Infinite smooth Lyndon words

Geneviève Paquin

somewhere between

LaCIM - Laboratoire de Combinatoire et d'Informatique Mathématique
UQAM - Université du Québec à Montréal

and

LAMA - Laboratoire de mathématiques
Université de Savoie

web : http://www.labmath.uqam.ca/~paquin/
email : genevievepaquin@gmail.com

The Prague Stringology Conference
September 2nd 2008

Motivation

- Lyndon words : class of words having lexicographical order properties.
- Smooth words : class of words, related to the Kolakoski word, that can be easily compressed.
- Some infinite smooth words are also Lyndon words.
- Is there other infinite smooth Lyndon words ?

Outline

- Notation
- Lyndon words
- Smooth words
- A characterization of smooth Lyndon words
- Idea of the proof
- Open problems

Notation

- $\mathcal{A} = \{a_0, a_1, \ldots, a_{k-1}\}$: a finite k-letter alphabet
- $\mathcal{A}^*$: set of finite words $w = w[0]w[1]\cdots w[n-1]$, $w[i] \in \mathcal{A}$
- $|w|$: length of w
- ε : the empty word, of length 0
- $\mathcal{A}^\omega$: (right-)infinite word over $\mathcal{A}$
- Let $w = pfs$ be a word. Then f is factor of w (if $p = \varepsilon$, f is prefix and if $s = \varepsilon$, f is suffix)

Lyndon words

Finite Lyndon word

$u \in \mathcal{A}^*$ is a *Lyndon word* if $u < v$ for all proper non-empty suffixes v of u. We write $u \in \mathcal{L}$.

Examples

$u = 112$ is a Lyndon word.
$v = 12112$ is not since $112 < 12112$.
$w = 112112$ is not since $112 < 112112$.

A word of length 1 is a Lyndon word.

Lyndon factorization (1/3)

Lyndon factorization of finite words [Chen, Fox, Lyndon - 1958]

Any non empty finite word w is uniquely expressed as a non increasing product (concatenation) of Lyndon words

$$w = \ell_0\ell_1\cdots\ell_n = \bigodot_{i=0}^{n} \ell_i, \text{ with } \quad \ell_i \in \mathcal{L}, \text{ and } \quad \ell_0 \geq \ell_1 \geq \cdots \geq \ell_n.$$

Duval (1983) described a linear algorithm that computes the Lyndon factorization of a Lyndon word of length n in $\mathcal{O}(n)$.

Example

$u = 1212112113112 1113$ has the Lyndon factorization

$$u = 12 \cdot 12 \cdot 112113 \cdot 112 \cdot 1113.$$

Indeed, $12 \geq 12 \geq 112113 \geq 112 \geq 1113$ and $12, 112113, 112, 1113$ are all Lyndon words.

Lyndon factorization (2/3)

Infinite Lyndon words [Siromoney et al - 1994]

The set $\mathcal{L}_\infty$ of *infinite Lyndon words* consists of infinite words smaller than any of their suffixes.

Theorem [Siromoney et al - 1994]

Any infinite word w is uniquely expressed as a non increasing product of Lyndon words, finite or infinite, in one of the two following forms :

i) either $w = \ell_0\ell_1\ell_2\cdots$ and for all k, $\ell_k \geq \ell_{k+1}$, with $\ell_i \in \mathcal{L}$,

ii) $w = \ell_0\ell_1\cdots\ell_m\ell_{m+1}$ and $\ell_0 \geq \cdots \geq \ell_m > \ell_{m+1}$,with $\ell_i \in \mathcal{L}$ for $0 \leq i \leq m$, $\ell_{m+1} \in \mathcal{L}_\infty$.

Lyndon factorization (3/3)

Examples

1) $u = 1211211121^{\omega} = 12 \cdot 112 \cdot 1112 \cdot 1 \cdot 1 \cdot 1 \cdots$.
2) $v = 3212^{\omega} = 3 \cdot 2 \cdot 12^{\omega}$.
3) $w = 121221222122221222222 \cdots$ is an infinite Lyndon word.

In what follows, we will only consider words over an ordered 2-letter alphabet.

Run-length encoding (1/3)

Every word $w \in \{a < b\}^\omega$ can be uniquely written as

$$w = a^{i_0} b^{i_1} a^{i_2} \cdots \text{ or } w = b^{i_0} a^{i_1} b^{i_2} \cdots , i_k \geq 1 .$$

Example

$w = 1221112121112 \cdots = 1^1 2^2 1^3 2^1 1^1 2^1 1^3 \cdots$

Definition

We define $\Delta : \mathcal{A}^\omega \longrightarrow \mathbb{N}^\omega$, by $\Delta(w) = i_1, i_2, i_3, \cdots$

Example

$\Delta(w) = [1, 2, 3, 1, 1, 1, 3, \cdots]$ which is usually written as

$$\Delta(w) = 1231113 \cdots$$

The operator Δ may be iterated until the coding alphabet changes.

Run-length encoding (2/3)

Example

Let $u = 11211212212112\cdots$
$\Delta^0(u) = 11211212212112\cdots$
$\Delta^1(u) = 212112112\cdots$
$\Delta^2(u) = 111212\cdots$
$\Delta^3(u) = 311\cdots$

Example

Let $v = 1121122121121221121121221211221221121121\cdots$
$\Delta^0(v) = \mathbf{1}121122121121221121121221211221221121121\cdots$
$\Delta^1(v) = \mathbf{2}12211211221211211221222121\cdots$
$\Delta^2(v) = \mathbf{1}122122112122121 1\cdots$
$\Delta^3(v) = \mathbf{2}212211211\cdots$
$\Delta^4(v) = \mathbf{2}1221\cdots$

Run-length encoding (3/3)

Example of the Kolakoski word

$$K_{(2,1)} = \underbrace{22}_{2}\,\underbrace{11}_{2}\,\underbrace{2}_{1}\,\underbrace{1}_{1}\,\underbrace{22}_{2}\,\underbrace{1}_{1}\,\underbrace{22}_{2}\,\underbrace{11}_{2}\cdots$$

$$K_{(1,2)} = \underbrace{1}_{1}\,\underbrace{22}_{2}\,\underbrace{11}_{2}\,\underbrace{2}_{1}\,\underbrace{1}_{1}\,\underbrace{22}_{2}\,\underbrace{1}_{1}\,\underbrace{22}_{2}\,\underbrace{11}_{2}\cdots = 1K_{(2,1)}$$

Δ has 2 fixpoints, since $\Delta(K_{(2,1)}) = K_{(2,1)}$ and $\Delta(K_{(1,2)}) = K_{(1,2)}$.

Smooth words (1/2)

Definition

The set of smooth words over $\mathcal{A} = \{a < b\}$, $a, b \in \mathbb{N}$, is defined as

$$\mathcal{K}_{\mathcal{A}} = \{w \in \mathcal{A}^\omega \mid \forall k \in \mathbb{N}, \Delta^k(w) \in \mathcal{A}^\omega\}.$$

Example

Over the alphabet $\mathcal{A} = \{1, 3\}$, $v = 1113111333131\cdots$ is not smooth, since $\Delta(v) = 313311\cdots$ and $\Delta^2(v) = 11\mathbf{2}\cdots$.

Definition

The bijection $\Phi : \mathcal{K}_{\mathcal{A}} \to \mathcal{A}^\omega$ is defined by

$$\Phi(w)[j] = \Delta^j(w)[0] \text{ for } j \geq 0$$

Example

If $v = 11211221211212211211212212112 2\cdots$, then $\Phi(v) = 12122\cdots$.

Smooth words (2/2)

Since Δ is not bijective, we define the pseudo-inverse functions

$$\Delta_a^{-1}, \Delta_b^{-1} : \mathcal{A}^\omega \longrightarrow \mathcal{A}^\omega$$

by

$$\Delta_\alpha^{-1}(u) = \alpha^{u[0]}\overline{\alpha}^{u[1]}\alpha^{u[2]}\overline{\alpha}^{u[3]}\cdots, \quad \text{for } \alpha \in \mathcal{A}.$$

Using the bijection Φ, any infinite word $w \in \mathcal{A}^\omega$ defines a unique smooth infinite word, using the pseudo-inverse functions as follows.

Example

Let $w = 11222\cdots$.
$\mathbf{1}21122122\cdots = \Delta_1^{-1}(112212\cdots)$
$\mathbf{1}12212\cdots = \Delta_1^{-1}(2211\cdots)$
$\mathbf{2}211\cdots = \Delta_2^{-1}(22\cdots)$
$\mathbf{2}2\cdots = \Delta_2^{-1}(2\cdots)$
$\mathbf{2}\cdots$
$\cdots$
Then $\Phi^{-1}(w) = 121122122\cdots$

Using this bijection, it takes $\mathcal{O}(\log n)$ to compute a prefix of length n and we can compress a prefix of length n in $\mathcal{O}(\log n)$.

Extremal smooth words (1/2)

Definition [Brlek, Melançon, P. - 2007]

The minimal (resp. maximal) infinite smooth word with respect to the lexicographic order is denoted $m_{\mathcal{A}}$ (resp. $M_{\mathcal{A}}$).

Examples

$$
\begin{aligned}
m_{\{1,2\}} &= 112112212112122112112122121122122112112 1221\cdots \\
M_{\{1,2\}} &= 2212211212212112212212112122112112212212112\cdots \\
m_{\{3,5\}} &= 3333555553333355533355533333555553333335553\cdots \\
m_{\{2,4\}} &= 2222444422224444224422442222444422224444224\cdots \\
m_{\{1,3\}} &= 1113111313111311131311131311131113131113111\cdots
\end{aligned}
$$

It takes time $\mathcal{O}(n^2)$ to compute a prefix of length n of an extremal smooth word.

Extremal smooth words (2/2)

Theorem [Brlek, Jamet, P. - 2008]

Over same parity alphabet, the only minimal infinite smooth words that are infinite Lyndon words are $m_{\{1<2b+1\}}$ and $m_{\{2a<2b\}}$, with $a, b \in \mathbb{N} \setminus \{0\}$.

Examples

$$m_{\{1,2\}} = 112112212112122 \cdot 1121121221211221221121121221 \cdots$$
$$m_{\{3,5\}} = 3333355555333335553335553333355555 \cdot 333335553 \cdots$$

are not Lyndon words while

$$m_{\{1,3\}} = 1113111313111311131311131311131113131113111 \cdots$$
$$m_{\{2,4\}} = 222244422224444224422442222444422224444224 \cdots$$

are so.

Result

Question

Is it possible to characterize the infinite smooth words, not necessarily minimal, that are also infinite Lyndon words ?

Answer

Over any 2-letter alphabet (same parity or not), the only infinite smooth words that are also infinite Lyndon words are $m_{\{2a<2b\}}$ and $m_{\{1<2b+1\}}$, for $a, b \in \mathbb{N} \setminus \{0\}$.

Idea of the proof

There are 4 cases to consider : $\mathcal{A} = \{a < b\}$, with

a) a even and b odd,

b) a and b even,

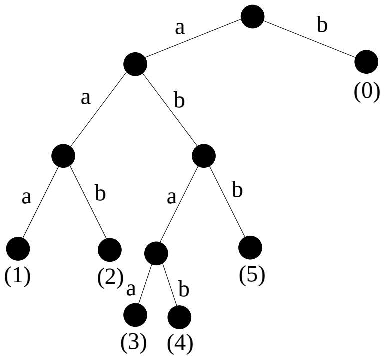
c) a odd and b even,

d) a and b odd.

For each case :

1) Consider all possible words p of length $\leq n$ s.t. $\Phi^{-1}(p)$ is prefix of an infinite smooth word w.
2) For each word p :
   - either we show that $\Phi^{-1}(p)$ can not be a prefix of a Lyndon word
   - or we describe an infinite smooth Lyndon word having $\Phi^{-1}(p)$ as prefix.

Case a) a even and b odd (1/2)



Case a) a even and b odd (2/2)

Case (0)

If $p = b$, then $\Phi^{-1}(p)$ can not be a prefix of an infinite smooth Lyndon word.

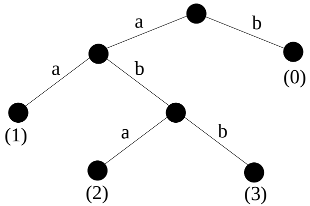
Case (4)

If $p = abab$, then

$$\Delta^0(w) = ((\mathbf{a}^b b^b)^{\frac{a}{2}})(a^a b^a)^{\frac{a}{2}})^{\frac{b-1}{2}} (a^b b^b)^{\frac{a}{2}} ((a^a b^a)^{\frac{b-1}{2}} a^a (b^b \underline{a}^b)^{\frac{b-1}{2}} b^b)^{\frac{b-1}{2}} \cdots$$
$$\Delta^1(w) = (\mathbf{b}^a a^a)^{\frac{b-1}{2}} b^a (a^b b^b)^{\frac{b-1}{2}} a^b \cdots$$
$$\Delta^2(w) = \mathbf{a}^b b^b \cdots$$
$$\Delta^3(w) = \mathbf{b} b \cdots$$

w has the prefix $(a^b b^b)^{\frac{a}{2}} a^a b^a$ and the smaller factor $f = (a^b b^b)^{\frac{b-1}{2}}$ contained in $(b^b a^b)^{\frac{b-1}{2}} b^b \implies w$ is not a Lyndon word.

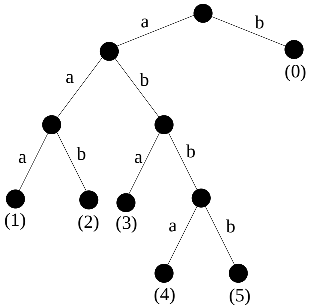
Case b) a and b even



Case (3) leads to an infinite smooth Lyndon word.

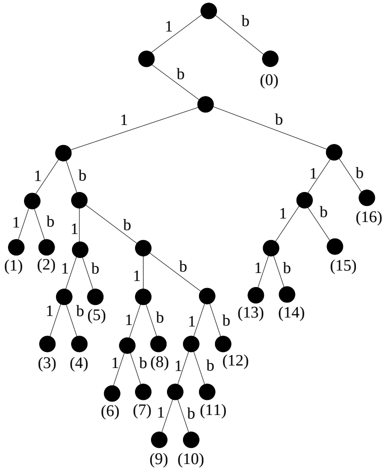
Case c) a odd and b even (1/3)

$a \neq 1$:



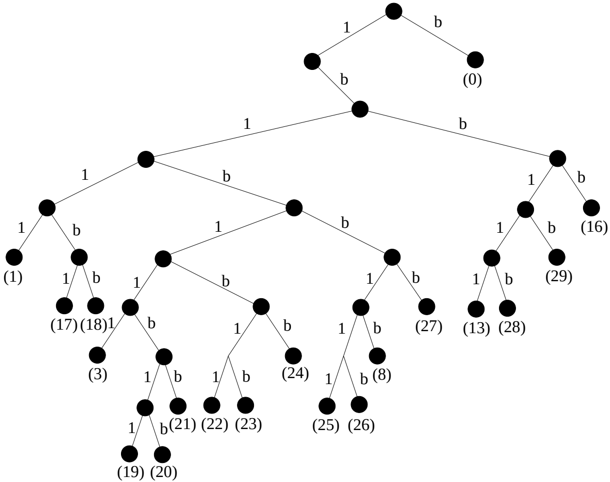
Case c) a odd and b even (2/3)

$a = 1$ and $b = 4n$:



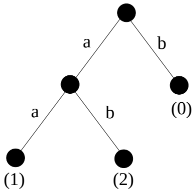
Case c) a odd and b even (3/3)

$a = 1$ and $b = 2(2n+1)$:



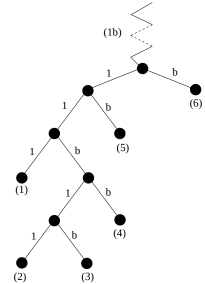
Case d) a and b odd (1/2)

$a \neq 1$:



Case d) a and b odd (2/2)

$a = 1$:



Only Case (5) leads to an infinite smooth Lyndon word.

Open problems

We proved that the only infinite smooth words having a trivial Lyndon factorization (only one factor) are $m_{\{2a<2b\}}$ and $m_{\{1<2b+1\}}$.

In 2000, Melançon characterized the Lyndon factorization of all standard Sturmian word s :

$$s = \prod_{n \geq 0} \ell_n^{c_{2n+1}}, \text{ with } \ell_n = (a\overline{s}_{2n+1})^{c_{2n}-1} a s_{2n} \overline{s}_{2n+1}.$$

- Characterize infinite smooth words that have a non trivial finite Lyndon factorization.
- Give an explicit computation of the Lyndon factorization, finite or infinite, of any infinite smooth words.