Combinatorics on Words

Gabriele Fici

CWI, Amsterdam — April 2024

Part 6: Classes of Finite Words

Lyndon Words

An important class of finite words is that of Lyndon words.

In what follows, we let $<$ denote the lexicographic order.

Proposition 1

For every word $w, u, v \in \Sigma^*$, we have $u < v$ if and only if $wu < wv$.

Moreover, if u is not a prefix of v, then for every word $z \in \Sigma^*$ we have $u < v$ if and only if $uw < vz$.

One could be tempted to state that $u < v$ if and only if $uw < vw$ for any word w. However, this is not true in general. For example, $01 < 010$ but $01 \cdot 1 \not< 010 \cdot 1$.

Lyndon Words

Definition 2

Let Σ be a totally ordered alphabet. A nonempty word $w \in \Sigma^*$ is a Lyndon word if it is lexicographically smaller than each of its proper suffixes, i.e., for every factorization $w = uv$ in two nonempty words, one has $w < v$.

In particular, letters are Lyndon words (trivially) and every Lyndon word is primitive.

Lyndon Words

Theorem 3

Let w be a primitive word. The following are equivalent:

1. *For every factorization $w = uv$ in two nonempty words, $uv < v$;*
2. *For every factorization $w = uv$ in two nonempty words, $uv < vu$;*
3. *For every factorization $w = uv$ in two nonempty words, $u < v$;*
4. *For every factorization $w = uv$ in two nonempty words, $u < vu$.*

Notice that the conditions:

- *For every factorization $w = uv$ in two nonempty words, $u < uv$;*

and

- *For every factorization $w = uv$ in two nonempty words, $v < vu$;*

are true for any word, not just for Lyndon words.

Remark 4

The statement of the previous theorem cannot be written as:

Let w be a primitive word, and let $w = uv$ be a factorization of w in two nonempty words. The following are equivalent:

1. $uv < v$;
2. $uv < vu$;
3. ...

Indeed, this is not true in general. Take for example $w = 01001$, and the factorization $w = uv$ with $u = 010$ and $v = 01$; then $uv = 01001$ is smaller than $vu = 01010$, yet it is not smaller than $v = 01$.

Lyndon Words

Theorem 5

Let w be a primitive word. The following are equivalent:

1. *For every factorization $w = uv$ in two nonempty words, $uv < v$;*
2. *For every factorization $w = uv$ in two nonempty words, $(uv)^\omega < (vu)^\omega$;*
3. *For every factorization $w = uv$ in two nonempty words, $(uv)^\omega < v^\omega$;*
4. *For every factorization $w = uv$ in two nonempty words, $(vu)^\omega < v^\omega$;*
5. *For every factorization $w = uv$ in two nonempty words, $u^\omega < v^\omega$;*
6. *For every factorization $w = uv$ in two nonempty words, $u^\omega < (vu)^\omega$;*
7. *For every factorization $w = uv$ in two nonempty words, $v^\omega < (vu)^\omega$;*
8. *For every factorization $w = uv$ in two nonempty words, $u^\omega < (uv)^\omega$.*

Remark 6

It is not true, in general, that $u < v$ implies $u^\omega < v^\omega$. For example, $01 < 010$, but $(010)^\omega < (01)^\omega$.

Lyndon Words

Every Lyndon word is not only primitive but also unbordered. Conversely, every nonempty primitive word has a conjugate that is a Lyndon word. As a byproduct, every primitive word has a conjugate that is unbordered (as we already saw).

But one can prove that reversing the order on Σ, the Lyndon conjugate with respect to the reverse order cannot coincide with the Lyndon conjugate of the previously fixed order, provided that the word is not of length 1. So, every primitive word of length > 1 has at least two unbordered conjugates.

For example, over $\Sigma = \{a, b\}$ with $a < b$, take the primitive word $abaabb$. Its conjugate $aabbab$ is Lyndon. For the order $b < a$, the Lyndon conjugate is $bbabaa$.

Lyndon Words

Example 7

The first few Lyndon words over $\Sigma = \{a, b\}$, with $a < b$, are:

a, b,
ab,
aab, abb,
$aaab$, $aabb$, $abbb$,
$aaaab$, $aaabb$, $aabab$, $aabbb$, $ababb$, $abbbb$.

Lyndon Words

Proposition 8

For every words u, v, with v nonempty, and every $n \geq 1$, $u^n v$ is a Lyndon word if and only if uv is a Lyndon word.

Proposition 9

If u and v are Lyndon words, and $u < v$, then uv is a Lyndon word. Conversely, for every Lyndon word w of length at least 2 there exist Lyndon words u, v, with $u < v$, such that $w = uv$.

For example, $aabb$ can be written as $a \cdot abb$. Note that $aabb$ can also be written as $aab \cdot b$, so the factorization in the previous proposition is not, in general, unique.

Generating Lyndon Words

The set of Lyndon words of length $\leq n$ can be constructed recursively by the following algorithm, which makes use of Proposition 8: Start with $X_1 = \Sigma$; assuming X_i is already constructed, take the least word x_i in X_i and define

$$X_{i+1} = X_i \backslash \{x_i\} \cup \{w : |w| \leq n \text{ and } w = x_i^k x_j \text{ for some } k \geq 1 \text{ and } j \neq i\}.$$

For example, over $\Sigma = \{a, b\}$, with $a < b$, for $n = 4$ one has:

$$X_1 = \{a, b\}$$
$$X_2 = \{aaab, aab, ab, b\}$$
$$X_3 = \{aab, ab, b\}$$
$$X_4 = \{aabb, ab, b\}$$
$$X_5 = \{ab, b\}$$
$$X_6 = \{abb, b\}$$
$$X_7 = \{abbb, b\}$$
$$X_8 = \{b\}$$

Then we stop since no new word is created after this iteration. The set of Lyndon words of length $\leq n$ is given by the union of the sets X_i.

Counting Lyndon Words

The number of Lyndon words of length n is equal to the number of conjugacy classes of primitive words of length n, and we saw that this number is given by Witt's formula

$$\frac{1}{n}\sum_{d|n}\mu(d)\ |\Sigma|^{n/d}$$

where μ is the Möbius function.

If $|\Sigma| = 2$, then number of Lyndon words of length n is asymptotically

$$\frac{2^n}{n}\left(1+O(2^{-n/2})\right).$$

The simple fact that there are exponentially many Lyndon binary words of length n follows immediately from the observation that for any word w of length n, the word $a^n wb$ is a Lyndon word of length $2n+1$.
Therefore, there are at least 2^n binary Lyndon words of length $2n+1$.

Lyndon Factors of a Lyndon Word

The minimum number of distinct Lyndon factors in a word of length n is 1 (consider for instance the word a^n). But what is the minimum number of distinct Lyndon factors in a Lyndon word?

Saari in 2014 proved that if w is a Lyndon word with $|w| \geq F_n$ for some $n \geq 3$, where F_n is the nth Fibonacci number, then w contains at least n distinct Lyndon factors. Therefore, every Lyndon word of length n contains at least $1 + \lceil \log_\varphi n \rceil$ distinct Lyndon factors.

This bound is tight, as it is realized by Lyndon factors of the Fibonacci word, which have length F_n and contain n distinct Lyndon factors.

For example, 00100101 is a Lyndon factor of the Fibonacci word; it has length $8 = F_6$ and contains 6 distinct Lyndon factors, namely $0, 1, 01, 001, 00101, 00100101$.

Conversely, all words of the form $a^n b^n$ (which are Lyndon) have the maximum number of Lyndon factors among all binary words of length $2n$.

Lyndon Sesquipowers

Powers of Lyndon words (e.g., $abab$ or bbb) are precisely those words w such that w is smaller than *or equal to* every its proper suffix (or conjugate). They are sometimes called necklaces (meaning that they can be chosen as representatives of a conjugacy class when the order is fixed), while Lyndon words are also called aperiodic necklaces.

Let us now consider periodic extensions (i.e., fractional powers) of Lyndon words, e.g., $aab \cdot aab \cdot a$. They are prefixes of powers of Lyndon words and are called Lyndon sesquipowers (or also preprime words, in those contexts in which Lyndon words are called prime words).

Let b be the largest letter of the alphabet. Then b^n, for $n > 1$, cannot be the prefix of a Lyndon word. But any other periodic extension of a Lyndon word is indeed a prefix of a Lyndon word. Let us call Lyndon prefixes the words that are prefixes of Lyndon words. So,

$$\text{Lyndon prefixes} \quad = \quad \text{Lyndon sesquipowers} \setminus \{b^n \mid n > 1\}.$$

Lyndon Sesquipowers

As a consequence, in order to construct the set of Lyndon sesquipowers of length n, or the set of Lyndon prefixes of length n (the latter is obtained from the former by removing the word b^n), it is sufficient to take the set of Lyndon words of length $\leq n$ and extend each of them them periodically up to a word of length n.

For example, taking the set of Lyndon words of length ≤ 4, we have that the set of Lyndon sesquipowers of length 4 over $\Sigma = \{a, b\}$, $a < b$, is

$$Y = \{aaaa, aaab, aaba, abab, aabb, abba, abbb, bbbb\}$$

and the set of Lyndon prefixes of length 4 is $Y \setminus \{bbbb\}$.

Remark 10

Notice that in the previous set, the Lyndon words are precisely the unbordered elements. Indeed, a prefix of a Lyndon word is a Lyndon word if and only if it is unbordered.

Lyndon Sesquipowers

Another corollary is the following.

Proposition 11

Let wa, $w \in \Sigma^*$, $a \in \Sigma$, be a prefix of a Lyndon word and b a letter greater than a. Then wb is a Lyndon word.

Standard Factorization

The factorization of Proposition 9 is not, in general, unique. So we define a standard factorization as follows.

Definition 12

The (right) standard factorization of a Lyndon word w of length > 1 is $w = uv$, where v is the lexicographically least proper suffix of w (or, equivalently, the longest proper suffix of w that is a Lyndon word).

For example, the standard factorization of $aabb$ is $a \cdot abb$.

Standard Factorization

Let us show that the longest proper suffix of w that is a Lyndon word, v_{Lyn}, coincides with the lexicographically least proper suffix of w, $v_{\min}$.

Since both are suffixes of w, one is a suffix of another. But $v_{\min}$ cannot be a proper suffix of v_{Lyn}, otherwise v_{Lyn} would have a proper suffix lexicographically smaller than itself, against the definition of Lyndon word. Hence v_{Lyn} is a suffix of $v_{\min}$. But $v_{\min}$ is a Lyndon word — since it is smaller than all its suffixes — and v_{Lyn}, the longest Lyndon suffix of w, cannot be shorter than $v_{\min}$, whence v_{Lyn} and $v_{\min}$ coincide.

Exercise 13

Prove that the words u and v in the standard factorization are both Lyndon words.

Standard Factorization

There is also a left standard factorization of a Lyndon word w (a.k.a. Viennot factorization).

It is the factorization $w = uv$, where u is the longest proper prefix of w that is a Lyndon word (but not the lexicographically least proper prefix of w, which is always a single letter).

The left and right standard factorizations do not coincide, in general. For example, the left standard factorization of $aabb$ is $aab \cdot b$.

Lyndon Tree

The standard factorization induces a binary tree structure on a Lyndon word w, in which the root is w and the children of a factor w' of length greater than 1 are the words in the standard factorization of w'. The leaves of the tree are single letters.

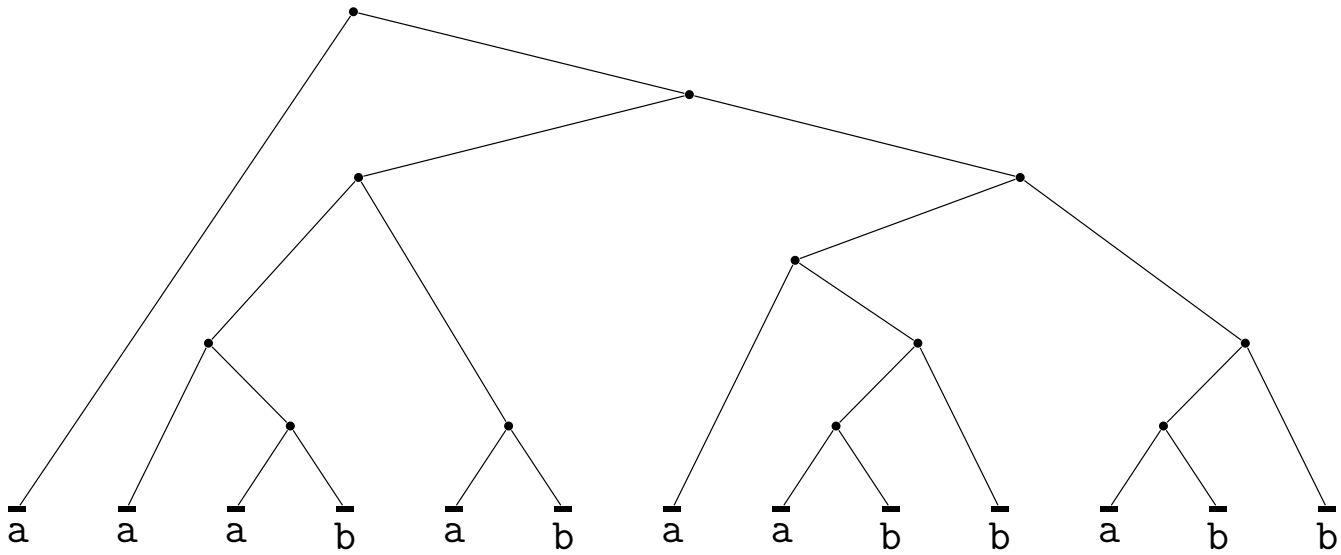


Figure: The Lyndon tree of the Lyndon word $w = aaababaabbabb$.

Lyndon Tree

Note that any word w can be made Lyndon by prepending to it a sentinel symbol # that is smaller than every other letter of the alphabet.

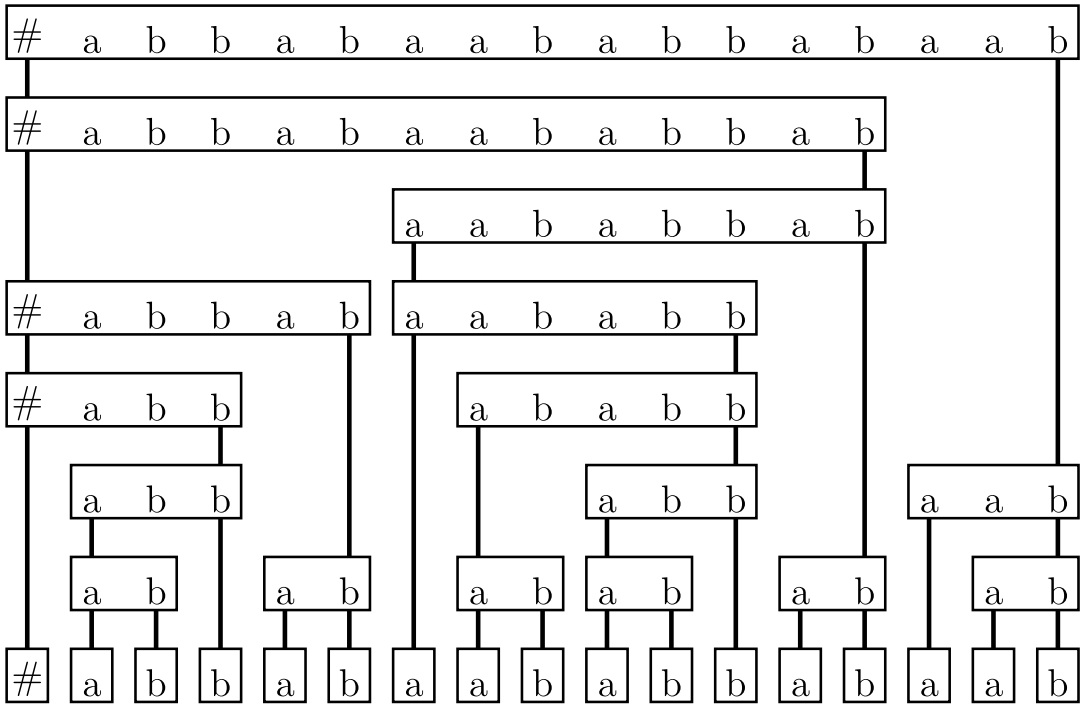
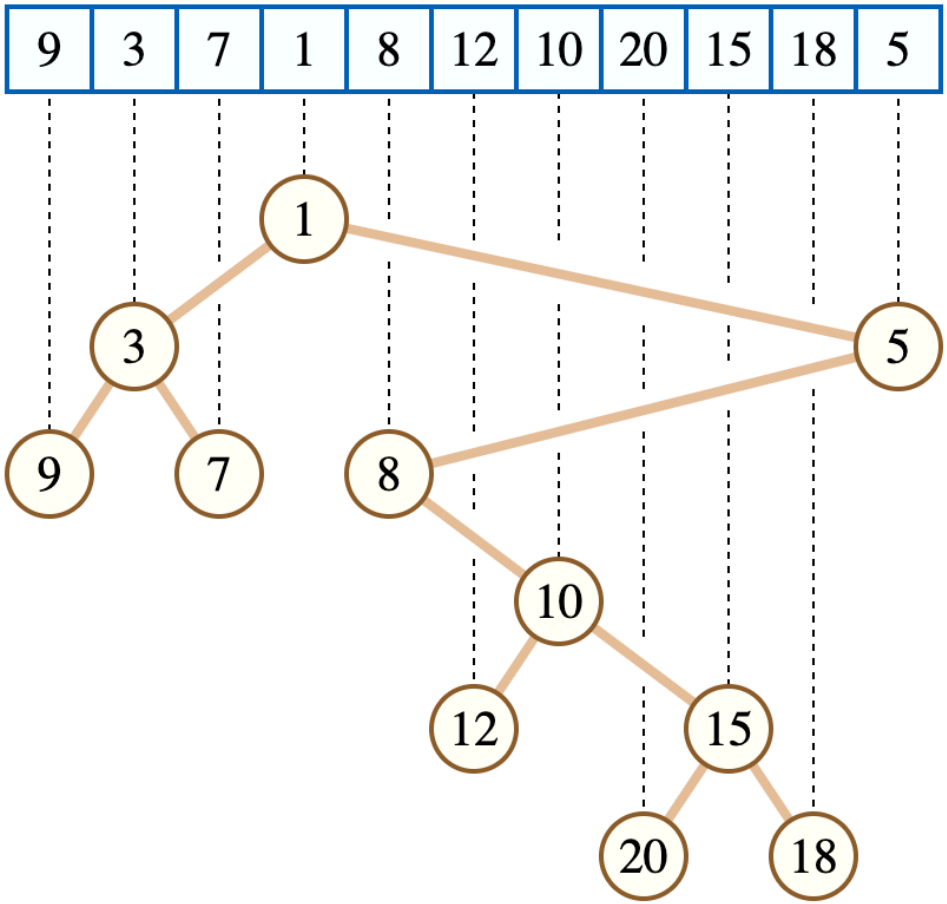


Figure: The Lyndon tree of the Lyndon word $w = \#abbabaababbabaab$.

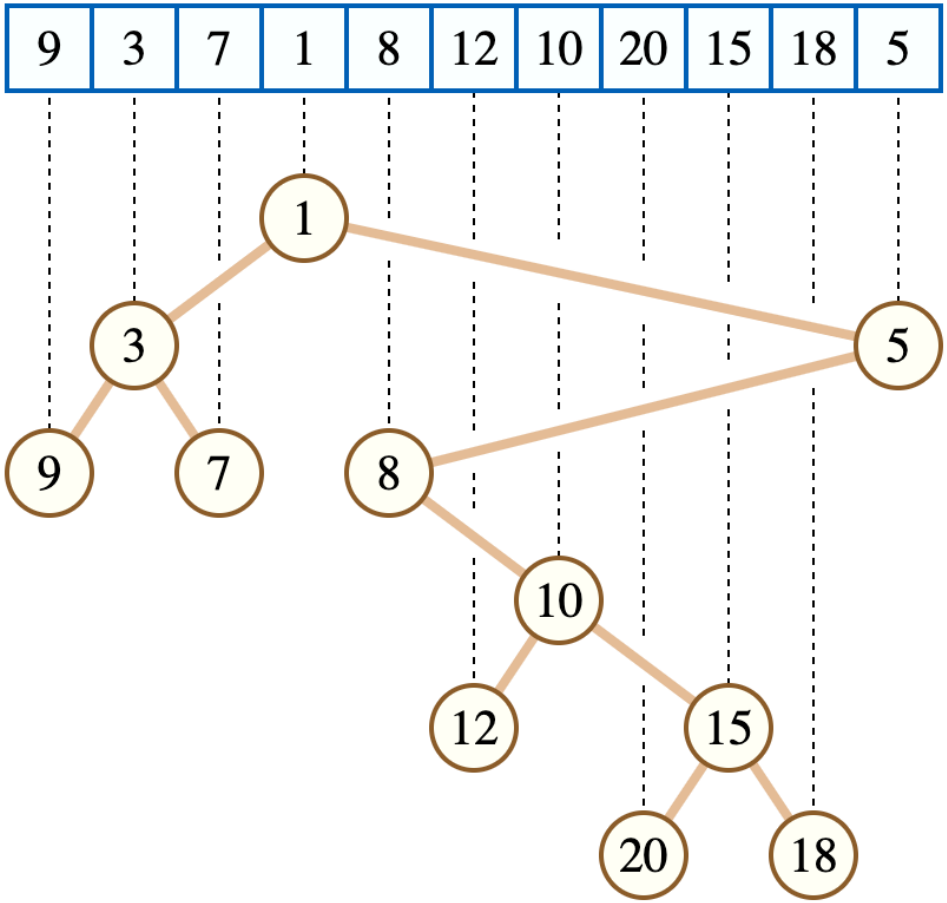
Lyndon Tree

Given a sequence S, one can define the Cartesian tree of S as the ordered binary tree whose nodes are the elements of S, the root is the node labeled by the least element, and an inorder traversal of the tree produces S.



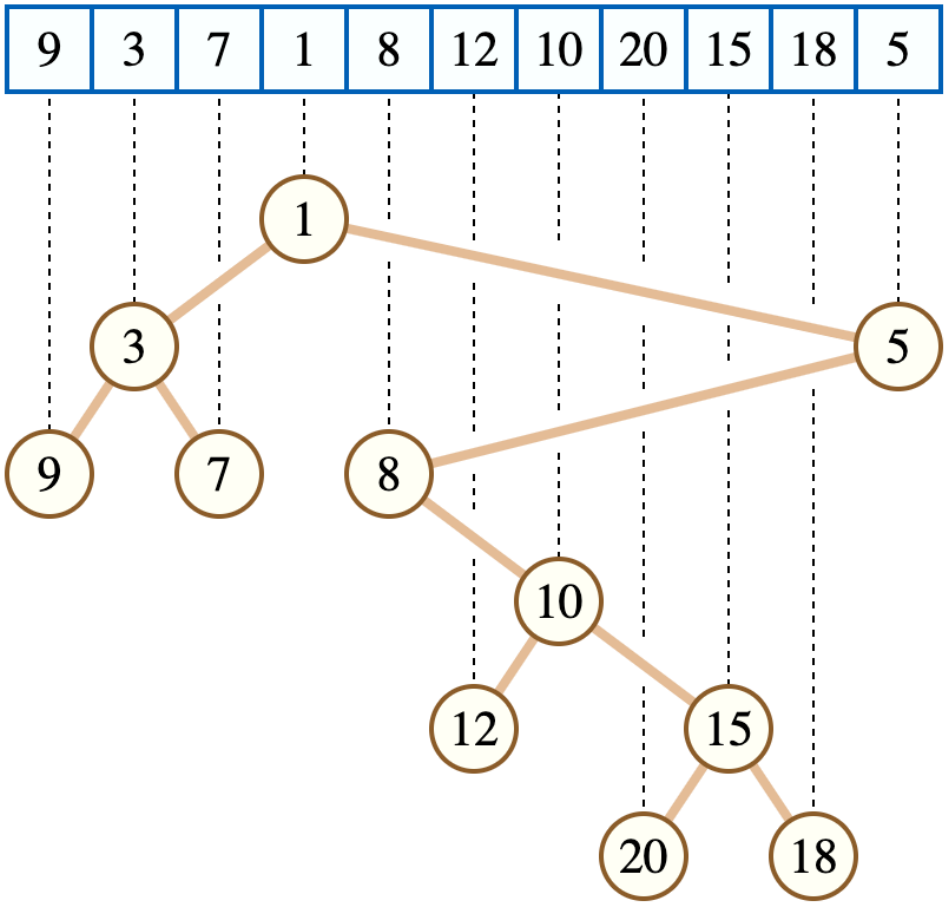
Lyndon Tree

Therefore, the left subtree of a node i contains all the elements of S that appear in S to the left of i, and the right subtree of i contains all the elements of S that appear in S to the right of i.



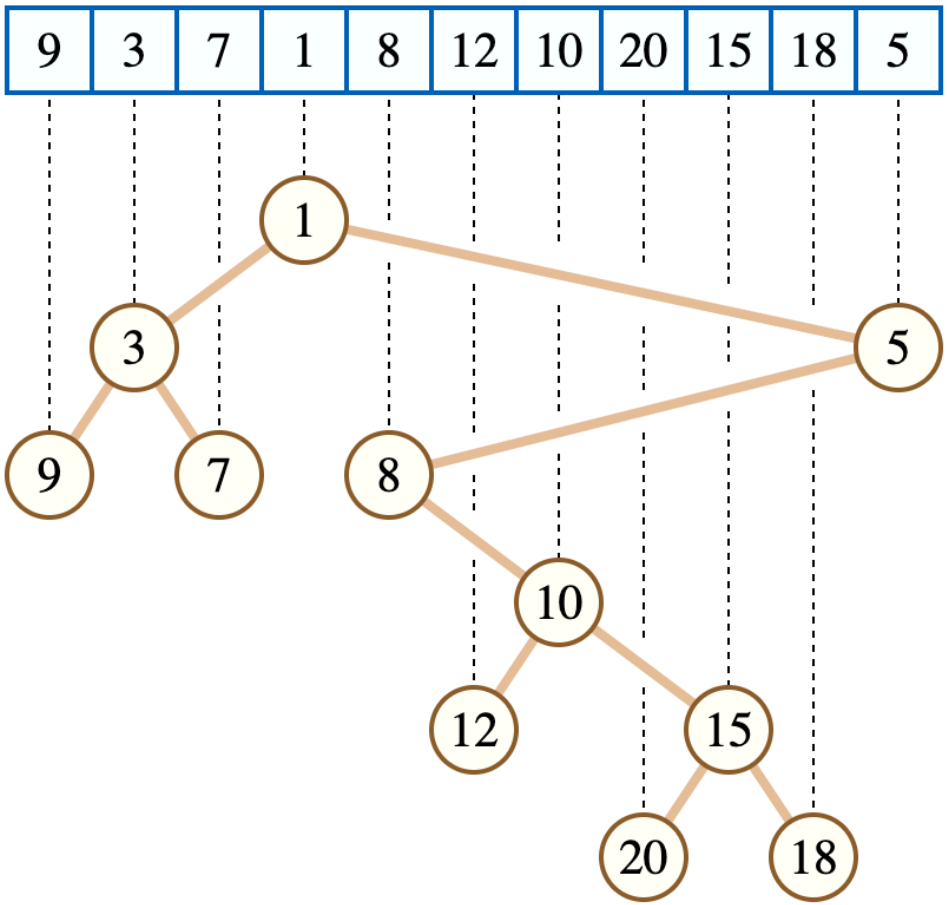
Lyndon Tree

As a consequence, the labels of the nodes in the subtree of i are all greater than i.



Lyndon Tree

Another property of the Cartesian tree is the following. Given i and j, the minimum value in the subsequence of S between i and j is the value of the node that is the lowest common ancestor between nodes i and j.



Lyndon Tree

Any sequence S can be associated with its corresponding Cartesian tree $CT(S)$ according to the following rules:

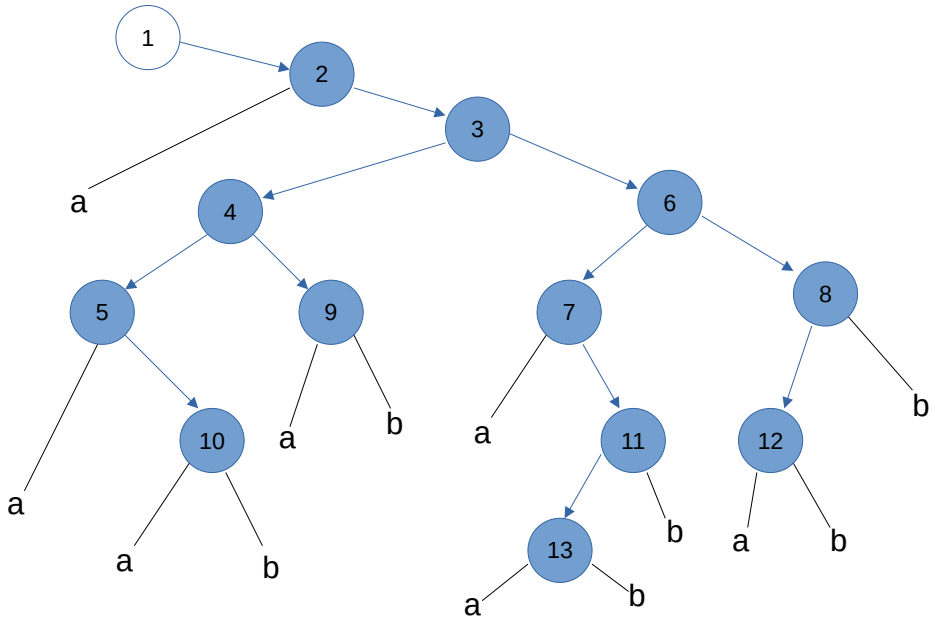
- If S is empty, then $CT(S)$ is an empty tree.
- If S is not empty and $S[i]$ is the minimum value in S, then $CT(S)$ is the tree with $S[i]$ as the root, $CT(S[1,\ldots,i-1])$ as the left subtree, and $CT(S[i+1,\ldots,n])$ as the right subtree. If there are two or more minimum values, we choose the leftmost one as the root.

Lyndon Tree

Theorem 14 (Hohlweg, Reutenauer, 2003)

Let w be a Lyndon word. Let 1π be the permutation associated with the ranking of the suffixes of w (a.k.a. the Inverse Suffix Array of w). Then, the Lyndon tree of w, after removing the leaves, coincides with the Cartesian tree of π.

For example, let $w = aaababaabbabb$. Its suffix ranking permutation is $ISA(w) = 1\pi = [1, 2, 5, 10, 4, 9, 3, 7, 13, 11, 6, 12, 8]$.



Lyndon Tree

Therefore, starting from the left and separating recursively the array $ISA(w)$ taking the smallest value in the range, we get the Lyndon tree of w, corresponding to the parentheses representation

$$(a)(aababaabbabb)$$

$$(a)((aabab)(aabbabb))$$

$$(a)(((aab)(ab))((aabb)(abb)))$$

$$(a)((((a)(ab))((a)(b)))(((a)(abb))((ab)(b))))$$

$$(a)((((a)((a)(b)))((a)(b)))(((a)((ab)(b)))(((a)(b))(b))))$$

$$(a)((((a)((a)(b)))((a)(b)))(((a)(((a)(b))(b)))(((a)b)(b))))$$

Remark 15

For binary words, it is possible to retrieve w from its ISA by writing for each position i the letter a if the value at i is smaller than the value at $i+1$ or b otherwise (and putting a b in the last position).

Lyndon Array

Given a word w, the Lyndon array $\lambda(w)$ is the array such that $\lambda[i]$ is the length of the longest Lyndon factor of w that starts at position i.

For example, if $w = abaababaab$, then $\lambda(w) = [2, 1, 5, 2, 1, 2, 1, 3, 2, 1]$.

Let us define the Next Smaller Value array of a permutation π of $[1 \dots n]$ as the array NSV whose i-th entry is the distance between i and the smallest index greater than i such that $\pi[j] < \pi[i]$, if such a j exists, or $NSV[i] = n + 1 - i$ (the distance between i and $n + 1$) otherwise.

Theorem 16

The Lyndon array of w coincides with the Next Smaller Value array of the permutation associated with the ranking of the suffixes of w (the ISA of w).

For example, let $\pi = ISA(abaababaab) = [5, 9, 2, 6, 10, 4, 8, 1, 3, 7]$.
Then

$$NSV[\pi] = [2, 1, 5, 2, 1, 2, 1, 3, 2, 1].$$

Lyndon Factorization

Theorem 17

Any word factorizes uniquely in non-increasing Lyndon words. This factorization is called the Lyndon factorization of w.

For example, let $w = abaaaabaaaaabaaaabaaaaaab$. The Lyndon factorization of w is

$$ab \cdot aaaab \cdot aaaaabaaaab \cdot aaaaaab.$$

The Lyndon factorization of a word w can be computed by taking the longest prefix that is a Lyndon word and recurse on the word obtained by removing this prefix.

Equivalently, it can be computed by taking the lexicographically smallest nonempty suffix and recurse on the word obtained by removing this suffix.

Lyndon Factorization

As a consequence of the theorem of Hohlweg and Reutenauer, the Lyndon factorization of w can be computed from its ISA, starting from the first position and by searching iteratively for the next position in which the value is smaller.

For example, for $w = abaababaab$, the ISA is $[5, 9, 2, 6, 10, 4, 8, 1, 3, 7]$. The Lyndon factorization of w is

$$ab \cdot aabab \cdot aab.$$

The ISA of $w = abaaaabaaaaabaaaabaaaaaab$ is

$$[20, 25, 6, 10, 14, 18, 23, 3, 7, 11, 15, 19, 24, 5, 9, 13, 17, 22, 1, 2, 4, 8, 12, 16, 21]$$

and the Lyndon factorization is

$$ab \cdot aaaab \cdot aaaaabaaaab \cdot aaaaaab.$$

Lyndon Factorization

The Lyndon factorization can be defined for infinite words as well. It is defined by taking the longest (possibly infinite) prefix that is Lyndon, and recurse on the suffix that remains.

For example, let $w_1 = 011$, $w_2 = 01$ and for every $n > 1$, w_{n+1} the word obtained by rotating by one position the word $\tau(w_n)$ (that is, removing the last letter and putting it in front of the word), where τ is the Thue–Morse morphism $0 \mapsto 01, 1 \mapsto 10$.

$$w_1 = 011$$
$$w_2 = 01$$
$$w_3 = 0011$$
$$w_4 = 00101101$$
$$w_5 = 0010110011010011$$

Then, the Lyndon factorization of the Thue–Morse word t is

$$t = \prod_{n \geq 1} w_n = 011 \cdot 01 \cdot 0011 \cdot 00101101 \cdot 0010110011010011 \cdots$$

de Bruijn Words

A de Bruijn word of order n on an alphabet Σ of size k is a circular word of length k^n such that every word of length n on k letters appears exactly once as a factor.

For example, $aaababbb$ is a de Bruijn word of order 3.

If one wants a linear word with the same property, it is sufficient to concatenate a de Bruijn word with its prefix of length $n-1$.

de Bruijn Words

One way to generate a de Bruijn word is given by the following remarkable theorem.

Theorem 18 (Fredricksen, Maiorana, 1978)

The lexicographically least de Bruijn word of order n is obtained by concatenating in increasing lexicographic order the Lyndon words of length dividing n.

For example, if $n = 4$ and $\Sigma = \{a, b\}$, then

$$aaaabaabbababbbb = a \cdot aaab \cdot aabb \cdot ab \cdot abbb \cdot b$$

is the least binary de Bruijn word of order 4.

de Bruijn Words

There is an interesting extension of the theorem of Fredericksen and Maiorana. A generalized de Bruijn word of order n on k letters is a circular word of length k^n such that every primitive word of length n on k letters appears exactly once as a factor.

Theorem 19 (Au, 2015)

The lexicographically least generalized de Bruijn word of order n is obtained by concatenating in increasing lexicographic order the Lyndon words of length n.

So, for example, if $n = 4$ and $\Sigma = \{a, b\}$, then

$$aaabaabbabbb = aaab \cdot aabb \cdot abbb$$

is the least generalized de Bruijn word of order 4 over Σ. The primitive words of length 4 over Σ are: $aaab$, $aaba$, $aabb$, $abaa$, $abba$, $abbb$, $baaa$, $baab$, $babb$, $bbaa$, $bbab$ and $bbba$.

de Bruijn Words

The de Bruijn graph of order $n > 1$ over an alphabet Σ of cardinality k is the directed graph whose nodes are the words over Σ of length n and there is an edge from u to v if removing the first letter of u produces a prefix of v (the label of the edge is the last letter of v).

The de Bruijn graph of order n has k^n nodes and k^{n+1} edges, it is strongly connected, and every node has indegree and outdegree k. Therefore, it is an Eulerian graph. It is also Hamiltonian.

Theorem 20

The set of de Bruijn words of order $n+1$ is equal to the set of labels of Eulerian cycles in the de Bruijn graph of order n. It is also equal to the set of labels of Hamiltonian cycles in the de Bruijn graph of order $n+1$.

Using the previous theorem, it is possible to prove that there are $\frac{(k!)^{k^{n-1}}}{k^n}$ distinct de Bruijn words of order n on k letters.

de Bruijn Words

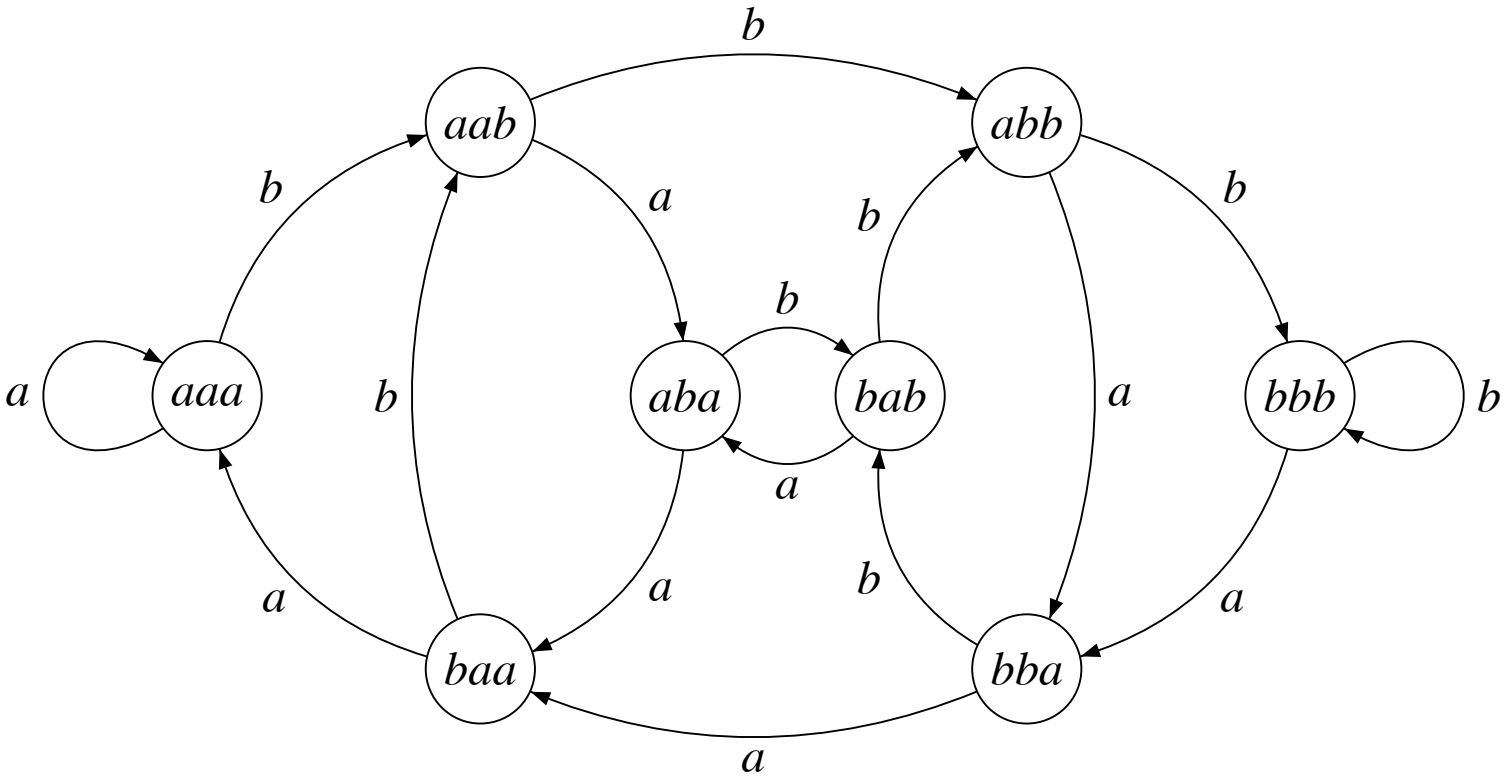


Figure: The de Bruijn graph of order 3 over $\Sigma = \{a, b\}$. One can verify that the de Brujn word of order 3 $aaababbb$ is indeed the label of a Hamiltonian cycle (starting from node bbb), and that the de Brujn word of order 4 $aaaabaabbbabaabbbb$ is indeed the label of an Eulerian cycle (always starting from node bbb).

Finite Sturmian Words

Definition 21

A finite or infinite word w over Σ_2 is C-balanced for an integer $C \geq 1$ if and only if for every factors of w of the same length u and v, one has $||u|_0 - |v|_0| \leq C$, that is, the number of 0s (or, equivalently, 1s) in two factors of the same length differ at most by C.

For example, the Thue–Morse word is 2-balanced but not 1-balanced; the Fibonacci word is 1-balanced.

We now take a closer look at finite 1-balanced words over Σ_2, i.e., finite factors of Sturmian words.

Finite Sturmian Words

Binary 1-balanced finite words are called finite Sturmian words.

We let *St* denote the set of finite Sturmian words.

de Luca and De Luca gave some characterizations of finite Sturmian words. For a nonempty word w, let ρ_w denote its fractional root, $\pi_w = |\rho_w|$ its minimum positive period, and R_w the least integer k such that w has no right special factor of length k.

Theorem 22

Let w be a nonempty word. The following conditions are equivalent:

- *w is a finite Sturmian word;*
- *ρ_w is a conjugate of a standard Sturmian word;*
- $\pi_w = 1 + R_{\rho_w^2}$.

Finite Sturmian Words

The following are characterizations of the words w such that w^2 is Sturmian (such words are sometimes called circularly balanced).

Proposition 23

The following conditions are equivalent:

1. w^2 *is Sturmian;*
2. w^n *is Sturmian for every* $n \geq 0$*;*
3. *every conjugate of* w^2 *is Sturmian;*
4. *every conjugate of* w *is Sturmian;*
5. w *is Sturmian and it is either non-primitive or a conjugate of a Lyndon Sturmian word.*

Finite Sturmian Words

On the other hand, we have a characterization of binary words that are not Sturmian.

Proposition 24

Let $w \in \Sigma_2^*$. Then w is not Sturmian if and only if there exists a palindrome v such that $0v0$ and $1v1$ are both factors of w.

The pair $(0v0, 1v1)$ of the previous proposition is called an unbalanced pair.

Proposition 25 (Dulucq, Gouyou-Beauchamps, 1987)

The language of binary words that are not Sturmian (i.e., that contain an unbalanced pair) is context-free.

Sturmian Morphisms

A Sturmian morphism is a morphism such that all images of finite Sturmian words are Sturmian.

Clearly, the identity morphism id and the morphism E that maps 0 to 1 and 1 to 0 are Sturmian morphisms. Moreover, a composition of Sturmian morphisms is a Sturmian morphism, so Sturmian morphisms constitute a monoid, called the Sturm monoid (or Sturmian monoid).

This monoid is generated by E, φ and $\widetilde{\varphi}$, where $\varphi : 0 \mapsto 01,\ 1 \mapsto 0$ and $\widetilde{\varphi} : 0 \mapsto 10,\ 1 \mapsto 0$ are, respectively, the Fibonacci morphism and the reverse Fibonacci morphism.

Sturmian Morphisms

Recall that the incidence matrix of the endomorphism of Σ_2 $\mu : 0 \mapsto u,\ 1 \mapsto v$ is $M_\mu = \begin{pmatrix} |u|_0 & |v|_0 \\ |u|_1 & |v|_1 \end{pmatrix}$.

Theorem 26

A matrix $M \in \mathbb{N}^{2\times 2}$ is the incidence matrix of a Sturmian morphism if and only if $\det(M) = \pm 1$, i.e., if and only if it is invertible.

So, the subset of matrices of $GL_2(\mathbb{Z})$ with nonnegative entries is a representation of the Sturm monoid.

Sturmian Morphisms

Sturmian morphisms have also the following local property:

Theorem 27 (Berstel, Séébold, 1994)

A morphism μ is Sturmian if and only if it is acyclic (i.e., $\mu(01) \neq \mu(10)$) and $\mu(100100101001 01)$ is Sturmian.

Sturmian morphisms can also be used to give another characterization of circularly balanced words.

Proposition 28

A primitive binary word w is circularly balanced if and only if $w = \mu(0)$ for some Sturmian morphism μ.

Central Words

Definition 29

A word having coprime periods p and q and length $p + q - 2$ is called a central word.

Note that a word having coprime periods p and q and length greater than $p + q - 2$ must be a power of a single letter by the theorem of Fine and Wilf.

Central words are in fact Sturmian words.

Central Words

Central words have several characterizations.

Proposition 30

Let v be a word in Σ_2. The following are equivalent:

1. *v is a central word;*
2. *v is a bispecial factor of some Sturmian word;*
3. *the words $0v1$ and $1v0$ are conjugate;*
4. *v is a palindrome and $v01$ is the product of two palindromes;*
5. *$0v1$ and $1v0$ are balanced;*
6. *v is a palindrome and $v0$ and $v1$ are balanced;*
7. *v is a power of a single letter or there exist P and Q such that $v = PxyQ = QyxP$, where $\{x, y\} = \Sigma_2$. Moreover, in this latter case, P and Q are central words.*

Central Words

As a consequence of the previous proposition, we have the following

Lemma 31

If v is a central word, then so is v^n for every $n \geq 1$.

To construct a central word with periods p and q, take p' and q', the multiplicative inverses of p and q modulo $p+q$, sort the positive multiples of p' and q' smaller than $p'q'$, then write 0 for each multiple of p' and 1 for each multiple of q'.

Example 32

Let $p = 4$, $q = 7$. Then $p' = 3$ and $q' = 8$, since $3 \cdot 4 = 1 \mod 11$ and $7 \cdot 8 = 1 \mod 11$. The central word having periods 4 and 7 (and length $4 + 7 - 2$) is, up to renaming letters, the word 001000100.

3	6	**8**	9	12	15	**16**	18	21
0	0	1	0	0	0	1	0	0

Standard Words

A standard word is a Sturmian word of the form vxy, with v a central word and $xy \in \{01, 10\}$.

Standard words are precisely the words that appear in some standard sequence (of a characteristic Sturmian word).

For example, Fibonacci finite words are standard words.

Standard Words

Another way to define standard words, in a recursive fashion, is by defining the standard pairs. The pair $(0, 1)$ is a standard pair; if (u, v) is a standard pair, then so are the pairs (u, uv) and (vu, v). Standard words are then those that appear in a standard pair.

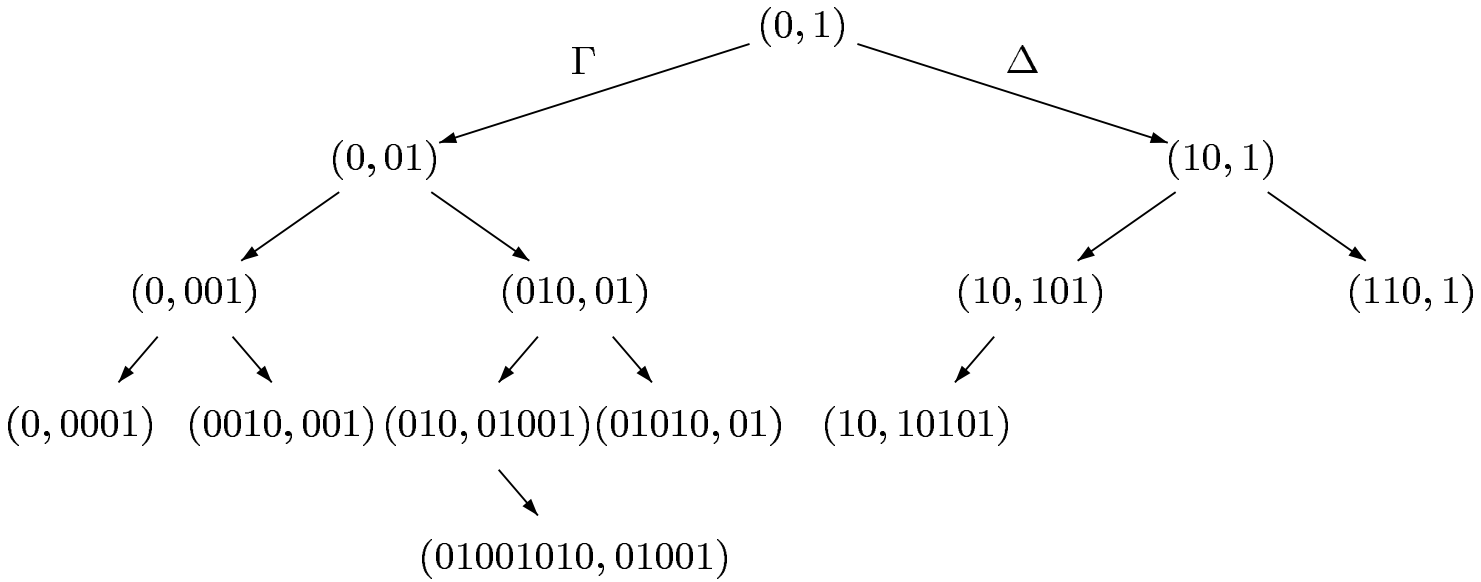


Figure: The tree of standard pairs.

Special Sturmian Words

If one considers extensibility within the set *St* of Sturmian words, one can define left special Sturmian words (resp., right special Sturmian words) as those words w over the alphabet $\Sigma_2 = \{0, 1\}$ such that $0w$ and $1w$ (resp., $w0$ and $w1$) are both Sturmian words.

For example, the word 001 is left special since 0001 and 1001 are both Sturmian words, but is not right special since 0011 is not a Sturmian word.

The Sturmian words that are both left and right special are called bispecial Sturmian words. They are of two kinds:

1. strictly bispecial Sturmian words (SBS), that are the words w such that $0w0$, $0w1$, $1w0$ and $1w1$ are all Sturmian words (e.g. 00), or
2. non-strictly bispecial Sturmian words (NBS) otherwise (e.g. 01).

Special Sturmian Words

Theorem 33 (Berstel, de Luca, 1997)

A word u is a strictly bispecial Sturmian word if and only if $0u1$ is a balanced Lyndon word.

This correspondence in fact holds more generally between bispecial Sturmian words and (powers of) balanced Lyndon words. More precisely, one has:

Theorem 34

A word u is a bispecial Sturmian word if and only if there exist letters x, y in $\{0, 1\}$ such that xuy is a power of a balanced Lyndon word or the reversal of a power of a balanced Lyndon word.

For example, $u = 01010010$ is bispecial but not strictly bispecial, since $1u0$ is not Sturmian; we have $0u1 = (00101)^2$.

Special Sturmian Words

As a corollary, we have that a bispecial Sturmian word is strictly bispecial if and only if it is a palindrome (hence a central word).

Since each central word of length n is associated with a pair (p, q) of coprime periods such that $p + q = n + 2$, there are $\varphi(n + 2)$ strictly bispecial Sturmian words of length n, where φ is the Euler totient function. That is,

$$SBS(n) = \varphi(n + 2).$$

Special Sturmian Words

Let w be a right special Sturmian word of length $n > 1$. If w is strictly bispecial, then $0w$ and $1w$ are right special Sturmian words of length $n+1$, otherwise only one between $0w$ and $1w$ is a right special Sturmian word of length $n+1$.

Therefore, the number $RS(n)$ of right special Sturmian words of length n verifies $RS(n+1) = SBS(n) + RS(n) = RS(n) + \varphi(n+2)$, hence

$$RS(n+1) = RS(1) + \sum_{i=2}^{n+1} \varphi(i+1).$$

Since $RS(1) = 2 = \varphi(1) + \varphi(2)$, we obtain

$$RS(n) = \sum_{i=1}^{n+1} \varphi(i).$$

Special Sturmian Words

Let w be a Sturmian word of length $n > 1$. If w is right special, then $w0$ and $w1$ are Sturmian words of length $n+1$, otherwise only one between $w0$ and $w1$ is a Sturmian word of length $n+1$.

Therefore, we have $St(n+1) = RS(n) + St(n)$, and hence, since $St(1) = 2$,

$$St(n) = 2 + \sum_{i=2}^{n} \sum_{j=1}^{i} \varphi(j) = 1 + \sum_{i=1}^{n} \sum_{j=1}^{i} \varphi(j) = 1 + \sum_{i=1}^{n} (n+1-i)\varphi(i).$$

Counting Sturmian Words

So we proved the following

Theorem 35

The number of balanced binary words (i.e., finite Sturmian words) of length n is

$$St(n) = 1 + \sum_{i=1}^{n} (n + 1 - i)\varphi(i)$$

where φ is the Euler's totient function, that is the function that counts the number of integers between 1 and n that are coprime with n.

Essentially, $St(n)$ is the sequence whose second difference is $\varphi(n+2)$, where φ is the Euler totient.

n	1	2	3	4	5	6	7	8	9	10	11	12
$St(n)$	2	4	8	14	24	36	54	76	104	136	178	224
$RS(n)$	2	4	6	10	12	18	22	28	32	42	46	58
$\varphi(n+2)$	2	2	4	2	6	4	6	4	10	4	12	6

Christoffel Words

Infinite Sturmian words can be viewed as the digital approximations of Euclidean straight lines in the plane.

Given a point (p, q) in the grid $\mathbb{Z} \times \mathbb{Z}$, with $p, q > 0$, there exists a unique path that approximates from below (resp., from above) the Euclidean segment joining the origin $(0, 0)$ to the point (p, q). If one encodes horizontal and vertical unitary segments with the letters 0 and 1 respectively, one obtains the lower (resp. upper) Christoffel word, denoted by $w_{p,q}$ (resp., $w'_{p,q}$), uniquely associated with the pair (p, q).

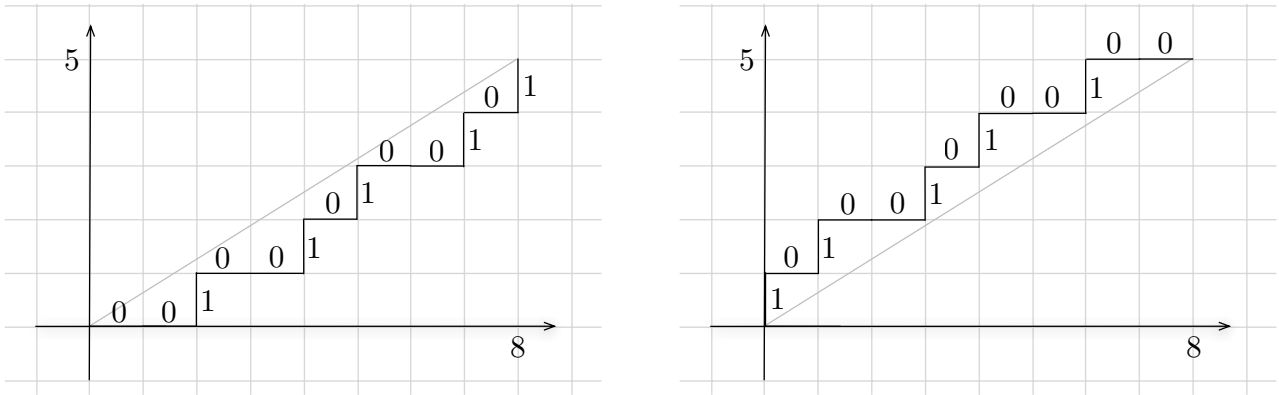


Figure: The lower Christoffel word $w_{8,5} = 0010010100101$ (left) and the upper Christoffel word $w'_{8,5} = 1010010100100$ (right).

Christoffel Words

By elementary geometrical considerations, one has that for any $p, q > 0$, $w_{p,q} = 0u1$ for some word u, and $w'_{p,q} = 1\widetilde{u}0$, where $\widetilde{u}$ is the reversal of u. If (and only if) p and q are coprime, the words $w_{p,q}$ and $w'_{p,q}$ are primitive.

If p and q are not coprime, the words $w_{p,q}$ and $w'_{p,q}$ are powers of primitive Christoffel words.

Lemma 36

For every pair of coprime integers (p, q) the upper Christoffel word $w'_{p,q}$ is the reversal of the lower Christoffel word $w_{p,q}$.

In a geometrical sense, Christoffel words are the finite approximations of mechanical words.

Christoffel Words

Actually, Christoffel words can be defined in a purely arithmetic way:

Definition 37

Let $n > 0$ and $p, q > 0$ be coprime integers such that $p + q = n$. The lower Christoffel word $w_{p,q} = w_1 w_2 \cdots w_n$ is the word defined by

$$w_i = \lfloor iq/(p+q) \rfloor - \lfloor (i-1)q/(p+q) \rfloor$$

i.e.,

$$w_i = \begin{cases} 0 & \text{if } iq \mod (n) > (i-1)q \mod (n) \\ 1 & \text{if } iq \mod (n) < (i-1)q \mod (n) \end{cases}$$

We call p/q the slope of $w_{p,q}$.[a]

[a] In the special case $q = 0$ we set the slope to be ∞.

Christoffel Words

Example 38

Let $p = 8$ and $q = 5$. We have

$$\{i5 \mod (13) \mid i = 0, 1, \ldots, 13\} = \{0, 5, 10, 2, 7, 12, 4, 9, 1, 6, 11, 3, 8, 0\}.$$

Hence, the lower Christoffel word of slope $5/8$ is $w_{8,5} = 0010010100101$.

Remark 39

Notice that at each step, either we add q, or we subtract p, and we have all integers between 1 and $n - 1$ exactly once.

Christoffel Words

Analogously, one can define the upper Christoffel word $w'_{p,q} = w'_1 w'_2 \cdots w'_n$ by

$$w'_i = \begin{cases} 0 & \text{if } ip \mod (n) < (i-1)p \mod (n) \\ 1 & \text{if } ip \mod (n) > (i-1)p \mod (n) \end{cases}$$

Of course, the upper Christoffel word $w'_{p,q}$ is the best grid approximation from above of the Euclidean segment joining $(0,0)$ to (p,q).

Example 40

Let $p = 8$ and $q = 5$. We have

$$\{i8 \mod (13) \mid i = 0, 1, \ldots, 13\} = \{0, 8, 3, 11, 6, 1, 9, 4, 12, 7, 2, 10, 5, 0\}.$$

(Notice that the numbers are the complements to n of the numbers in the sequence of the lower Christoffel word.) Hence, the upper Christoffel word of slope $5/8$ is $w'_{8,5} = 10100101001 00$.

Christoffel Words

From the definition, it follows that every point in the grid that belongs to the path encoded by a primitive Christoffel word of slope q/p has Euclidean distance smaller than $\sqrt{2}$ from the Euclidean segment joining $(0,0)$ to (p,q).

Consider the sequence $\{iq \mod (p+q)\}$, for $i = 0, 1, \ldots, p+q$, defining the lower Christoffel word $w_{p,q}$. Each subsequent number in the sequence is obtained by either adding q or subtracting p.

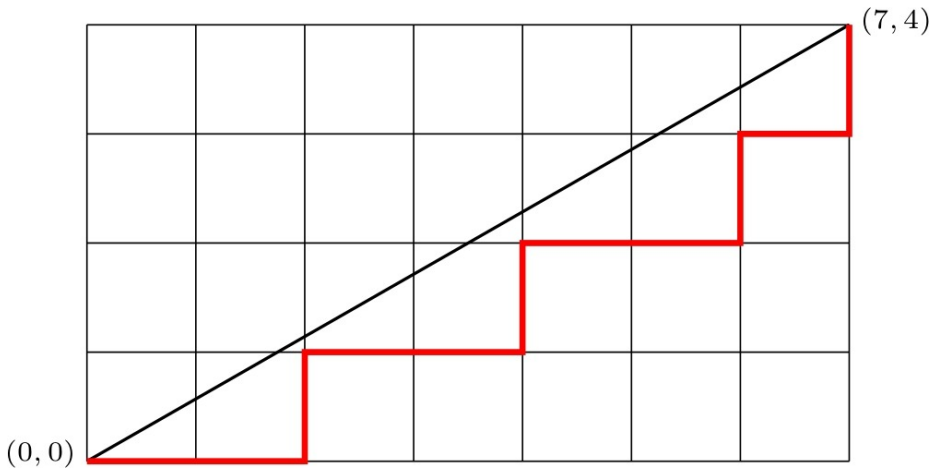
If we divide each term in the sequence by p, we get the sequence of vertical distances between the endpoints of paths encoded by prefixes of $w_{p,q}$ and the Euclidean segment joining $(0,0)$ to (p,q); if instead we divide by $\sqrt{p^2+q^2}$ (the length of the Euclidean segment) we get the sequence of Euclidean distances.

Christoffel Words

For example, for $w_{7,4}$, the sequence $\{4i \mod 11\}$ is

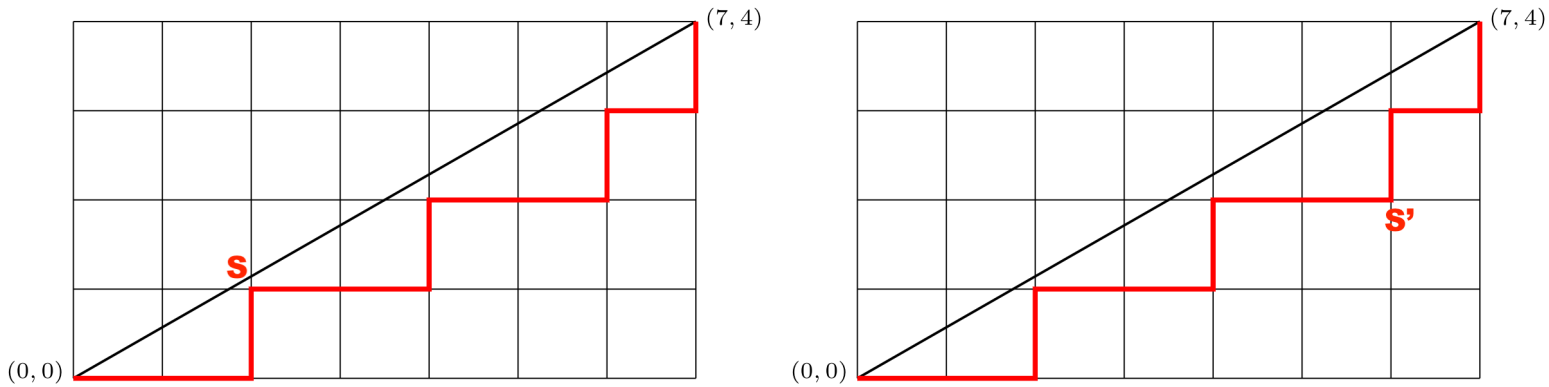
$$\{0, 4, 8, 1, 5, 9, 2, 6, 10, 3, 7, 0\}$$

Taking for example $i = 5$, we get that the point $(4, 1)$, which is the endpoint of the path corresponding to the prefix 00100 of $w_{7,4}$, has vertical distance $9/7 \approx 1.286$ and Euclidean distance $9/\sqrt{65} \approx 1.116$ from the Euclidean segment joining $(0, 0)$ and $(7, 4)$.



Christoffel Words

In particular, the point S on the path representing $w_{p,q}$ that is closest to the Euclidean segment (without lying on the segment itself) is at distance $1/\sqrt{p^2+q^2}$, whereas the point S' that is farthest from the segment is always at distance $(p+q-1)/\sqrt{p^2+q^2}$.



$$\{0, 4, 8, 1, 5, 9, 2, 6, 10, 3, 7, 0\}$$

Christoffel Words

Christoffel words have several characterizations.

Proposition 41

Let avb be a word in Σ_2, with $\{a, b\} = \Sigma_2$. The following are equivalent:

1. *avb is a (lower or upper) primitive Christoffel word;*
2. *v is a central word;*
3. *avb is balanced and unbordered;*
4. *$0v1$ is balanced and Lyndon (for the order $0 < 1$);*
5. *avb is a conjugate of bva.*

So, Christoffel words are precisely the unbordered Sturmian words, and lower Christoffel words are precisely the Lyndon Sturmian words.

Christoffel Words

So, central words are the "central" factors of primitive Christoffel words of length ≥ 2.

Moreover, we have:

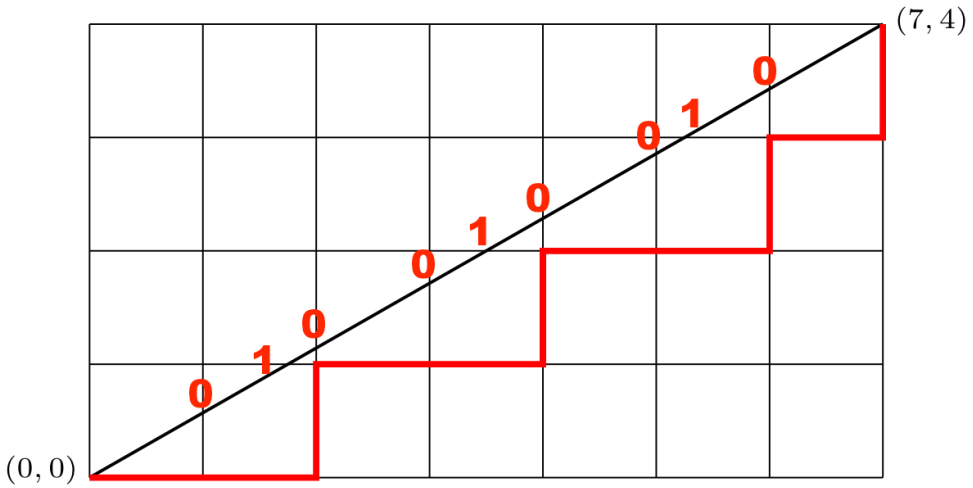
Proposition 42

Let $w_{p,q} = 0v1$ be a primitive lower Christoffel word. The central word v has periods p' and q', the multiplicative inverses of p and q modulo $p+q$, respectively (and length $p' + q' - 2 = p + q - 2$).

Christoffel Words

A geometric interpretation of the central word v in $w_{p,q} = 0v1$ is the following: it encodes the intersections of the Euclidean segment joining $(0,0)$ to (p,q) (0 for a vertical intersection and 1 for a horizontal intersection).

That is, the word v is the cutting sequence of the Euclidean segment joining $(0,0)$ to (p,q).



Christoffel Words

From the geometrical point of view, the lower and the upper Christoffel words with a given Parikh vector encode the frontiers of the region containing all Sturmian words with that Parikh vector (but there are also other words).

Theorem 43

Let $p, q > 0$ and $n = p + q$. Every finite Sturmian word with Parikh vector (p, q) encodes a path that is contained in the region delimited by the lower and the upper Christoffel words, i.e., stays at distance smaller than $\sqrt{2}$ from the Euclidean segment joining $(0, 0)$ to (p, q).

Christoffel Words

Example 44

Out of the $\binom{7+4}{4} = 330$ binary words with Parikh vector $(7, 4)$, only 112 of them encode paths that lie no more than $\sqrt{2}$ away from the Euclidean segment joining $(0, 0)$ to $(7, 4)$, i.e., are contained in the region delimited by the lower and the upper Christoffel words of Parikh vector $(7, 4)$.

In particular, all 19 balanced words of slope $4/7$ are among such approximations of the segment.

Christoffel Words

Proposition 45 (Borel, Laubie, 1993)

For every pair (p, q) of coprime positive integers, the lower Christoffel word $w_{p,q}$ is the greatest (in the lexicographic order) Lyndon word having Parikh vector (p, q).

For example, the Lyndon words of Parikh vector $(7, 4)$ are, in lexicographic order: 00000001111, 00000010111, 00000011011, 00000011101, 00000100111, 00000101011, 00000101101, 00000110011, 00000110101, 00000111001, 00001000111, 00001001011, 00001001101, 00001010011, 00001010101, 00001011001, 00001100011, 00001100101, 00001101001, 00001110001, 00010001011, 00010001101, 00010010011, 00010010101, 00010011001, 00010100011, 00010100101, 00010101001, 00011001001, $00100100101 = w_{7,4}$.

Christoffel Words

The following proposition, which follows from Theorem 43, is in some sense dual to Proposition 45.

Proposition 46

For every pair (p, q), the lower Christoffel word $w_{p,q}$ is the smallest (in the lexicographic order) finite Sturmian word having Parikh vector (p, q).

For example, the 19 Sturmian words with Parikh vector $(7, 4)$, in lexicographic order, are:
$w_{7,4} = 00100100101$, 00100101001, 00101001001, 00101001010, 00101010010, 00101010100, 01001001001, 01001001010, 01001010010, 01001010100, 01010010010, 01010010100, 10001001001, 10010001001, 10010010001, 10010010010, 10010010100, 10010100100, 10100100100

Christoffel Words

Recall that a word is conjugate to its reversal if and only if it is the concatenation of two palindromes and that a primitive word cannot have two factorizations as concatenations of two nonempty palindromes.

So we have:

Proposition 47

Every primitive Christoffel word has a unique factorization as a concatenation of two palindromes.

This factorization is called the palindromic factorization.

For example, the palindromic factorization $w_{7,4} = 00100100101$ is $00100100 \cdot 101$.

Christoffel Words

But since primitive lower Christoffel words are Lyndon words, we also have that every primitive lower Christoffel word longer than 1 has a unique factorization as a concatenation of two Lyndon words (actually, two Christoffel words).

This factorization is called the standard factorization.

For example, the standard factorization $w_{7,4} = 00100100101$ is $001 \cdot 00100101$.

Christoffel Words

Let $w_{p,q} = 0v1$ be a lower Christoffel word. If the central word v is not a power of a single letter, then there exist central words P and Q such that $v = P01Q = Q10P$ so that $w_{a,b} = 0v1 = 0P0 \cdot 1Q1 = 0Q1 \cdot 0P1$.

Hence, we have the factorizations:

1. $0v1 = 0P0 \cdot 1Q1$ (palindromic factorization);
2. $0v1 = 0Q1 \cdot 0P1$ (standard factorization).

If instead $v = 0^n$ (the case $v = 1^n$ is analogous) we have:

1. $0v1 = 0^{n+1} \cdot 1$ (palindromic factorization);
2. $0v1 = 0 \cdot 0^n 1$ (standard factorization).

Christoffel Words

The lengths of the two factors in both the palindromic and the standard factorization of the primitive lower Christoffel word $w_{p,q} = 0v1$ are precisely the two coprime periods of the central word v whose sum is $p + q$, i.e., the multiplicative inverses of p and q modulo $p + q$.

The two factorizations determine the point S and S', respectively.

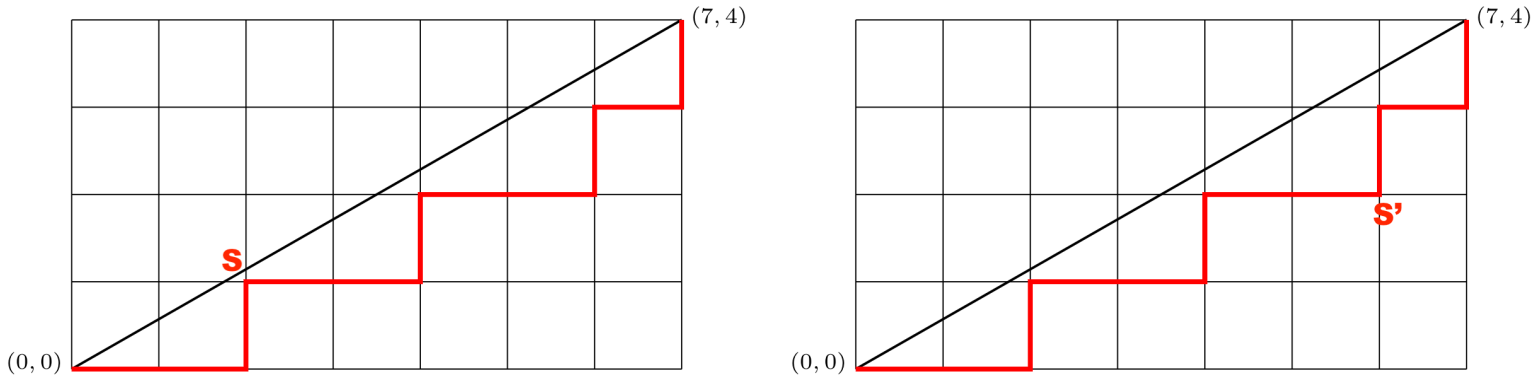


Figure: The standard factorization $0Q1 \cdot 0P1 = 001 \cdot 00100101$ (left) and the palindromic factorization $0P0 \cdot 1Q1 = 00100100 \cdot 101$ (right) of the lower Christoffel word $w_{7,4}$. The point S determined by the standard factorization is the closest to the Euclidean segment, while the point S' determined by the palindromic factorization is the farthest.