Linear Computation of Unbordered Conjugate

J.-P. Duval, T. Lecroq and A. Lefebvre

Journées du GDR IM - Groupe COMATEGE
Marne-la-Vallée
26 - 27 novembre 2012

Basic definitions

$$w = \mathsf{abaababaabaab}$$

Basic definitions

$$w = \underset{\substack{\uparrow \\ \text{prefix}}}{\boxed{\mathsf{a\,b\,a}}}\mathsf{a\,b\,a\,b\,a}\underset{\substack{\uparrow \\ \text{suffix}}}{\boxed{\mathsf{a\,b\,a\,a\,b}}}$$

Basic definitions

$$w = \underset{\substack{\uparrow \\ \text{prefix}}}{\boxed{\mathsf{a\,b\,a}}}\mathsf{a\,b\,a\,b\,a}\underset{\substack{\uparrow \\ \text{suffix}}}{\boxed{\mathsf{a\,b\,a\,a\,b}}}$$

Definition

$\texttt{Pref}(w)$ is the set of proper prefixes of w.
$\texttt{Suff}(w)$ is the set of proper suffixes of w.

Basic definitions

$$w = \underset{\substack{\uparrow \\ \text{prefix}}}{\boxed{\mathsf{a\,b\,a}}}\,\mathsf{a\,b\,a\,b\,a}\,\underset{\substack{\uparrow \\ \text{suffix}}}{\boxed{\mathsf{a\,b\,a\,a\,b}}}$$

Definition

A word w is unbordered iff $\mathtt{Pref}(w) \cap \mathtt{Suff}(w) = \emptyset$.

Basic definitions

$$w = \boxed{\mathsf{abaab}}\,\mathsf{aba}\,\boxed{\mathsf{abaab}}$$

border

Definition

A word w is unbordered iff $\texttt{Pref}(w) \cap \texttt{Suff}(w) = \emptyset$.

Pure words

Definition

A word is *pure* iff it is unbordered or the integer power of an unbordered word.

Pure words

Definition

A word is *pure* iff it is unbordered or the integer power of an unbordered word.

Example

- abaabb
- TINTIN = $(\texttt{TIN})^2$

Conjugates

Definition

A word w' is a conjugate of word w iff it is equal to a circular permutation of symbols of w.

Example

`babaab` and `bbabaa` are conjugates of `abaabb`.

Our problem

Given word w, our problem is to find a pure conjugate of w (without considering any lexicographic order).

Example

- `babaababaabaa` is a pure conjugate of `abaababaabaab`.
- `TINTIN` is a pure conjugate of itself.

Relation between words

Definition

Given two words x and y, $x \ll y$ iff the following conditions hold:

- y is unbordered;
- $|x| \geqslant |y|$;
- the prefix of x of length $|y|$ is a concatenation of prefixes of y.

Example

$x = aababab$ and $y = abb$

Relation between words

Proposition

Given u an unbordered word and $v \in \mathtt{Pref}^+(u)$. The following properties hold:

- $v \cdot u$ *is an unbordered word;*
- $v \cdot u \ll u$.

Corollary

Given an unbordered word u and $v \in \mathtt{Pref}^+(u)$, for all integer $q \geqslant 1$, $v \cdot u^q$ is unbordered.

Relation between words

Lemma

Let x and y be two unbordered words such that $x \ll y$. If $y' \in \mathit{Suff}(y)$ then $x \cdot y'$ is unbordered.

Corollary

Let x be a unbordered word and $z = z_1 \cdot z_2 \cdots z_i \cdots z_m$ be a concatenation of unbordered words or suffix of unbordered words such that $x \ll z_i$ for $1 \leqslant i \leqslant m$. Then $x \cdot z$ is unbordered.

Decomposition in unbordered words

Definition

We say that $(u_1, q_1) \cdots (u_m, q_m) \cdot (u, q) \cdot (v)$ is a decomposition of w in unbordered words iff $u_1, \ldots, u_m$ and u are unbordered words and $w = u_1^{q_1} \cdots u_m^{q_m} \cdot u^q \cdot v$ with $v \in \texttt{Pref}^*(u)$.

Canonical decomposition in unbordered words

We call *canonical decomposition in unbordered words* of w the decomposition defined as follows.

$$w[1\,..\,1] = (w[1], 1) \cdot ()$$

If $w[1\,..\,j] = (u_1, q_1) \cdots (u_m, q_m) \cdot (u, q) \cdot (v)$ and $w[j+1] = a$ then

$$w[1\,..\,j+1] =$$
$$\begin{cases} (u_1, q_1) \cdots (u_m, q_m) \cdot (u, q+1) \cdot () & \text{if } v \cdot a = u & (1) \\ (u_1, q_1) \cdots (u_m, q_m) \cdot (u, q) \cdot (v \cdot a) & \text{if } v \cdot a \in \mathtt{Pref}(u) & (2) \\ (u_1, q_1) \cdots (u_m, q_m) \cdot (u, q) \cdot (v \cdot a) & \text{if } v \cdot a \in \mathtt{Pref}^{+}(u) \setminus \mathtt{Pref}(u) & (3) \\ (u_1, q_1) \cdots (u_m, q_m) \cdot (u, q) \cdot (v \cdot a, 1)() & \text{if } v \cdot a \in \mathtt{Pref}^{+}(u) \cdot u & (4) \\ (u_1, q_1) \cdots (u_m, q_m) \cdot (u^q \cdot v \cdot a, 1)() & \text{if } v \cdot a \notin \mathtt{Pref}^{+}(u) \cup \mathtt{Pref}^{*}(u) \cdot u & (5) \end{cases}$$

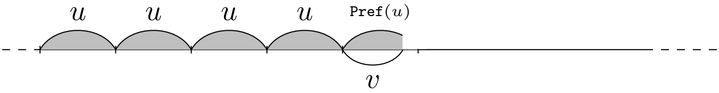
Morris-Pratt failure function

- This decomposition is computed using a Morris-Pratt failure function.
- This function enables to compute, for each prefix of a word, the length of its longest border.
- Its complexity is linear.

Canonical decomposition in unbordered words

u is unbordered and $v \in \texttt{Pref}(u)$.

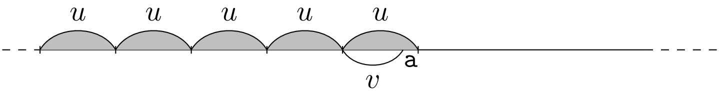
(1)



Canonical decomposition in unbordered words

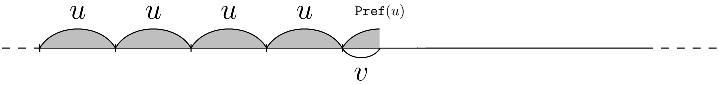
u is unbordered and $v \in \texttt{Pref}(u)$.

(1)



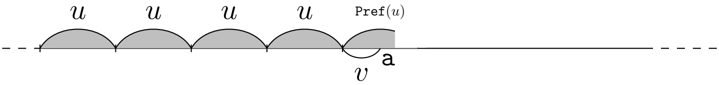
Canonical decomposition in unbordered words

u is unbordered and $v \in \texttt{Pref}(u)$.

(2) 

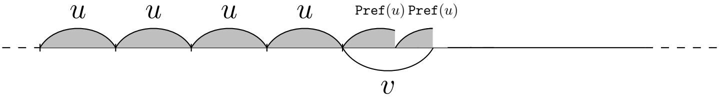
Canonical decomposition in unbordered words

u is unbordered and $v \in \texttt{Pref}(u)$.

(2) 

Canonical decomposition in unbordered words

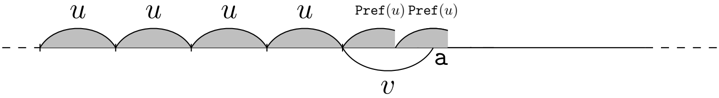
u is unbordered and $v \in \texttt{Pref}^{+}(u)$.

(3) 

Canonical decomposition in unbordered words

u is unbordered and $v \in \texttt{Pref}^+(u)$.

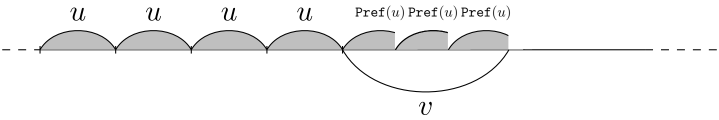
(3)



Canonical decomposition in unbordered words

u is unbordered and $v \in \texttt{Pref}^+(u)$.

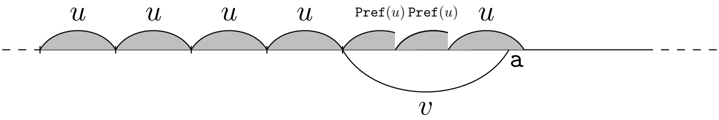
(4)



Canonical decomposition in unbordered words

u is unbordered and $v \in \mathtt{Pref}^+(u)$.

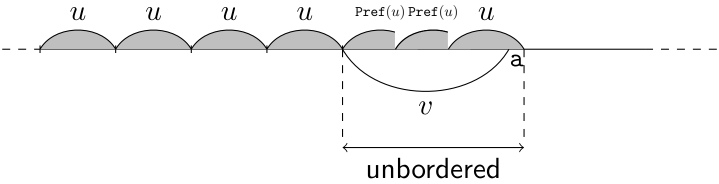
(4)



Canonical decomposition in unbordered words

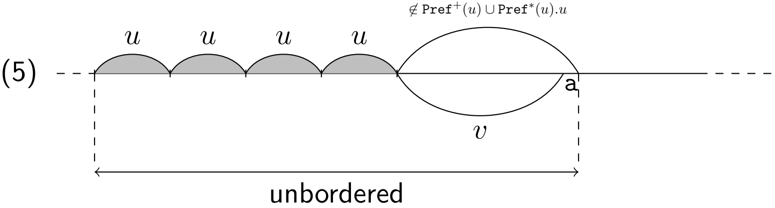
u is unbordered and $v \in \mathtt{Pref}^{+}(u)$.

(4)



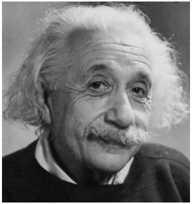
Canonical decomposition in unbordered words

u is unbordered and $v \in \texttt{Pref}^+(u)$.



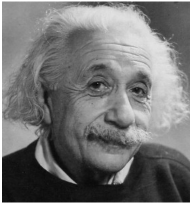
Example:

Example: "EINSTEIN"



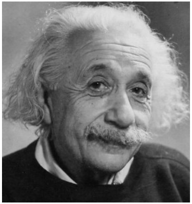
EINSTEIN

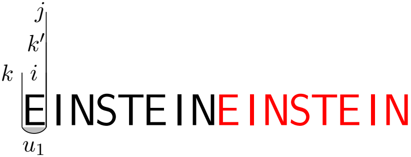
Example: "EINSTEIN"



EINSTEINEINSTEIN

Example: "EINSTEIN"





j

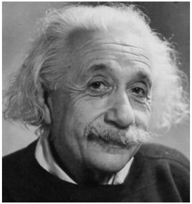
k'

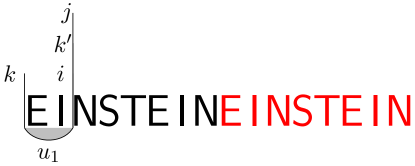
k i

EINSTEINEINSTEIN

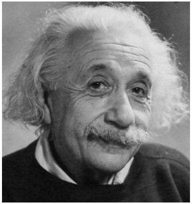
u_1

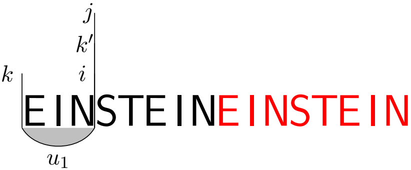
Example: "EINSTEIN"



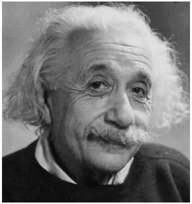


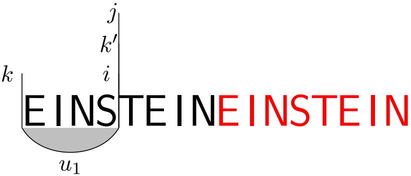
Example: "EINSTEIN"



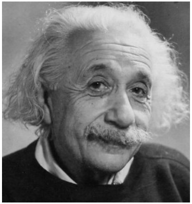


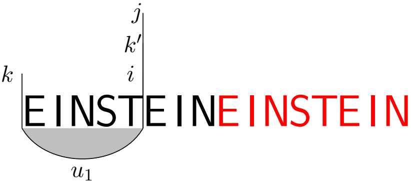
Example: "EINSTEIN"



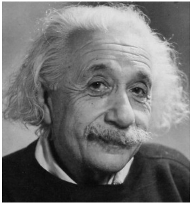


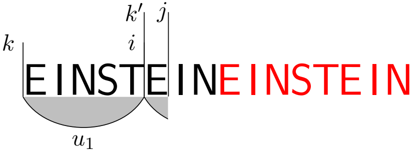
Example: "EINSTEIN"



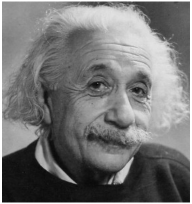


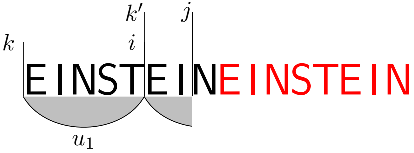
Example: "EINSTEIN"





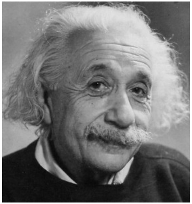
Example: "EINSTEIN"

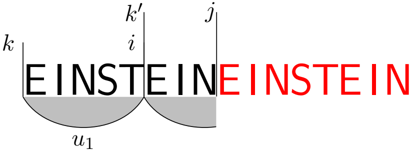




EINSTEINEINSTEIN

Example: "EINSTEIN"





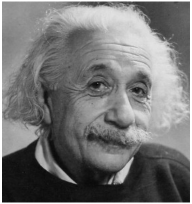
k' j

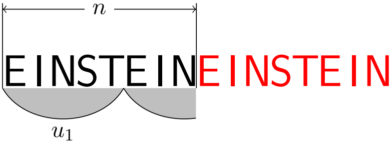
k i

EINSTEINEINSTEIN

u_1

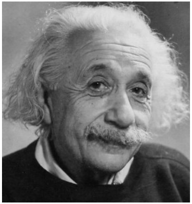
Example: "EINSTEIN"

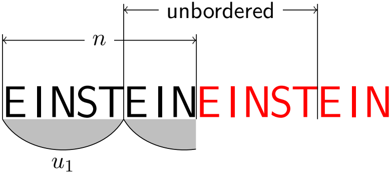




EINSTEINEINSTEIN

Example: "EINSTEIN"





A classical case:

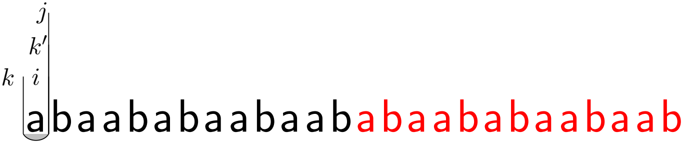
A classical case: "Fibonacci"

abaababaabaab

A classical case: ”Fibonacci”

abaababaabaababaababaabaab

A classical case: ”Fibonacci”



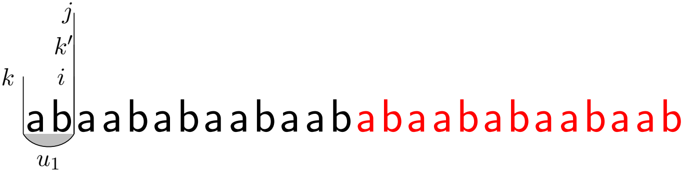
j

k'

k i

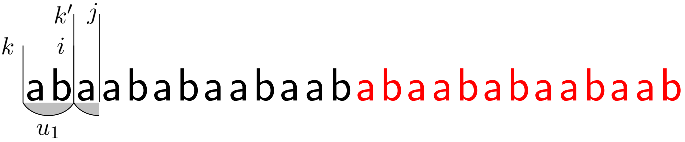
abaababaabaab abaababaabaab

A classical case: "Fibonacci"



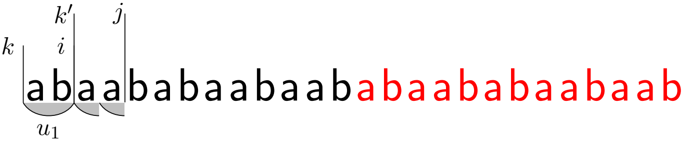
abaababaabaababaababaabaab

A classical case: "Fibonacci"

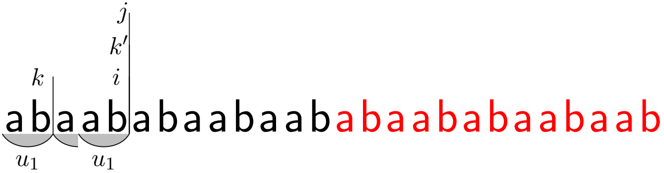


abaababaabaababaababaabaab

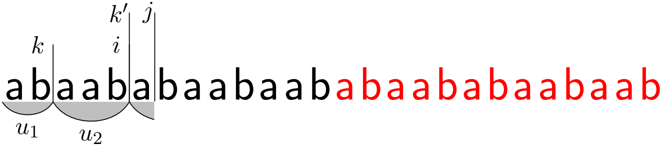
A classical case: "Fibonacci"



A classical case: "Fibonacci"

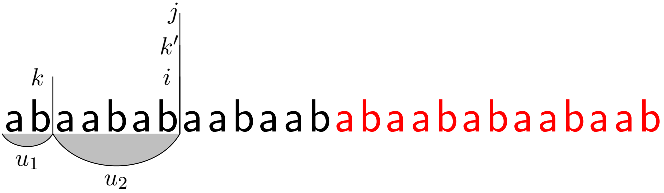


A classical case: "Fibonacci"



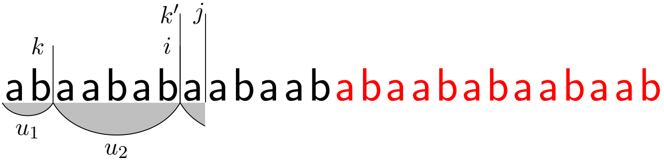
$$u_2 \ll u_1$$

A classical case: "Fibonacci"



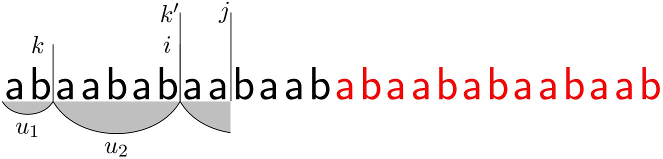
$u_2 \ll u_1$

A classical case: "Fibonacci"



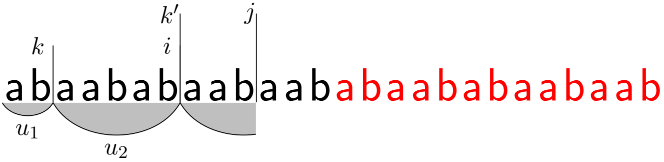
$u_2 \ll u_1$

A classical case: "Fibonacci"



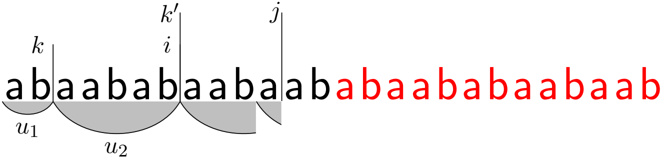
$$u_2 \ll u_1$$

A classical case: "Fibonacci"



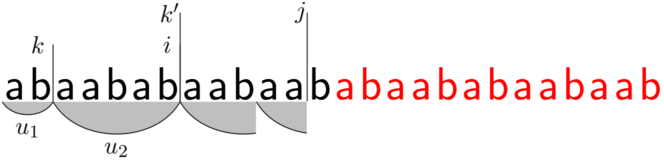
$$u_2 \ll u_1$$

A classical case: "Fibonacci"



$u_2 \ll u_1$

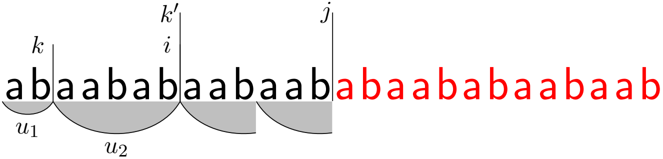
A classical case: "Fibonacci"



abaababaabaababaababaabaab

$u_2 \ll u_1$

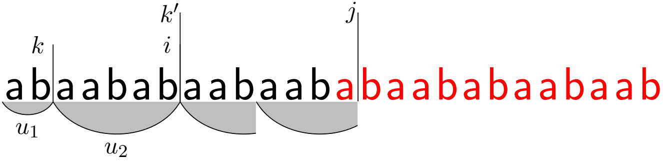
A classical case: "Fibonacci"



abaababaabaababaababaabaab

$u_2 \ll u_1$

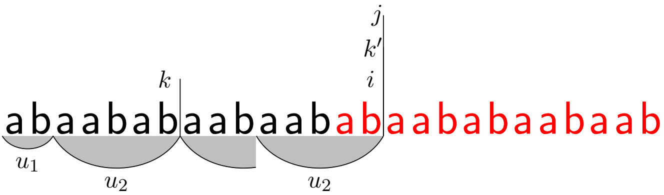
A classical case: "Fibonacci"



abaababaabaababaababaabaab

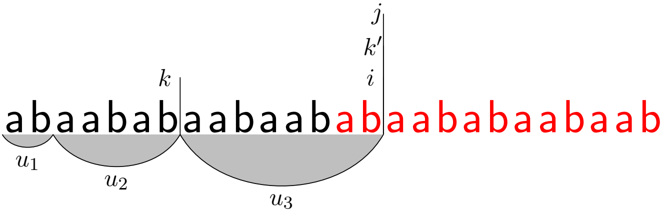
$u_2 \ll u_1$

A classical case: "Fibonacci"



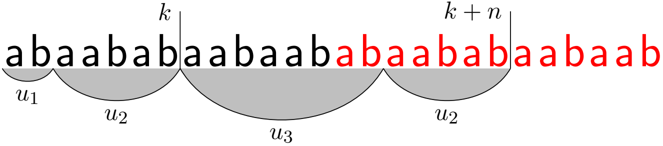
$u_2 \ll u_1$

A classical case: "Fibonacci"



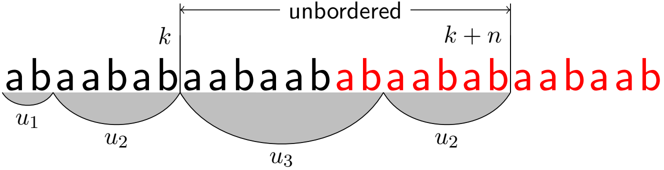
$$u_3 \ll u_2 \ll u_1$$

A classical case: "Fibonacci"



$$u_3 \ll u_2 \ll u_1$$

A classical case: "Fibonacci"



unbordered

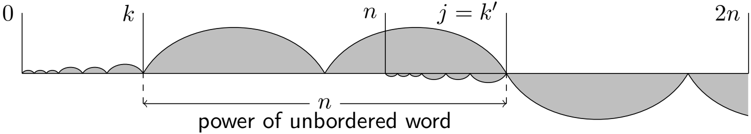
k $k+n$

abaababaabaab<span style="color:red">abaababaabaab</span>

u_1 u_2 u_3 u_2

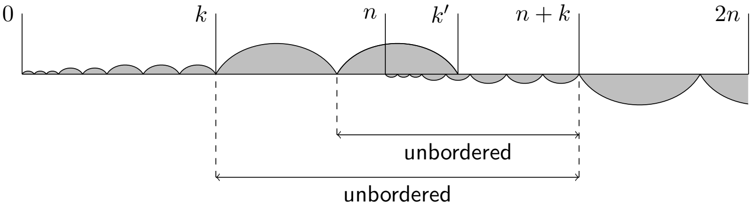
$$u_3 \ll u_2 \ll u_1$$

Stop case 1



Case where $w[k+1\,.\,.\,k+n]$ is a power of an unbordered word.

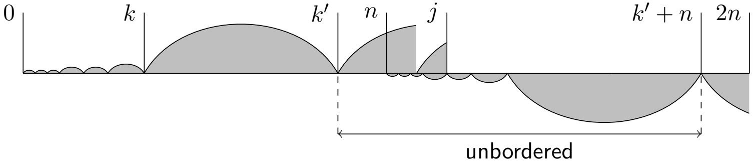
Stop case 2



Corollary

Let x be a unbordered word and $z = z_1 \cdot z_2 \cdots z_i \cdots z_m$ be a concatenation of unbordered words or suffix of unbordered words such that $x \ll z_i$ for $1 \leqslant i \leqslant m$. Then $x \cdot z$ is unbordered.

Stop case 3



Corollary

Given an unbordered word u and $v \in \mathtt{Pref}^{+}(u)$, for all integer $q \geqslant 1$, $v \cdot u^{q}$ is unbordered.

Complexity

- two copies of w are necessary
- the exact number of symbols comparisons perfomed by the Morris-Pratt failure function f' on a word of length n is bounded by $2n - 2 - f'(n) - r$ where $r = \#\{j | f'(j) = 0\}$
- the total complexity is bounded by $4n$

The end

Questions?