Nyldon Words

Émilie Charlier

joint work with Manon Philibert and Manon Stipulanti

Département de mathématiques, ULiège

17th journées montoises d'informatique théorique
Bordeaux, September 2018

Lyndon words

A word is Lyndon if it is primitive and lexicographically minimal among its conjugates.

Thus, w is Lyndon iff for all non-trivial factorizations $w = xy$, we have $w <_{lex} yx$.

- $0, 1$
- 01
- $001, 011$
- $0001, 0011, 0111$
- $00001, 00011, 00101, 00111, 01011, 01111$
- $\ldots$

The Chen-Fox-Lyndon theorem

Every word w can be uniquely factorized as

$$w = \ell_1 \ell_2 \cdots \ell_k$$

where $k \in \mathbb{N}$ and $\ell_1 \geq_{\text{lex}} \ell_2 \geq_{\text{lex}} \cdots \geq_{\text{lex}} \ell_k$ are Lyndon words.

Some Lyndon factorizations:

- $0100010110 = (01)(0001011)(0)$
- $000100111001 = (000100111001)$
- $0110101 = (011)(01)(01)$

Recursive definition of Lyndon words

- The letters are Lyndon.
- A word is Lyndon if and only if it cannot be factorized as a decreasing sequence of shorter Lyndon words.

NB: In this talk, "decreasing" means "nonincreasing".

Nyldon words

- The letters are Nyldon.
- A word is Nyldon if and only if it cannot be factorized as an increasing sequence of shorter Nyldon words.

NB: In this talk, "increasing" means "nondecreasing".

Mathoverflow

Nyldon words were defined in a post on Mathoverflow by Grinberg in November 2014, together with 2 conjectures:

- Is it true that any word can be uniquely factorized as an increasing sequence of shorter Nyldon words?
- Is it true that Nyldon words form a set of representatives of the primitive conjugacy classes?

Let's look at them

- $0, 1$

Let's look at them

- $0, 1$
- $0, 1$

Let's look at them

- $0, 1$
- $00, 01, 10, 11$

- $0, 1$

Let's look at them

- $0, 1$
- $00, 01, 10, 11$

- $0, 1$
- 10

Let's look at them

- $0, 1$
- $00, 01, 10, 11$
- $000, 001, 010, 011,$
  $100, 101, 110, 111$

- $0, 1$
- 10

Let's look at them

- $0, 1$
- $00, 01, 10, 11$
- $000, 001, 010, 011,$
  $100, 101, 110, 111$

- $0, 1$
- 10
- $100, 101$

Let's look at them

- 0, 1
- 00, 01, 10, 11
- 000, 001, 010, 011,
  100, 101, 110, 111
- 0000, 0001, 0010, 0011,
  0100, 0101, 0110, 0111,
  1000, 1001, 1010, 1011,
  1100, 1101, 1110, 1111

- 0, 1
- 10
- 100, 101

Let's look at them

- $0, 1$
- $00, 01, 10, 11$
- $000, 001, 010, 011,$
  $100, 101, 110, 111$
- $0000, 0001, 0010, 0011,$
  $0100, 0101, 0110, 0111,$
  $1000, 1001, 1010, 1011,$
  $1100, 1101, 1110, 1111$

- $0, 1$
- 10
- $100, 101$
- $1000, 1001, 1011$

Let's look at them

- $0, 1$
- 10
- $100, 101$
- $1000, 1001, 1011$
- $10000, 10001, 10010, 10011, 10110, 10111$
- $100000, 100001, 100010, 100011, 100110, 100111,$
  $101100, 101110, 101111$
- $1000000, 1000001, 1000010, 1000011, 1000100, 1000110,$
  $1000111, 1001100, 1001110, 1001111, 1011000, 1011001,$
  $1011100, 1011101, 1011110, 1011111$
- $\ldots$

If we count them, we find: $2, 1, 2, 3, 6, 9, 18, 30, 56, 99, 186, \ldots$

First observations and properties

- Nyldon words are not maximal among their conjugates: for example, 101 is Nyldon.
- Except 0 and 1, all binary Nyldon words start with 10.
- If one can prove the unicity of the Nyldon factorization, then we know that they are as many Nyldon words as Lyndon words of each length.

Nyldon words start with the prefix 10: why is that?

Nyldon words can't start with the letter 0 (except for 0 itself).

- Let $w = 0u$, with $u \neq \varepsilon$.
- Write

$$u = n_1 \cdots n_k, \qquad k \geq 1, \quad n_1 \leq_{\text{lex}} \cdots \leq_{\text{lex}} n_k$$

- But $0 \leq_{\text{lex}} n_1$ and 0 is Nyldon.
- Thus $w = (0)(n_1) \cdots (n_k)$ is a factorization of w into increasing shorter Nyldon words.
- By definition, w is not Nyldon.

Nyldon words start with the prefix 10: why is that?

Nyldon words can't start with 11.

- Let $w = 11u$.
- Write

$$1u = n_1 \cdots n_k, \qquad k \geq 1, \quad n_1 \leq_{\text{lex}} \cdots \leq_{\text{lex}} n_k$$

- Now observe that n_1 begins in 1.
- Then $1 \leq_{\text{lex}} n_1$ and 1 is Nyldon.
- This shows that $w = 11u = (1)(n_1)\cdots(n_k)$ factors into increasing shorter Nyldon words.
- By definition, w is not Nyldon.

Forbidden prefixes

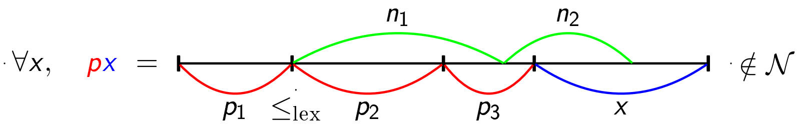
We can extend this argument in order to show that many other prefixes are not allowed in the family of Nyldon words:

- 0
- $11, 1010, 100100, \ldots$ (in general 10^k10^k)
- $1001011, 10001011, \ldots$ (in general 10^k1011)
- $1011011, 101110111, \ldots$ (in general 101^k01^k)
- $10011011, 1000110011, \ldots$ (in general $10^{k+1}110^k11$)
- $\ldots$

Forbidden prefixes

All the examples we have are from the following family:

$$F = \{p \in A^* : p = p_1 p_2 p_3,\ p_1 \text{ Nyldon},\ p_1 \leq_{\mathrm{lex}} p_2,$$
$$\forall x \in A^*,\ p_2 p_3 x = n_1 \cdots n_k \implies |n_1| \geq |p_2|\}$$



- All elements of F are forbidden prefixes.
- Question: Are there other forbidden prefixes?

Suffixes, rather than prefixes

- Prefixes of Lyndon and Nyldon words behave in a very different manner.
- On the opposite, we can show that suffixes of both families share some properties.

Theorem (Lyndon)

Let $w \in A^*$. Then the following assertions are equivalent:

- *w is Lyndon*
- *w is smaller than all its proper suffixes*
- *w is smaller than all its Lyndon proper suffixes.*

Theorem (Nyldon)

A Nyldon word is greater than all its Nyldon proper suffixes.

But the condition is not sufficient in the Nyldon case, even if we ask it to be true for all proper suffixes: 110 is not Nyldon.

Nyldon suffixes of Nyldon words

Nyldon	Lyndon
1000000, 1000001, 1000010,	0000001, 0000011, 0000101,
1000011, 1000100, 1000110,	0000111, 0001001, 0001101,
1000111, 1001100, 1001110,	0001111, 0010011, 0011101,
1001111, 1011000, 1011001,	0011111, 0001011, 0011011,
1011100, 1011101, 1011110,	0010111, 0110111, 0101111,
1011111	0111111

The longest Nyldon suffix of a word

Theorem (Lyndon)

- *The longest Lyndon suffix of a word is the right-most factor of its Lyndon factorization.*
- *The longest Lyndon prefix of a word is the left-most factor of its Lyndon factorization.*

Theorem (Nyldon)

The longest Nyldon suffix of a word is the right-most factor of any of its Nyldon factorizations.

NB: There is no similar condition on prefixes for Nyldon words.

Unicity of the Nyldon factorization

Every word w over A can be uniquely factorized as

$$w = n_1 n_2 \cdots n_k$$

where $k \in \mathbb{N}$ and $n_1 \leq_{\text{lex}} n_2 \leq_{\text{lex}} \cdots \leq_{\text{lex}} n_k$ are Nyldon words.

Lyndon vs Nyldon factorizations:

- $0100010110 = (01)(0001011)(0) = (0)(1000)(10110)$
- $000100111001 = (000100111001) = (0)(0)(0)(100111001)$
- $0110101 = (011)(01)(01) = (0)(1)(10)(101)$

Faster algorithm for computing Nyldon words

Theorem

If $w = PS$ where S is the longest Nyldon proper suffix of w and if $P = n_1 \cdots n_k$ is the Nyldon factorization of P, then w is Nyldon iff $n_k >_{\text{lex}} S$.

- This result allows us to compute the Nyldon factorization of w from right to left easily.

Compute the Nyldon factorization of w from right to left

$$1010010110 = (10)(100)(10110)$$

1	101001011.0
2	10100101.1.0, 10100101.10
3	1010010.1.10
4	101001.0.1.10
5	10100.1.0.1.10, 10100.10.1.10, 10100.101.10, 10100.10110
6	1010.0.10110
7	101.0.0.10110
8	10.1.0.0.10110, 10.10.0.10110, 10.100.10110
9	1.0.100.10110
10	.1.0.100.10110, .10.100.10110

Standard factorization

Theorem (Lyndon)

- *Let S be the longest Lyndon proper suffix of a word w and let $w = PS$. Then w is Lyndon iff P is Lyndon and $P <_{\text{lex}} S$.*
- *Let P be the longest Lyndon proper prefix of a word w and let $w = PS$. Then w is Lyndon iff S is Lyndon and $P <_{\text{lex}} S$.*

Theorem (Nyldon)

Let S be the longest Nyldon proper suffix of a word w and let $w = PS$. Then w is Nyldon iff P is Nyldon and $P >_{\text{lex}} S$.

NB: There is no similar condition on prefixes for Nyldon words.

From Lyndon to Nyldon: lost properties

Many results for Lyndon words don't have analogues in the case of Nyldon words.

- Results concerning prefixes.
- Let u, v be Lyndon words such that $u <_{\text{lex}} v$. Then uv is Lyndon.
- For example, $10010 >_{\text{lex}} 1$ but $100101 = (100)(101)$ is not Nyldon.
- The longest Lyndon suffix of a word is also the suffix that is lexicographically minimal.
- For example, the longest Nyldon suffix of 110 is 10 although it is not lex. maximal (nor lex. minimal).

Complete factorizations of the free monoid

A totally ordered family $F \subseteq A^*$ is called a complete factorization of A^* if each $w \in A^*$ can be uniquely factorized as

$$w = x_1 x_2 \cdots x_k \tag{1}$$

where $k \in \mathbb{N}$ and $(x_i)_{1 \le i \le k}$ is a decreasing sequence of F.

- Lyndon words are a complete factorization for the order $<_{\mathrm{lex}}$.
- Nyldon words are a complete factorization for the order $>_{\mathrm{lex}}$.

Schützenberger's theorem (1965)

Quote from the webpage of Dominique Perrin: "The following result has no known elementary proof."

Theorem (Schützenberger 1965)

Let A be an alphabet, $F \subseteq A^*$ and $\prec$ be a total order on F. Then any two of the following three conditions imply the third:

- *Each $w \in A^*$ admits at least one factorization* (1).
- *Each $w \in A^*$ admits at most one factorization* (1).
- *All elements of F are primitive and each primitive conjugacy class of A^+ contains exactly one element of F.*

Nyldon words and primitivity

As a consequence of the unicity of the Nyldon factorization and Schützenberger's theorem, we obtain:

Corollary

The Nyldon words form a set of representatives of the primitive conjugacy classes.

A simpler proof of the primitivity of Nyldon words?

Lyndon words are primitive by definition.

But suppose for a minute that you only know Lyndon words from their recursive definition.

(Of course, also suppose that you don't know Schützenberger's theorem.)

We can obtain that Lyndon words are smaller than all their Lyndon proper suffixes and the unicity of the Lyndon factorization.

Then, we easily deduce that Lyndon word are primitive and that each primitive conjugacy class contains exactly one Lyndon word.

Lyndon word are primitive and each primitive conjugacy class contains exactly one Lyndon word

- By induction on the length n of the words.
- Base case: all letters are Lyndon.
- Let w be a word of length $n \geq 2$ which is a power: $w = x^m$ with x primitive and $m \geq 2$.
- By induction hypothesis, we know that x has a Lyndon conjugate: $y = vu$ is Lyndon and $x = uv$.
- Then $w = x^m = (uv)^m = u(vu)^{m-1}v = uy^{m-1}v$.
- Let $u = u_1 \cdots u_k$ and $v = v_1 \cdots v_\ell$ be (the) Lyndon factorizations of u and v.
- Then $u_k \geq_{\text{lex}} y \geq_{\text{lex}} v_1$.
- Therefore $w = u_1 \cdots u_k \cdot \underbrace{y \cdots y}_{m-1} \cdot v_1 \cdots v_\ell$ is not Lyndon.

Lyndon word are primitive and each primitive conjugacy class contains exactly one Lyndon word.

- We have shown that Lyndon words of length n are primitive.
- Now suppose that there exist distinct Lyndon words x, y of length n in the same conjugacy class: $x = uv$ and $y = vu$.
- Then x^2 has two Lyndon factorizations:

$$\begin{aligned} x^2 &= x \cdot x \\ &= u_1 \cdots u_k \cdot y \cdot v_1 \cdots v_\ell \end{aligned}$$

which is a contradiction.

The same reasoning doesn't work for Nyldon words.

If $x = uv$ and $y = vu$ is Nyldon, then

$$x^m = u_1 \cdots u_k \cdot \underbrace{y \cdots y}_{m-1} \cdot v_1 \cdots v_\ell$$

is not the Nyldon factorization of x^m.

- For example, if $x = 01$ then $y = 10$ and $x^2 = (0)(101)$.
- For example, if $x = 001100101$ then $y = 100101001$ and $x^2 = (0)(0)(1)(10010)(1001100)(101)$.

Lazard factorizations of the free monoid

A Lazard factorization of A^* is a set $F \subseteq A^*$ satisfying the following conditions:

1. F is totally ordered by some order $\prec$
2. For every $n \geq 1$, if

   $$F \cap A^{\leq n} = \{u_1, \ldots, u_k\} \text{ with } u_1 \prec \cdots \prec u_k$$

   and if

   $$Y_1 = A \text{ and } Y_{i+1} = u_i^*(Y_i \setminus u_i) \text{ for all i,}$$

   then we have

   (i) for every $i \in \{1, \ldots, k\}$, $u_i \in Y_i$
   (ii) $Y_k \cap A^{\leq n} = \{u_k\}$.

Lyndon is Lazard

The set $\mathcal{L}$ is a Lazard factorization for the order $<_{\text{lex}}$.

For $A = \{0, 1\}$, the words of length ≤ 4 in the lexicographic order are $0, 0001, 001, 0011, 01, 011, 0111, 1$.

- $Y_1 = \{0, 1\}$
- $Y_2 = 0^*(Y_1 \setminus 0) = \{1, 01, 001, 0001\}$
- $Y_3 = (0001)^*(Y_2 \setminus 0001) = \{1, 01, 001\}$
- $Y_4 = (001)^*(Y_3 \setminus 001) = \{1, 01, 0011\}$
- $Y_5 = (0011)^*(Y_4 \setminus 0011) = \{1, 01\}$
- $Y_6 = (01)^*(Y_5 \setminus 01) = \{1, 011\}$
- $Y_7 = (011)^*(Y_6 \setminus 011) = \{1, 0111\}$
- $Y_8 = (0111)^*(Y_7 \setminus 0111) = \{1\}$

- u_1
- u_2
- u_3
- u_4
- u_5
- u_6
- u_7
- u_8

We observe that $u_i = \min Y_i$.

Nyldon is not Lazard

Theorem (Viennot 1978)

A complete factorization $(F, \prec)$ *of* A^* *is a Lazard factorization if and only if*

$$\forall x, y \in F,\ xy \in F \implies x \prec xy.$$

- Since the Lyndon words are a complete factorization w.r.t. $<_{\mathrm{lex}}$, this result confirms that they form a Lazard factorization.
- Since the Nyldon words are a complete factorization w.r.t. $>_{\mathrm{lex}}$, this result tells us that they do *not* form a Lazard factorization.

But Nyldon is right-Lazard

For $A = \{0, 1\}$, the Nyldon words of length ≤ 4 in the (increasing) lexicographic order are $0, 1, 10, 100, 1000, 1001, 101, 1011$.

- $Y_1 = \{0, 1\}$
- $Y_2 = (Y_1 \setminus 0)\, 0^* = \{1, 10, 100, 1000\}$
- $Y_3 = (Y_2 \setminus 1)\, 1^* = \{10, 100, 1000, 101, 1011, 1001\}$
- $Y_4 = (Y_3 \setminus 10)(10)^* = \{100, 1000, 101, 1011, 1001\}$
- $Y_5 = (Y_4 \setminus 100)(100)^* = \{1000, 101, 1011, 1001\}$
- $Y_6 = (Y_5 \setminus 1000)(1000)^* = \{101, 1011, 1001\}$
- $Y_7 = (Y_6 \setminus 1001)(1001)^* = \{101, 1011\}$
- $Y_8 = (Y_7 \setminus 101)(101)^* = \{1011\}$

- u_1
- u_2
- u_3
- u_4
- u_5
- u_6
- u_7
- u_8

We observe that $u_i = \min Y_i$.

Right Lazard factorizations of the free monoid

Lazard factorization $\rightarrow$ left Lazard factorization

A right Lazard factorization of A^* is a set $F \subseteq A^*$ satisfying the following conditions:

1. F is totally ordered by some order $\prec$
2. For every $n \geq 1$, if

   $$F \cap A^{\leq n} = \{u_1, \ldots, u_k\} \text{ with } u_1 \succ \cdots \succ u_k$$

   and if

   $$Y_1 = A \text{ and } Y_{i+1} = (Y_i \setminus u_i)\, u_i^* \text{ for all i,}$$

   then we have

   (i) for every $i \in \{1, \ldots, k\}$, $u_i \in Y_i$
   (ii) $Y_k \cap A^{\leq n} = \{u_k\}$.

Regular factorization = left and right Lazard factorization.

Right factorizations of the free monoid

Theorem (Viennot 1978)

A complete factorization $(F, \prec)$ *of* A^* *is a right Lazard factorization if and only if*

$$\forall x, y \in F,\ xy \in F \implies xy \prec y.$$

- Lyndon words are a right Lazard factorization for $<_{\mathrm{lex}}$.
- Nyldon words are a right Lazard factorization for $>_{\mathrm{lex}}$.

Four types of Lazard factorizations

The min and max choices show an asymetric situation for left and right Lazard factorizations:

Left Lazard		Right Lazard	
lexmin	$\mathcal{L}$	lexmin	$\mathcal{N}$
lexmax	$\bar{\mathcal{L}}$	lexmax	$\mathcal{L}$

References

- J. Berstel, D. Perrin and C. Reutenauer, Codes and Automata, *Encyclopedia of Mathematics and its Applications* **129**, Cambridge University Press, 2010.
- E. Charlier, M. Philibert and M. Stipulanti, *Nyldon words*, Arxiv 2018.
- D. Grinberg, "Nyldon words": understanding a class of words factorizing the free monoid increasingly, *Mathoverflow* 2014.
- M. Lothaire, Combinatorics on words, *Encyclopedia of Mathematics and its Applications* **17**, Addison-Wesley Publishing Co.,1983.
- M.-P. Schützenberger, On a Factorisation of Free Monoids, *Proc. Amer. Math. Soc.*, **16**, 21–24, 1965.
- G. Viennot, Algèbres de Lie Libres et Monoïdes libres, *Lecture Notes in Mathematics*, **691**, Springer-Verlag, 1978.