of the fractional repetition. For example, $w = 01001010010$ has periods 5 and 8 so it is both a $11/5$-power and a $11/8$-power.

So, unless otherwise specified, when we refer to the exponent of a fractional repetition, we refer to the ratio between its length and its minimum positive period.

Squares

Squares are the simplest kind of repetition. A natural question is how many distinct square factors a word can contain.

The following result was conjectured by Fraenkel and Simpson in 1998 and then proved by Brlek and Li in 2022:

Theorem

A word of length n contains less than n distinct square factors.

Squares

About the occurrences of squares in a word, we have the following result, due to Crochemore and Rytter:

Lemma (Three Squares Lemma)

Suppose u is primitive, and v is not a power of u. If u^2 is a prefix of v^2, in turn a proper prefix of w^2, then $|w| \geq |u| + |v|$.

The Fibonacci word demonstrates that this result is best possible.

Indeed, the square prefixes of the Fibonacci word are the squares of the Fibonacci finite words, starting from 010 (square prefixes ending at positions 6, 10, $16 = 6 + 10$, $26 = 10 + 16$, etc.): $(010)^2$, $(01001)^2$, $(01001010)^2$, $(0100101001001)^2$, etc.

Squares

Any sufficiently long binary word must contain a square factor.

Indeed, let x be a word over Σ_2. Suppose x starts with 0 (if x starts with 1 the reasoning is the same). If x does not contain squares, then the second letter of x must be 1. The third letter of x must therefore be 0, since otherwise x contains the square 11. But then any letter following 010 creates a square factor (either 0^2 or $(01)^2$).

In other words, we say that squares are unavoidable with two letters.

Squares

The Thue–Morse word contains infinitely many square factors.

In fact, Brlek proved that w^2 is a square factor of the Thue–Morse word if and only if w is of one of the following forms: $\tau^i(0)$, $\tau^i(1)$, $\tau^i(010)$, $\tau^i(101)$, for some $i \geq 0$.

So the square factors of the Thue–Morse word have length either a power of 2 or 3 times a power of 2.

Overlaps

What if we can use three letters then? Axel Thue in early 1900's proved that it is possible to construct arbitrarily long words over an alphabet with three or more letters with the property that no factor is a square. So squares are avoidable with three letters.

He also proved that there exist infinite binary words in which no factor has exponent larger than 2. An overlap is a word of the form $auaua$, for a (possibly empty) word u and a letter a. An overlap is therefore a word whose exponent is larger than 2, but since $auaua = (au)^{2+1/|au|}$, the exponent of an overlap can be smaller than $2 + \varepsilon$ for any real number ε.

Overlap-Freeness

A word is overlap-free if it avoids overlaps, that is, if none of its factors is an overlap.

Theorem

The Thue–Morse word t is overlap-free.

Therefore, any square in t is followed by the letter different from its first letter.

Overlap-Freeness

Séébold proved this remarkable result:

Theorem (Séébold, 1985)

The Thue–Morse word t and its complement $\bar{t}$ (which is the fixed point starting with 1 of the morphism $\tau : 0 \mapsto 01,\ 1 \mapsto 10$) are the only pure morphic overlap-free binary words.

The twisted Thue–Morse word $tt = 0100110100101100110100110010110 10\cdots$, although is not pure morphic, is overlap-free too.

Overlap-Freeness

The Thue–Morse word is a paradigmatic example of binary overlap-free word also in view of the following theorem of Restivo and Salemi:

Theorem (Restivo, Salemi, 1985)

Every binary overlap-free finite word is, up to removing one or two letters at the beginning and/or at the end, the image under the Thue–Morse morphism τ of another binary overlap-free word.

Moreover, this decomposition is unique if the length of the word is at least 7.

For example, consider the overlap-free word $w = 0010011$. Then, $w = 00\tau(10)1$.

The previous theorem holds true in the more general case of words avoiding $7/3$-powers.

Overlap-Freeness

There is also a version for infinite words:

Theorem

Every binary overlap-free infinite word x is equal to $u\tau(y)$ where $u \in \{\varepsilon, 0, 1, 00, 11\}$, τ is Thue–Morse morphism and y is another binary overlap-free word.

So for example, the twisted Thue–Morse word tt is equal to $0\tau(\overline{tt})$.

Overlap-Freeness

About larger alphabets, Allouche and Shallit proved the following result about generalized Thue–Morse words.

Theorem

Let $s_k(n)$ denote the sum of the digits in the base-k representation of n. The word $t_{k,m}$, defined by taking the sequence of sums of the digits in the base-k representation of n modulo m, is overlap-free if and only if $m \geq k$.

Moreover, $t_{k,m}$ always contains infinitely many squares, and contains infinitely many palindromes if and only if $m \leq 2$.

Morton and Mourant proved that $t_{k,m}$ is ultimately periodic if and only if m divides $k - 1$.

Square-Freeness

Let x be an infinite word over Σ_2. Define the word x' by $x'_n = 1 + x_{n+1} - x_n$. Notice that x' is a word over Σ_3.

Applying this transformation to the Thue–Morse word t one obtains the word

$$vtm = 2102012101202102012021012102012101202 10 \cdots$$

It can be proved that the word vtm is the word whose nth letter is the number of 1's between the nth and the $(n+1)$th occurrence of 0 in t.

We already mentioned that it is also the fixed point of the morphism $0 \mapsto 1,\ 1 \mapsto 20,\ 2 \mapsto 210$.

Square-Freeness

A similar word is the one whose nth letter is the number of 0's between the nth and the $(n+1)$th occurrence of 1 in t:

$$vtm' = 012021012102012021020121012021012102012\cdots$$

Clearly, since t does not contain 000 nor 111, the words vtm and vtm' are the same word up to exchanging 0 and 2. So, t' is generated by the morphism $0 \mapsto 012,\ 1 \mapsto 02,\ 2 \mapsto 1$.

Proposition

The words vtm and vtm' are square-free. Hence, there exist infinite ternary square-free words.

Proof.

We prove that the word vtm is square-free. If $w_1 w_2 \cdots w_n w_1 w_2 \cdots w_n$, $w_i \in \{0,1,2\}$, were a factor of vtm, then $01^{w_1}01^{w_2}0\cdots 1^{w_n}01^{w_1}01^{w_2}0\cdots 1^{w_n}0$ would be a factor of t, which contradicts the overlap-freeness of t. □

Square-Freeness

Remark

Taking either of vtm or vtm' modulo 2, one obtains the period-doubling word:

$$d = 0100010101000100010001010100010101000 10 \cdots$$

Instead, the complement of the period-doubling word

$$\overline{d} = 1011101010111011101110101011101010111 01 \cdots$$

can be obtained by taking the absolute value of the consecutive differences in the Thue–Morse word. Or, equivalently, by taking the consecutive sums modulo 2 in t.

Square-Freeness

Do ternary square-free words exist that are fixed point of a *uniform* morphism? The answer is yes: an example is the fixed point of the Leech morphism, which has length 13:

$$0 \mapsto 0121021201210$$
$$1 \mapsto 1202102012021$$
$$2 \mapsto 2010210120102$$

Avoiding Long Squares

Squares are unavoidable with two letters, and the Thue–Morse word contains infinitely many distinct squares. So, does an infinite word over Σ_2 exist with only a finite number of square factors? The answer to this question is positive.

For example, the paperfolding word contains only 8 distinct square factors: 0^2, $(001)^2$, $(00110)^2$, $(011)^2$, $(10)^2$, $(10011)^2$, 1^2, and $(110)^2$.

Carpi showed that the Oldenburger–Kolakoski word contains no square longer than 54.

Entringer, Jackson and Schatz proved that the image under the morphism

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

of any ternary square-free word is a binary word containing only the squares 0^2, 1^2, $(01)^2$, $(10)^2$ and $(11)^2$.

Avoiding Long Squares

Actually, it is possible to construct an infinite word over the alphabet Σ_2 containing as square factors only 0^2, 1^2 and $(01)^2$.

This is optimal both in number and in lengths of the squares.

An example of such a word is given by applying to any other ternary square-free word the morphism of Harju and Nowotka:

$$0 \mapsto 11100011001011000111001 0$$
$$1 \mapsto 11100010110001110010110001 0$$
$$2 \mapsto 11100011001011000101110010110 0.$$

On the contrary, rich words cannot avoid long square factors:

Proposition

A recurrent rich infinite word contains arbitrarily long squares.

k-Freeness

Definition

An infinite word x is said to be k-power-free, or k-free, if there exists an integer $k \geq 2$ such that for every finite factor u of x, one has that u^k is not a factor of x.

An infinite word x is said to be ω-free if for every finite factor u of x there exists an integer $k \geq 2$ (depending on u) such that u^k is not a factor of x.

Of course, if a word is k-free for some integer k, then it is ω-free, but the converse is not always true.

k-Freeness

Since an aperiodic uniformly recurrent word cannot contain powers of arbitrary order of the same factor, we have the following

Theorem

Every uniformly recurrent word is either purely periodic or ω-free.

k-Freeness

We already mentioned that fixed points of primitive morphisms are linearly recurrent. Brigitte Mossé proved the following result.

Theorem (Mossé, 1992)

Every fixed point of a primitive morphism is either purely periodic or k-free for some k.

k-Freeness

About Toeplitz words (recall that the paperfolding word p is generated by the partial word $P = 0?1?$; d is generated by $P = 010?$; lnd_3 is generated by $P = 12?$) we have the following beautiful result:

Theorem (Boccuto, Carpi, 2020)

Let x be a Toeplitz word generated by the sequence of partial words (P_n) of maximal length k. Then x is k-free.

Notice that the value of k in the previous statement is not always the smallest possible, as witnessed by the the paperfolding word p, which is generated by $P = 0?1?$ but is indeed 3-free and not only 4-free.

k-Free Morphisms

Definition

A k-free morphism is one that sends k-free words to k-free words.

An overlap-free morphism is defined analogously.

About square-free morphisms, Crochemore proved that:

1. A uniform morphism is square-free if and only if the images of square-free words of length 3 are square-free.
2. A morphism over Σ_3 is square-free if and only if the images of square-free words of length 5 are square-free;

k-Free Morphisms

Carpi proved that the Thue morphism

$$0 \mapsto 01201$$
$$1 \mapsto 020121$$
$$2 \mapsto 0212021$$

is the shortest (in terms of sum of the lengths of images) square-free morphism over three letters.

k-Free Morphisms

Notice that the morphism $0 \mapsto 012, 1 \mapsto 02, 2 \mapsto 1$ is not square-free, even if its fixed point vtm' is square-free. Indeed, applied to the square-free word 010, it produces the word 01202012, which contains the square 2020.

It was shown by Brandenburg that the smallest square-free uniform morphism has length 11. An example of such a morphism is given by

$$0 \mapsto 01021012102$$
$$1 \mapsto 01021202102$$
$$2 \mapsto 01210120212.$$

k-Free Morphisms

About endomorphisms of Σ_2, Karhumäki proved that a morphism μ prolongable on 0 has a cube-free fixed point if and only if $\mu^{10}(0)$ is cube-free.

For general k, we have the following result.

Theorem (Richomme, Wlazinski, 2007)

A uniform morphism μ on Σ is k-free for an integer $k \geq 3$ if and only if the images by μ of all k-free words of length at most $k|\Sigma| + k + 1$ are k-free.

If the morphism is not uniform, however, there is no finite test-set that can be defined if the alphabet size is greater than 2.

k-Free Morphisms

For overlap-free morphisms, already Thue in 1912 proved that the only overlap-free endomorphisms of Σ_2 are of the form τ^n or $E \circ \tau^n$, $n \geq 0$, where τ is the Thue–Morse morphism and E is the automorphism exchanging 0 and 1.

Berstel and Séébold proved an equivalent condition: An endomorphism μ of Σ_2 is overlap-free if and only if the word $\mu(01101001)$ is overlap-free.

k-Free Morphisms

For general alphabets, Séébold proved that the morphism $\mu : a \mapsto a(a+1)$ (modulo k) for every $a \in \Sigma_k$ is overlap-free if and only if k is odd (i.e., the cardinality of the alphabet is even).

As another construction, taking any permutation σ of Σ_k and defining the morphism that maps each letter i to the rotation of σ that starts with i, one obtains an overlap-free infinite word.

For example, the fixed point

$$013524135240352401524013240135401352135240352401524 0$$

of $0 \mapsto 013524,\ 1 \mapsto 135240,\ 2 \mapsto 240135,\ 3 \mapsto 352401,\ 4 \mapsto 401352,\ 5 \mapsto 524013$ is an overlap-free word over 6 letters.

k-Free Morphisms

In the case of uniform morphisms, Richomme and Wlazinski gave the following characterization:

Theorem (Richomme, Wlazinski, 2004)

A uniform endomorphism of Σ_k, $k \geq 3$, is overlap-free if and only if it is overlap-free on words of length up to $k+2$.

Critical Exponent

Cubes can be avoided over Σ_2, as the Thue–Morse word avoids overlaps.

Another example of binary cube-free word is the word

$$lnd_3 = 1211221211211221221211221211211221211211\cdots$$

fixed point of $1 \mapsto 121, 2 \mapsto 122$ and also simple Toeplitz word generated by $P = 12?$.

However, this word is not overlap-free (for example, it contains 12112112). In fact, this word contains $(3-\varepsilon)$-powers, for arbitrarily small ε. (This can be easily seen from its Toeplitz definition.)

The Fibonacci word f contains cubes
(e.g. $00101001 \cdot 00101001 \cdot 00101001$) but avoids 4th powers.

So the question one can ask is: What is the least real number β such that no factor of a given word has exponent larger than β? This value is called the critical exponent.

Critical Exponent

Definition

The critical exponent of the infinite word x is the real number

$$ce(x) = \sup\{\beta \mid v^\beta \in \mathrm{Fact}(x), v \neq \varepsilon\}.$$

For example:

- The critical exponent of the Thue–Morse word t is 2;
- The critical exponent of the period-doubling word d is 4;
- The critical exponent of the Rudin–Shapiro word rs is 4;
- All paperfolding words have the same critical exponent, which is equal to 3;
- All Stewart words have the same critical exponent, which is equal to 3.

Critical Exponent

The critical exponent needs not to be an integer. For example, Carpi proved that the critical exponent of the Oldenburger–Kolakoski word k is $8/3$.

Actually, the critical exponent needs not to be a rational number either. For example, Mignosi and Pirillo proved that the critical exponent of the Fibonacci word f is $2 + \varphi \approx 3.618$, where $\varphi = (1 + \sqrt{5})/2$ is the golden ratio.

The critical exponent of the Pell word is $3 + \sqrt{2} \approx 4.4142$.

Tan and Wen proved that the critical exponent of the Tribonacci word is $\chi \approx 3.1915$, real solution of $2x^3 - 12x^2 + 22x - 13 = 0$.

Notice that the critical exponent of an infinite word can be infinite. This is the case, for example, of any (ultimately) periodic word.

Critical Exponent

Definition

The asymptotic critical exponent of an infinite word x is the real number

$$\begin{aligned} ace(x) &= \limsup_{n\to\infty}\{\beta \mid v^\beta \in \mathrm{Fact}(x), |v| = n\} \\ &= \sup\{\beta \mid \exists \text{ arbitrarily long } v \in \mathrm{Fact}(x) \text{ s.t. } v^\beta \in \mathrm{Fact}(x)\}. \end{aligned}$$

Clearly, $ace(x) \leq ce(x)$ and the two coincide if $ce(x)$ is not rational.

The asymptotic critical exponent of the m-bonacci word is $2 + 1/(\lambda_m - 1)$ where $2 - \frac{1}{m} < \lambda_m < 1$ is the unique positive real root of the polynomial $x^m - x^{m-1} - \ldots - x - 1$. Hence, the asymptotic critical exponent of the m-bonacci word is greater than 3 but smaller than $3 + \frac{1}{m-1}$.

It has been recently proved by Lubomíra Dvořáková and Edita Pelantová that for m-bonacci words the asymptotic critical exponent and the critical exponent coincide.

Critical Exponent

Definition

Let β be a rational number. We say that an infinite word x is β-free if none of its factors has exponent equal to β; and β^+-free if none of its factors has exponent larger than β (but can have factors of exponent equal to β).

For example:

- The Thue–Morse word t is 2^+-free (i.e., overlap-free) but not 2-free;
- The paperfolding word p is 3^+-free but not 3-free (however, the only cubes it contains are 000 and 111);
- The Rudin–Shapiro word rs is 4^+-free but contains 4-powers.

On the other hand:

- The period-doubling word d is 4-free but contains $(4-\varepsilon)$-powers, for arbitrarily small ε;
- All Stewart words are cube-free but contain $(3-\varepsilon)$-powers, for arbitrarily small ε.

Critical Exponent

Now, a natural question is: Does *every* real number larger than 1 is the critical exponent of some word?

The following theorem was proved by Dalia Krieger and Jeffrey Shallit.

Theorem

The following statements hold:

1. *For every real number $\beta > 1$ there exists an infinite word over some alphabet whose critical exponent is β;*
2. *For every real number $\beta \geq 2$ there exists an infinite binary word whose critical exponent is β.*

Repetition Threshold

So, in order to avoid repetitions, the size of the alphabet matters. For example, no ternary word exists with critical exponent $3/2$. This leads us to the following definition.

Definition

For every $k \geq 2$, the repetition threshold $RT(k)$ is the minimum of the critical exponents of infinite words over Σ_k.

For example, we know that the Thue–Morse word avoids overlaps, so it avoids any word whose exponent is larger than 2. Since any infinite binary word contains squares, we have $RT(2) = 2$.

Repetition Threshold

Over Σ_3, we know that squares are avoidable. But what is then the smallest exponent that can be avoided? The exact answer is $7/4$.

Indeed, there are only finitely many ternary words avoiding $7/4$-powers, but it is possible to construct an infinite ternary word in which no factor has exponent larger than $7/4$, hence $RT(3) = 7/4$.

An example of $7/4^+$-free ternary word is the fixed point of the morphism

$$0 \mapsto 012021201021012102120210$$
$$1 \mapsto 120102012102120210201021$$
$$2 \mapsto 201210120210201021012102$$

Repetition Threshold

For general size of the alphabet, the following important theorem, whose statement was conjectured in 1972 by Françoise Dejean, has been finally proved in 2011 with the effort of many researchers.

Theorem (Threshold Theorem)

One has $RT(2) = 2$, $RT(3) = 7/4$, $RT(4) = 7/5$ *and, for every* $k > 4$, $RT(k) = k/(k-1)$.

Several people, including Dejean herself, proved the statement for small values of k. The breakthrough was a result of Carpi, who proved in 2007 that the conjecture holds true for every $k \geq 33$. The last cases were proved in 2011 by Currie and Rampersad and independently by Rao.

Avoiding Palindromes

About avoidability of other kinds of patterns, let us see what happens if one wants to avoid palindromes (of length > 1 of course).

The set of palindromic factors of an infinite word can be finite or infinite.

For example, the word $(01)^\omega$ contains arbitrarily long palindromic factors, whereas the word $(012)^\omega$ does not contain any palindromic factor of length greater than 1.

Avoiding Palindromes

But there also exist aperiodic words having finitely many palindromic factors. For example:

- the Chacon word does not contain palindromes of length ≥ 13;
- the Rudin–Shapiro word does not contain palindromes of length ≥ 15;
- any paperfolding word contains only 29 palindromes, the longest of which has length 13;
- any Stewart word contains only 15 palindromes, the longest of which has length 7.

So, a natural question arises here: what is the minimum number of palindromes that an infinite word must contain?

Avoiding Palindromes

The word $(012)^\omega$ contains 4 palindromes: $\varepsilon, 0, 1, 2$. However, an infinite word over Σ_2 must contain at least 9 palindromes. For example, the palindromes in $(001011)^\omega$ are: $\varepsilon, 0, 1, 00, 11, 010, 101, 0110, 1001$.

In the aperiodic case, we have the following result.

Theorem (Fici, Zamboni, 2013)

Every aperiodic word contains at least 5 palindromes.

Every aperiodic binary word contains at least 11 palindromes.

For example, the image of the Fibonacci word f under the morphism $\mu : 0 \mapsto 0,\ 1 \mapsto 12$, $\mu(f) = 0120012012001200120120012 \cdots$, contains only 5 palindromes, namely: $\varepsilon, 0, 1, 2$ and 00.

The 11 palindromes in $\psi(f)$, where $\psi : 0 \mapsto 0, 1 \mapsto 01101$ are $\varepsilon, 0, 1, 00, 11, 000, 010, 101, 0110, 1001, 10001$.

Avoiding Palindromes

The previous examples are not closed under reversal.

Berstel et al. exhibited a uniformly recurrent word over a four-letter alphabet closed under reversal and containing only 5 palindromes (the letters and the empty word):

$$01231023013210230123103201321 0\cdots$$

defined as the limit of the sequence defined by $U_0 = 01$ and $U_{n+1} = U_n 23 \widetilde{U_n}$.

An aperiodic binary word closed under reversal, instead, must contain at least 13 palindromes. An example is given by the limit of the sequence defined by $U_0 = 0100110100011001 0$ and for $n \geq 0$

$$\begin{cases} U_{2n+1} = U_{2n} 1100 U_{2n} \\ U_{2n+2} = U_{2n+1} 0011 U_{2n+1} \end{cases}$$

whose set of palindromes is $\{\varepsilon, 0, 00, 000, 00100, 001100, 010, 0110, 1, 10001, 1001, 101, 11\}$.

The Ruler Word

One can define words over an infinite alphabet, for example by taking $\Sigma = \mathbb{N}$.

An example of such a word is the infinite Zimin word, also called *ruler sequence*, or infini-bonacci word:

$$r_2 = 0102010301020104010201030102010501020\cdots$$

It can be defined as the limit of the sequence of finite words $z_0 = 0$, $z_n = z_{n-1}(n)z_{n-1}$, for $n > 0$.

The first few values of the sequence are: $z_0 = 0$, $z_1 = 010$, $z_2 = 0102010$, $z_3 = 01020103010201 0$, etc.

The words z_n are usually called Zimin words.

The Ruler Word

$$r_2 = 0102010301020104010201030102010501020\cdots$$

The word r_2 has several remarkable properties:

First of all, its n-th term is $\nu(n)$, the 2-adic valuation of n, i.e., the position from right to left of the last 1 in the binary representation of n. In other words, $\nu(n)$ is the largest integer k such that 2^k divides n.

The Ruler Word

$$r_2 = 0102010301020104010201030102010501020\cdots$$

But it is also the lexicographically least word avoiding squares.

In the word r_2, 0 occurs in every odd position (starting from 1). Deleting all the 0s in r_2, one obtains a word isomorphic to r_2 (actually, the word obtained mapping i to $i+1$ for all i), which is precisely the sequence of 2-adic valuations of $2n$, and is the lexicographically least word avoiding squares over the alphabet $\mathbb{N}_0 = \{1, 2, 3, \ldots\}$.

It is worth mentioning that Cobham proved the following

Theorem

Let x be a morphic word defined on the finite alphabet Σ. Let $V \subset \Sigma$ and x' be the word obtained from x by erasing all occurrences of the letters belonging to V. Then the word x' is either finite or morphic.

The Ruler Word

But what is the lexicographically least word avoiding squares over the alphabet Σ_3?

Problem

Characterize the lexicographically least word avoiding squares over the alphabet Σ_3.

The word exists since it can be constructed by the following algorithm: Start form 0102012; for every $n > 7$, define p_n as the prefix of length n of the least square-free word over $\{0, 1, 2\}$ of length $2n$. Then take the limit as n goes to infinity.

Indeed, it can be proved that if $xy \in \Sigma_3^*$ is square-free and $|y| = |x|$, then x can be extended to an infinite square-free word over Σ_3.

The first few letters are:

$$0102012021012010201202102010210120102012021012010210\cdots$$

The Ruler Word

On the other hand, the lexicographically least word avoiding cubes over the alphabet Σ_2 is:

$$0010010100100110010010100100110010010100101100100 1 \cdots$$

Again, one can prove that this word exists but no algorithm is known for constructing it.

Problem

Characterize the lexicographically least word avoiding cubes over the alphabet Σ_2.

The Ruler Word

More generally, if one takes the sequence of p-adic valuations (the largest integer k such that p^k divides n) of positive integers, then one obtains the lexicographically least word (over the alphabet $\mathbb{N}$) avoiding p-powers.

For example, the sequence of the 3-adic valuations of n:

$$r_3 = 0010010020010010020010010030010010020010010020010010030 0 \cdots$$

is the lexicographically least word avoiding cubes. It is isomorphic to the word obtained from it by removing all 0s, which is the lexicographically least word avoiding cubes over the infinite alphabet $\mathbb{N}_0 = \{1, 2, \ldots\}$.

The word r_3 is the fixed point of the (infinite) morphism

$$0 \mapsto 001,\ 1 \mapsto 002,\ \ldots, (n-1) \mapsto 00n,\ \ldots$$

The Ruler Word

The word r_3 is the fixed point of the (infinite) morphism

$$0 \mapsto 001,\ 1 \mapsto 002,\ \ldots, (n-1) \mapsto 00n,\ \ldots$$

For any $k \geq 2$, taking the word r_3 modulo k one obtains the fixed point of the morphism $0 \mapsto 001,\ 1 \mapsto 002,\ \ldots,\ (k-1) \mapsto 000$. For example, for $k = 2$, one obtains the word

$$0010010000100100001001001001000001001000001\cdots$$

whereas for $k = 3$, one obtains the word

$$0010010020010010020010010000010010020010010020\cdots$$

The Ruler Word

n	ternary	3-adic	mod 2	mod 3
1	1	0	0	0
2	2	0	0	0
3	10	1	1	1
4	11	0	0	0
5	12	0	0	0
6	20	1	1	1
7	21	0	0	0
8	22	0	0	0
9	100	2	0	2
10	101	0	0	0
11	102	0	0	0
12	110	1	1	1
13	111	0	0	0
14	112	0	0	0
15	120	1	1	1
16	121	0	0	0

The Ruler Word

The word r_2 is a Toeplitz word generated by the sequence of patterns $0?$, $1?$, $\ldots$, $n?$, $\ldots$

There is a similar word, that is generated by the sequence of patterns $0??$, $1??$, $\ldots$, $n??$, $\ldots$

$$r_2' = 012031042015031026014021037 0 \cdots$$

The word r_2' can also be obtained starting from 0120 and applying the 2-letter substitution $ab \mapsto 0(a+1)(b+1)$, for every $a, b \in \mathbb{N}$.

The Ruler Word

The word r_2 is also the fixed point of the (infinite) morphism

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

Recall that fixing an integer $m > 1$, the (primitive) morphism

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

generates the m-bonacci word.

This is why the word r_2 is sometimes called the infini-bonacci word.

The Ruler Word

The ruler word r_2 can also be obtained iterating the right palindromic closure operator.

Given an infinite directive sequence $a = (a_1, a_2, a_3, \ldots)$ one builds an infinite word as the limit of the sequences of words w_n defined by: $w_1 = a_1$ and for $n > 1$, w_n is the shortest palindrome that begins in $w_n a_n$. The directive sequence that generates the ruler word is $\mathbb{N}$.

Taking as directive sequence the natural numbers modulo 2, $a = (0, 1, 0, 1, 0, 1, 0, \ldots)$, then the right palindromic closure operator generates the Fibonacci word $01001010010\cdots$; taking as directive sequence the natural numbers modulo 3, $a = (0, 1, 2, 0, 1, 2, \ldots)$, then the right palindromic closure operator generates the Tribonacci word $01020100102010\cdots$, and so on, for every m-bonacci word.

The Ruler Word

Another way to construct the m-bonacci word, $m \geq 2$, is the following: Take the sequence of the 2-adic valuations of the positive integers whose binary representation does not contain 0^m.

n	binary	2-adic	Fibonacci	Tribonacci
1	1	0	0	0
2	10	1	1	1
3	11	0	0	0
4	100	2		2
5	101	0	0	0
6	110	1	1	1
7	111	0	0	0
8	1000	3		
9	1001	0		0
10	1010	1	1	1
11	1011	0	0	0
12	1100	2		2

The Ruler Word

Taking the ruler sequence r_2 modulo k, one obtains the fixed point of the morphism $0 \mapsto 01$, $1 \mapsto 02$, $\ldots$, $(k-1) \mapsto 00$, also called k-th generalized period-doubling word d_k.

The same word can be obtained as a Toeplitz word generated by the pattern $P = z_k?$, where z_k is the kth Zimin word.

For example, taking the word r_2 modulo 3, one obtains the generalized period-doubling word

$$d_3 = 0102010001020101020100010201020 1 \cdots$$

fixed point of the morphism $0 \mapsto 01$, $1 \mapsto 02$, $2 \mapsto 00$ and Toeplitz word generated by $01012010?$.

The Ruler Word

n	binary	2-adic	d	d_3
1	1	0	0	0
2	10	1	1	1
3	11	0	0	0
4	100	2	0	2
5	101	0	0	0
6	110	1	1	1
7	111	0	0	0
8	1000	3	1	0
9	1001	0	0	0
10	1010	1	1	1
11	1011	0	0	0
12	1100	2	0	2
13	1101	0	0	0
14	1110	1	1	1
15	1111	0	0	0

The Ruler Word

The period-doubling word d, fixed point of $0 \mapsto 01,\ 1 \mapsto 00$ is also the Toeplitz word generated by $010?$.

The period-doubling word d is also the sequence whose n-th element is the parity of the 2-adic valuation of n; whereas its binary complement

$$\overline{d} = 10111010101110111011010101110\cdots$$

is the sequence whose n-th element is the parity of the 2-adic valuation of $2n$.

Equivalently, $\overline{d}$ can be obtained from d by erasing every other term. We have seen that $\overline{d}$ is also equal to und_4 modulo 2.

The Ruler Word

By the way, there are other binary words with the property that erasing every other term one obtains the complement of the original word.

One example is the Thue–Morse word $t = 0110100110010110\cdots$.

Indeed, we can write t as $t \amalg \bar{t}$, i.e., t is the perfect shuffle[1] of itself with its complement:

$$0\,\mathbf{1}\,1\,\mathbf{0}\,1\,\mathbf{0}\,0\,\mathbf{1}\,1\,\mathbf{0}\,0\,\mathbf{1}\,0\,\mathbf{1}\,1\,\mathbf{0}\ldots$$

[1] The perfect shuffle of two words $a_1a_2a_3\cdots$ and $b_1b_2b_3\cdots$ is the word $a_1b_1a_2b_2a_3b_3\cdots$

The Ruler Word

There are other words that can be written as perfect shuffles. For example, we have seen that the ruler word r_2 can be written as $r_2 = 0^\omega \sqcup\!\sqcup (r_2 + 1)$:

$$0\,\mathbf{1}\,0\,\mathbf{2}\,0\,\mathbf{1}\,0\,\mathbf{3}\,0\,\mathbf{1}\,0\,\mathbf{2}\,0\,\mathbf{1}\,0\,\mathbf{4}\,0\,\mathbf{1}\,0\,\mathbf{2}\,0\,\mathbf{1}\,0\,\mathbf{3}\,0\,\mathbf{1}\,0\,\mathbf{2}\,0\,\mathbf{1}\,0\,\mathbf{5}\cdots$$

The Ruler Word

We also have:

- the period-doubling word: $d = 0^\omega \sqcup\!\sqcup \overline{d}$

  $0\,\mathbf{1}\,0\,\mathbf{0}\,0\,\mathbf{1}\,0\,\mathbf{1}\,0\,\mathbf{1}\,0\,\mathbf{0}\,0\,\mathbf{1}\,0\,\mathbf{0}\,0\,\mathbf{1}\,0\,\mathbf{0}\,0\,\mathbf{1}\,0\mathbf{1}\,0\,\mathbf{1}\,0\,\mathbf{0}\,0\,\mathbf{1}\,0\,\mathbf{0}\,\cdots$

- the regular paperfolding word: $p = (01)^\omega \sqcup\!\sqcup p$

  $0\,\mathbf{0}\,1\,\mathbf{0}\,0\,\mathbf{1}\,1\,\mathbf{0}\,0\,\mathbf{0}\,1\,\mathbf{1}\,0\,\mathbf{1}\,1\,\mathbf{0}\,0\,\mathbf{0}\,1\,\mathbf{0}\,0\,\mathbf{1}\,1\,\mathbf{1}\,0\,\mathbf{0}\,1\,\mathbf{1}\,0\mathbf{1}\,1\,\mathbf{0}\,\cdots$

- the alternate paperfolding word: $a = (01)^\omega \sqcup\!\sqcup \overline{a}$

  $0\,\mathbf{1}\,1\,\mathbf{0}\,0\,\mathbf{0}\,1\,\mathbf{1}\,0\,\mathbf{1}\,1\,\mathbf{1}\,0\,\mathbf{0}\,1\,\mathbf{0}\,0\,\mathbf{1}\,1\,\mathbf{0}\,0\,\mathbf{0}\,1\,\mathbf{0}\,0\,\mathbf{1}\,1\,\mathbf{1}\,0\,\mathbf{0}\,1\,\mathbf{1}\cdots$

The Ruler Word

Remark

Let x and y be infinite words. Then $y = x \sqcup\!\sqcup \overline{x}$ if and only if $y = \tau(x)$, where τ is the Thue–Morse morphism.

So for example, we already mentioned that the Stewart–Thue–Morse word $stm = 01011001011001\cdots$ is the perfect shuffling of the Stewart choral word $st = 001001011\cdots$, fixed point of the morphism $0 \mapsto 001,\ 1 \mapsto 011$, with its complement $\overline{st}$.

While one has that $y = 0^\omega \sqcup\!\sqcup \overline{x}$ if and only if $y = \delta(x)$, where δ is the period-doubling morphism.

The Ruler Word

Another word that is related to the 2-adic valuation of n is the von Neumann word

$$v = 0010011001001110010011001001111001001100 1 \cdots$$

Indeed,

$$v = \prod_{n \geq 1} 01^{\nu(n)} = 01^0 01^1 01^0 01^2 01^0 01^1 01^0 01^3 01^0 01^1 01^0 01^2 01^0 01^1 01^0 01^4 \cdots$$

In other words, the sequence obtained from v by taking the number of 1's between two consecutive 0's is r_2.

By comparison, recall that the word vtm is the word whose nth letter is the number of 1's between two consecutive 0's in the Thue–Morse word t.

Sums of Blocks of Length 2

The Thue–Morse word is a concatenation of blocks of the form 01 or 10

$$t = 01 \cdot 10 \cdot 10 \cdot 01 \cdot 10 \cdot 01 \cdot 01 \cdots$$

that alternate exactly as 0 and 1 alternate in t.

If we take the sequence of the sums of these blocks we obtain of course the periodic word $1^\omega = 111\cdots$

However, if we take the sequence of sums of blocks of length 2 skipping the first letter (i.e., the sequence $t(2i) + t(2i+1)$, $i > 0$) we obtain the word

$$vtm = 2102012101202102012021012102012101202101210201 20\cdots$$

which, taken modulo 2, is the period-doubling word

$$d = 0100010101000100010001010100010101000 10\cdots$$

Sums of Blocks of Length 2

Now, if we take the sequence of sums of blocks of length 2 in the Rudin–Shapiro word $rs = 00010010000011\cdots$, always skipping the first letter (i.e., the sequence $rs(2i) + rs(2i+1)$, $i > 0$) we obtain the word:

$$01100211011120100110021221 1\cdots$$

which, taken modulo 2, is nothing else than the alternate paperfolding word

$$a = 01100011011100100110001001 1100110\cdots$$

Sums of Blocks of Length 2

		sum	mod 2		sum	mod 2
	Thue–Morse	vtm	d	Rudin–Shapiro		a
1	1			0		
10	1	2	0	0	0	0
11	0			1		
100	1	1	1	0	1	1
101	0			0		
110	0	0	0	1	1	1
111	1			0		
1000	1	2	0	0	0	0
1001	0			0		
1010	0	0	0	0	0	0
1011	1			1		
1100	0	1	1	1	2	0
1101	1			1		
1110	1	2	0	0	1	1
1111	0			1		
10000	1	1	1	0	1	1

Table: The sequence of sums of blocks of two letters in the Thue–Morse word and in the Rudin–Shapiro word.

Runlength Encoding and Self-Generating Sets

Another interesting remark about the Thue–Morse word $t = 01101001100101101\cdots$. Writing the runlength encoding of t one gets the word

$$\Delta(t) = 12112221121121122211221 1\cdots$$

(which can also be defined as the fixed point of $1 \mapsto 121,\ 2 \mapsto 12221$) whose partial sums $1, 3, 4, 5, 7, 9, 11, 12, 13, 15, 16, 17, 19, 20\ldots$ form the smallest set S of positive integers (for the lexicographic order) such that $n \in S$ if and only if $2n \notin S$.

The set S can be constructed by a min-excluded algorithm:

S	1	3	4	5	7	9	11	12	13
$\overline{S}$	2	6	8	10	14	18			

Runlength Encoding and Self-Generating Sets

The characteristic sequence of S is precisely the word $\overline{d} = 10111010101110110\cdots$ whose n-th element is the parity of the 2-adic valuation of $2n$.

Recall that $\overline{d}$ is also the complement of the period-doubling word $d = 01000101010001000 1\cdots$ which is the sequence of the parity of 1s between two consecutive 0s (or, equivalently, the parity of 0s between two consecutive 1s) in the Thue–Morse word.

The word d (resp., $\overline{d}$) can be obtained from the word $\Delta(t)$, the runlength encoding of the Thue–Morse word, by applying the morphism $1 \mapsto 0,\ 2 \mapsto 10$ (resp., $1 \mapsto 1,\ 2 \mapsto 01$).

Runlength Encoding and Self-Generating Sets

Le us now take the set defined by: $0 \in S$; if $x \in S$ then $3x \in S$ and $3x + 1 \in S$.

S	0	1	3	4	9	10	12	13	27	28	30
$\overline{S}$	2	5	6	7	8						

The integers in S are precisely those whose ternary expansion does not contain 2.

The sequence of elements of S taken modulo 2 is precisely the Thue–Morse word t.

Runlength Encoding and Self-Generating Sets

Another self-generating set is the following: Let T be defined by the rules: $1, 2 \in T$; if $x \in T$ then $2x + 1 \in T$ and $4x + 2 \in T$.

T	1	2	3	5	6	7	10	11	13	14	15
$\overline{T}$	4	8	9	12							

The sequence of elements in T taken modulo 2 forms the word $\overline{f} = 1011010110110\cdots$, which is the binary complement of the Fibonacci word $f = 0100101001001\cdots$

Indeed, the elements of T are the positive integers whose binary representation does not contain 00.

In fact, the Fibonacci word can be obtained taking the last digit of the binary representation of positive integers whose binary representation does not contain 00 (or by taking the last digit of the binary representation of nonegative integers whose binary representation does not contain 11).

Runlength Encoding and Self-Generating Sets

n	binary	2-adic	Fibonacci	Tribonacci
1	1	0	0	0
2	10	1	1	1
3	11	0	0	0
4	100	2		2
5	101	0	0	0
6	110	1	1	1
7	111	0	0	0
8	1000	3		
9	1001	0		0
10	1010	1	1	1
11	1011	0	0	0
12	1100	2		2
13	1101	0	0	0
14	1110	1	1	1
15	1111	0	0	0
16	10000	4		