UNIVERSITÀ
di **VERONA**
Dipartimento
di **INFORMATICA**

On Infinite Prefix Normal Words

Ferdinando Cicalese, Zsuzsanna Lipták and ***Massimiliano Rossi***

Department of Computer Science – University of Verona

SOFSEM 2019

27-30 Jan 2019, Nový Smokovec

Prefix Normal Words

Definition

Prefix Normal Words

binary words

no factor (substring) has more 1s than the prefix of the same length

Fici, Lipták, "On prefix normal words". DLT ***[2011]***

$w = 110000000000$ prefix normal

$u = 110100111110$ not prefix normal

111 has more 1s than **110**

$v = 110100110110$ not prefix normal

11011 has more 1s than **11010**

Binary jumbled indexing problem

Given a binary string

100101001011011110010

Build an INDEX to answer the following type of queries:

Is there a substring of length 8 that contains:

3 0s, 5 1s?

Burcsi, Cicalese, Fici, Lipták, "Algorithms for jumbled pattern matching in strings". *IJFund.Comp.Sci.* ***[2012]***

Moosa, Rahman, "Sub-quadratic time and linear space data structures for permutation matching in binary strings". *J. Discr. Algorithms* ***[2012]***

Amir, Chan, Lewenstein, Lewenstein, "On Hardness of Jumbled Indexing". *ICALP* ***[2014]***

Chan, Lewenstein, "Clustered integer 3SUM via additive combinatorics". *STOC* ***[2015]***

Gagie, Hermelin, Landau, Weimann, "Binary jumbled pattern matching on trees and tree-like structures". *Algorithmica* ***[2015]***

Amir, Apostolico, Hirst, Landau, Lewenstein, Rozenberg, "Algorithms for jumbled indexing, jumbled border and jumbled square on run-length encoded strings". *TCS* ***[2016]***

Cunha, Dantas, Gagie, Wittler, Kowada, Stoye, "Fast and simple jumbled indexing for binary run-length encoded strings". *CPM* ***[2017]***

Kociumaka, Radoszewski, Rytter, "Efficient indexes for jumbled pattern matching with constant-sized alphabet". *Algorithmica* ***[2017]***

Binary jumbled indexing problem

Given a binary string

1001010010110111110010

Build an INDEX to answer the following type of queries:

Is there a substring of length 8 that contains:

3 0s, 5 1s?

Several applications in bioinformatics e.g.,

Protein identification with mass spectrometry and gene clusters.

Prefix normal words can be used as an index for the binary jumbled indexing problem.

This paper: Infinite prefix normal words

Infinite prefix normal words

Definition

Infinite Prefix Normal Words

infinite binary words

no factor has more 1s than the prefix of the same length

*Cicalese, Lipták, Rossi, "Bubble-Flip–A new generation algorithm for prefix normal words". LATA **[2018]** TCS **[2018]***

An infinite word x can be:

- Periodic – if x can be written as $x = u^\omega$, where u is a finite string.

$u = 1100$ $\quad x = u^\omega = 1100110011001100110011001100\cdots$ p.n.

$u' = 1101$ $\quad x = u'^\omega = 1101110111011101110111011101\cdots$ not p.n.

- Ultimately periodic – if x can be written as $x = vu^\omega$.

$v = 1110$ $\quad x = vu^\omega = 1110110011001100110011001100\cdots$ p.n.

$v' = 1000$ $\quad x = v'u^\omega = 1000110011001100110011001100\cdots$ not p.n.

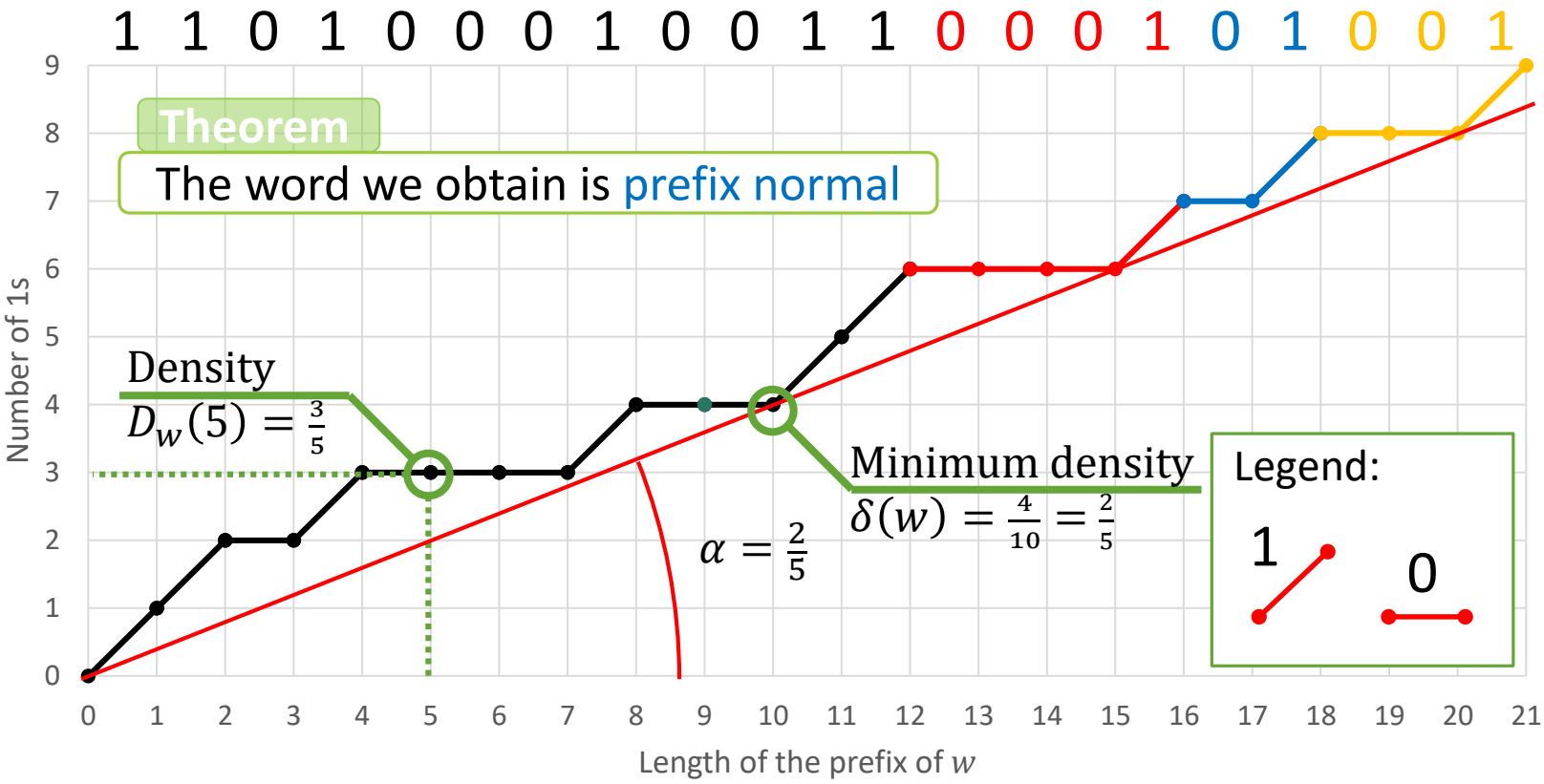
- Aperiodic – if x is not periodic or ultimately periodic.

$x = \lim_{n\to\infty} 1010^2\cdots10^n = 10100100010000100000100000\cdots$ p.n.

$x = 01101001100101101001011001101001100101100\cdots$ not p.n.

Extension of prefix normal words

Given $w = 110100010011$



We extend w with the maximum number of 0s followed by a 1 such that the word we obtain has minimum density equals to $\delta(w)$.

Sturmian words

Definition

Sturmian Words

infinite aperiodic binary words

exactly n+1 factors of length n, for all n.

They have the lowest factor complexity for an aperiodic word.

Widely studied.

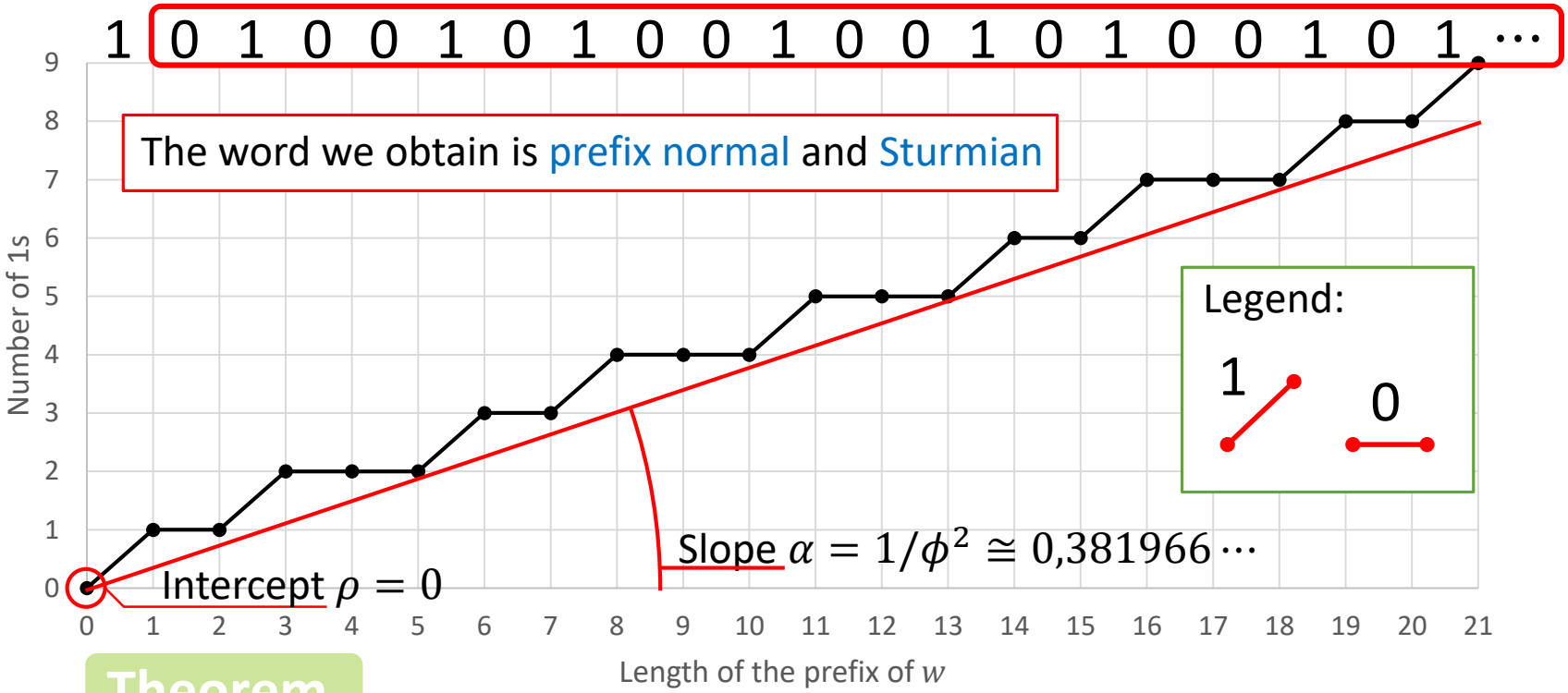
Several equivalent characterizations.

Theorem

Given an infinite binary word w, then w is Sturmian if and only if w is irrational mechanical.

Mechanical words

Fibonacci word



Theorem

Given an irrational slope α then there exists exactly one Sturmian word with slope α which is prefix normal, the one with intercept 0 and first character 1. ("Upper mechanical sequence")

The Thue-Morse word

$$T_0 = 0$$

$$T_n = T_{n-1}\overline{T_{n-1}} \quad \text{for all } n \geq 1$$

n	T_n
0	0
1	01
2	0110
3	01101001
4	0110100110010110

$$\mathfrak{tm} = \lim_{i \to \infty} T_i = 01101001100101101001011001 10 \cdots$$

Abelian complexity

$\mathfrak{tm} = 01101001100101101001011001 10 \cdots$

For each substring u of $\mathfrak{tm}$ we write the pair $(|u|_0, |u|_1)$.

E.g.: 011 $\mapsto$ (1,2)

$|u|_0$: no. of 0s in u
$|u|_1$: no. of 1s in u

List them for all lengths

n	1	2	3	4	5	6	7	8
	(1,0)	(2,0)	(2,1)	(3,1)	(3,2)	(4,2)	(4,3)	(5,3)
	(0,1)	(1,1)	(1,2)	(2,2)	(2,3)	(3,3)	(3,4)	(4,4)
		(0,2)		(1,3)		(2,4)		(3,5)
$\psi_{\mathfrak{tm}}(n)$	2	3	2	3	2	3	2	3

$\psi_{\mathfrak{tm}}(n)$: no. of pairs $(|u|_0, |u|_1)$ of substrings u of length n.

Theorem

$$\psi_{\mathfrak{tm}}(n) = \begin{cases} 2 & \text{if } n \text{ is odd} \\ 3 & \text{if } n \text{ is even} \end{cases}$$

*Richomme, Saari, Zamboni, "*Abelian complexity of minimal subshift*". J. London Math. Soc.* ***[2011]***

Abelian complexity

*Richomme, Saari, Zamboni, "*Abelian complexity of minimal subshift*". J. London Math. Soc.* ***[2011]***

*Madill, Rampersad, "*The abelian complexity of the paperfolding word*". Discrete Math.* ***[2013]***

*Blanchet-Sadri, Rampersad, "*On the asymptotic abelian complexity of morphic words*". Adv. Appl. Math.* ***[2014]***

*Turek, "*Abelian complexity function of the Tribonacci word*". J. Integer Seq.* ***[2015]***

*Cassigne, Kaboré, "*Abelian complexity and frequencies of letters in infinite words*". IJFund.Comp.Sci.* ***[2016]***

*Kaboré, Kientéga "*Abelian complexity of the Thue-Morse word over a ternary alphabet*". WORDS* ***[2017]***

Prefix normal forms

Theorem

Given an infinite binary word w, there exists two infinite prefix normal words s and t such that

$$\psi_w(n) = |s_1 \cdots s_n|_1 - |\overline{t_1 \cdots t_n}|_1 + 1$$

We call s and $\bar{t}$ the prefix normal forms of w.

$$\mathfrak{tm} = 01101001100101101001011001 10 \cdots$$

$$s = \overbrace{1111}^{4\ 1\text{s}}0101010101010101010101 \cdots$$

$$\bar{t} = \overbrace{111}^{3\ 1\text{s}}0010101010101010101010 \cdots$$

From abelian complexity to prefix normal forms

Theorem

Given a binary word w, if for all substrings u of w, it holds that $\overline{u}$ or $\overline{u^{rev}}$ is a substring of w, then

$$s = t \text{ and } |s_1 \cdots s_n|_1 = (\psi_w(n) + n - 1)/2$$

$$\mathfrak{tm} = 01101001100101101001011001 10 \cdots$$

$$\psi_{\mathfrak{tm}}(n) = \begin{cases} 2 & \text{if } n \text{ is odd} \\ 3 & \text{if } n \text{ is even} \end{cases} \qquad |s_1 \cdots s_n|_1 = \frac{\psi_{\mathfrak{tm}}(n)+n-1}{2} = \begin{cases} \frac{1+n}{2} & \text{if } n \text{ is odd} \\ \frac{2+n}{2} & \text{if } n \text{ is even} \end{cases}$$

n	1	2	3	4	5	6	7	8	$\cdots$
$\lvert s_1 \cdots s_n \rvert_1$	1	2	2	3	3	4	4	5	$\cdots$
s	1	1	0	1	0	1	0	1	$\cdots$
$\bar{t}$	0	0	1	0	1	0	1	0	$\cdots$

Summary

Full charachterization of prefix normal Sturmian words.

Connections between prefix normal forms and abelian complexity.

In the paper we also present:

Connections between prefix normal words and lexicographic order (e.g. infinite Lyndon words, Pirillo's max-min words)

THANK YOU!