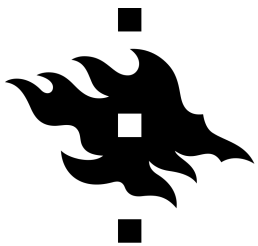
Tighter Bounds for the Sum of Irreducible LCP Values

Juha Kärkkäinen[1] Dominik Kempa[1] **Marcin Piątkowski**[2]

[1] University of Helsinki



[2] Nicolaus Copernicus University

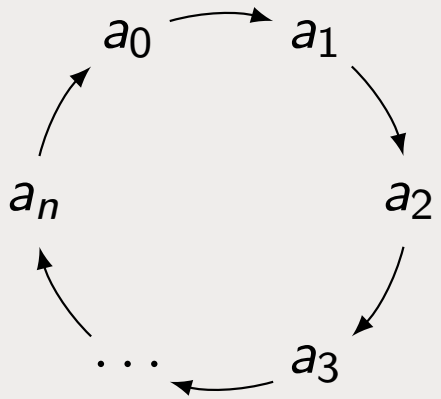
Outline

1. Cyclic words
2. Irreducible LCP values
3. Upper bound for the sum of irreducible values
4. Lower bound for the sum of irreducible values

Cyclic suffixes

$$W = \{\!\{v_1, v_2, v_3\}\!\} = \{\!\{aab, aab, ab\}\!\}$$

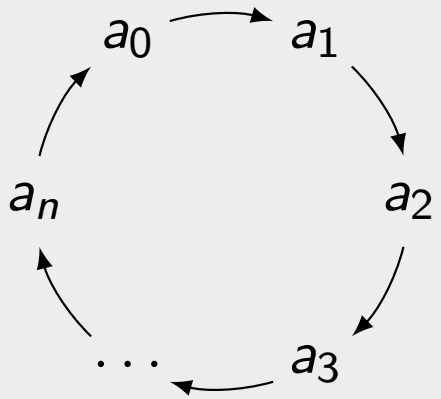
	$\mathrm{suf}(W)$
$\langle 1,0\rangle$	**a a b** a a b $\cdots$
$\langle 1,1\rangle$	**a b a** a b a $\cdots$
$\langle 1,2\rangle$	**b a a** b a a $\cdots$
$\langle 2,0\rangle$	**a a b** a a b $\cdots$
$\langle 2,1\rangle$	**a b a** a b a $\cdots$
$\langle 2,2\rangle$	**b a a** b a a $\cdots$
$\langle 3,0\rangle$	**a b** a b a b $\cdots$
$\langle 3,1\rangle$	**b a** b a b a $\cdots$



Cyclic suffix array

$$W = \{\!\{v_1, v_2, v_3\}\!\} = \{\!\{aab, aab, ab\}\!\}$$

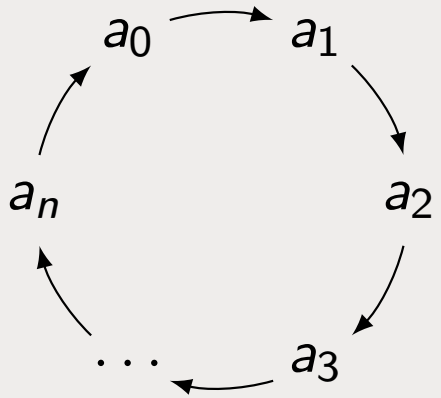
SA_W	$\mathrm{suf}(W)$
$\langle 1,0\rangle$	**a a b** a a b $\cdots$
$\langle 2,0\rangle$	**a a b** a a b $\cdots$
$\langle 1,1\rangle$	**a b a** a b a $\cdots$
$\langle 2,1\rangle$	**a b a** a b a $\cdots$
$\langle 3,0\rangle$	**a b** a b a b $\cdots$
$\langle 1,2\rangle$	**b a a** b a a $\cdots$
$\langle 2,2\rangle$	**b a a** b a a $\cdots$
$\langle 3,1\rangle$	**b a** b a b a $\cdots$



Cyclic suffix array

$$W = \{\!\{v_1, v_2, v_3\}\!\} = \{\!\{aab, aab, ab\}\!\}$$

SA_W	$\mathrm{suf}(W)$
$\langle 1,0\rangle$	**a a b** a a b $\cdots$
$\langle 2,0\rangle$	**a a b** a a b $\cdots$
$\langle 1,1\rangle$	**a b a** a b a $\cdots$
$\langle 2,1\rangle$	**a b a** a b a $\cdots$
$\langle 3,0\rangle$	**a b** a b a b $\cdots$
$\langle 1,2\rangle$	**b a a** b a a $\cdots$
$\langle 2,2\rangle$	**b a a** b a a $\cdots$
$\langle 3,1\rangle$	**b a** b a b a $\cdots$



Longest common prefix and distinguishing prefix arrays

$$W = \{\!\{v_1, v_2, v_3\}\!\} = \{\!\{aab, aab, ab\}\!\}$$

SA_W	$\mathrm{suf}(W)$	LCP_W	DP_W
$\langle 1,0\rangle$	**a a b** a a b $\cdots$	$-$	$-$
$\langle 2,0\rangle$	**a a b** a a b $\cdots$	∞	∞
$\langle 1,1\rangle$	**a b a** a b a $\cdots$	1	2
$\langle 2,1\rangle$	**a b a** a b a $\cdots$	∞	∞
$\langle 3,0\rangle$	**a b** a b a b $\cdots$	3	4
$\langle 1,2\rangle$	**b a a** b a a $\cdots$	0	1
$\langle 2,2\rangle$	**b a a** b a a $\cdots$	∞	∞
$\langle 3,1\rangle$	**b a** b a b a $\cdots$	2	3

Burrows-Wheeler transform

$$W = \{\!\{v_1, v_2, v_3\}\!\} = \{\!\{aab, aab, ab\}\!\}$$

SA_W	$\mathrm{suf}(W)$	LCP_W	DP_W	$BWT(W)$
$\langle 1,0\rangle$	**a a b** a a b $\cdots$	–	–	b
$\langle 2,0\rangle$	**a a b** a a b $\cdots$	∞	∞	b
$\langle 1,1\rangle$	**a b a** a b a $\cdots$	1	2	a
$\langle 2,1\rangle$	**a b a** a b a $\cdots$	∞	∞	a
$\langle 3,0\rangle$	**a b** a b a b $\cdots$	3	4	b
$\langle 1,2\rangle$	**b a a** b a a $\cdots$	0	1	a
$\langle 2,2\rangle$	**b a a** b a a $\cdots$	∞	∞	a
$\langle 3,1\rangle$	**b a** b a b a $\cdots$	2	3	a

Cyclic equivalence

Cyclic equivalence

Two multisets of words V and W are cyclically equivalent if

$$\mathrm{suf}(V) = \mathrm{suf}(W).$$

Example multiset

$W = \{\!\{aab,\ aab,\ ab\}\!\}$

$aabaabaabaab\cdots$
$abaabaabaaba\cdots$
$baabaabaabaa\cdots$
$aabaabaabaab\cdots$
$abaabaabaaba\cdots$
$baabaabaabaa\cdots$
$abababababab\cdots$
$babababababa\cdots$

Equivalence class

$\{\!\{aab,\ aab,\ ab\}\!\}$	$\{\!\{aab,\ aab,\ ba\}\!\}$
$\{\!\{aba,\ aab,\ ab\}\!\}$	$\{\!\{aba,\ aab,\ ba\}\!\}$
$\{\!\{baa,\ aab,\ ab\}\!\}$	$\{\!\{baa,\ aab,\ ba\}\!\}$
$\{\!\{aab,\ aba,\ ab\}\!\}$	$\{\!\{aab,\ aba,\ ba\}\!\}$
$\vdots$	$\vdots$
$\{\!\{aabaab,\ ab\}\!\}$	$\{\!\{aabaab,\ ba\}\!\}$
$\{\!\{abaaba,\ ab\}\!\}$	$\{\!\{abaaba,\ ba\}\!\}$
$\{\!\{baabaa,\ ab\}\!\}$	$\{\!\{baabaa,\ ba\}\!\}$

Cyclic equivalence

Lemma

Let $W = \{\!\{w_i\}\!\}_{i=1}^{s}$ be a multiset of cyclic words. Then:

1. There exists a set of cyclic words $V = \{v_i\}_{i=1}^{t}$ such that $\mathrm{suf}(W) = \mathrm{suf}(V)$.
2. There exists a multiset of primitive cyclic words $U = \{\!\{u_i\}\!\}_{i=1}^{p}$ such that $\mathrm{suf}(W) = \mathrm{suf}(V)$.

Cyclic equivalence

Lemma

Let $W = \{\!\{w_i\}\!\}_{i=1}^{s}$ be a multiset of cyclic words. Then:

1. There exists a set of cyclic words $V = \{v_i\}_{i=1}^{t}$ such that $\mathrm{suf}(W) = \mathrm{suf}(V)$.
2. There exists a multiset of primitive cyclic words $U = \{\!\{u_i\}\!\}_{i=1}^{p}$ such that $\mathrm{suf}(W) = \mathrm{suf}(V)$.

Remark

If two multisets of words V and W are cyclically equivalent, then $\mathrm{LCP}_V = \mathrm{LCP}_W$, $\mathrm{DP}_V = \mathrm{DP}_W$ and $\mathrm{BWT}_V = \mathrm{BWT}_W$.

Cyclic equivalence

Lemma

Let $W = \{\!\{w_i\}\!\}_{i=1}^{s}$ be a multiset of cyclic words. Then:

1. There exists a set of cyclic words $V = \{v_i\}_{i=1}^{t}$ such that $\mathrm{suf}(W) = \mathrm{suf}(V)$.
2. There exists a multiset of primitive cyclic words $U = \{\!\{u_i\}\!\}_{i=1}^{p}$ such that $\mathrm{suf}(W) = \mathrm{suf}(V)$.

Remark

If two multisets of words V and W are cyclically equivalent, then $\mathrm{LCP}_V = \mathrm{LCP}_W$, $\mathrm{DP}_V = \mathrm{DP}_W$ and $\mathrm{BWT}_V = \mathrm{BWT}_W$.

Theorem (Mantaci, Restivo, Rosone, Sciortino – 2007)

The mapping from a word v to the cyclical equivalence class of $\mathrm{IBWT}(v)$ is a bijection

Irreducible LCP and DP values

Irreducible values

A value $\mathrm{LCP}_W[i]$ (respectively $\mathrm{DP}_W[i]$) is irreducible if $\mathrm{BWT}_W[i] \neq \mathrm{BWT}_W[i-1]$.

SA_W	$\mathrm{suf}(W)$	LCP_W	DP_W	$BWT(W)$
$\langle 1,0\rangle$	**a a b** a a b $\cdots$	–	–	b
$\langle 2,0\rangle$	**a a b** a a b $\cdots$	∞	∞	b
$\langle 1,1\rangle$	**a b a** a b a $\cdots$	**1**	**2**	a
$\langle 2,1\rangle$	**a b a** a b a $\cdots$	∞	∞	a
$\langle 3,0\rangle$	**a b** a b a b $\cdots$	**3**	**4**	b
$\langle 1,2\rangle$	**b a a** b a a $\cdots$	**0**	**1**	a
$\langle 2,2\rangle$	**b a a** b a a $\cdots$	∞	∞	a
$\langle 3,1\rangle$	**b a** b a b a $\cdots$	2	3	a

Sum of irreducible LCP values

Irreducible values

- $W = \{\!\{w_i\}\!\}_{i=1}^{s}$ and $\sum_{i=1}^{n} |w_i| = n$
- $\Sigma\mathrm{ilcp}(W)$ – sum of irreducible LCP values
- $\Sigma\mathrm{idp}(W)$ – sum of distinguishing prefixes lengths

Theorem (Kärkkäinen, Manzini, Puglisi – 2009)

$$\Sigma\mathrm{ilcp}(W) = O(n \log n)$$

New upper bounds for the sum of irreducible lcp values

Theorem

For any multiset W of words of total length $n > 0$, we have

$$\Sigma\mathrm{ilcp}(W) \leqslant \Sigma\mathrm{idp}(W) \leqslant n\lceil \lg n \rceil - 2^{\lceil \lg n \rceil} + 1$$

Theorem

For any multiset W of words of total length $n > 0$ such that $\mathrm{BWT}(W)$ has r runs, we have

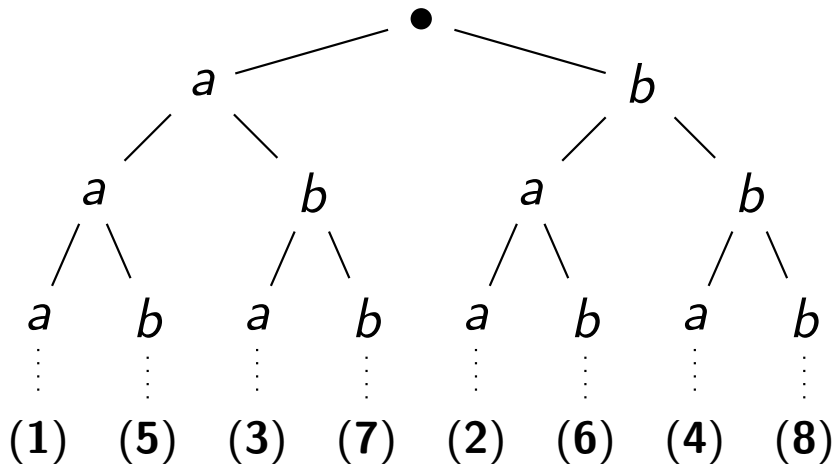
$$\Sigma\mathrm{ilcp}(W) + r - 1 = \Sigma\mathrm{idp}(W) \leqslant n\lceil \lg r \rceil - 2^{\lceil \lg r \rceil} + 1$$

Reverse suffixes

$W = \{\!\{a, aab, abb, b\}\!\}$

(1) $aaaaaa\cdots$
(2) $baabaa\cdots$
(3) $abaaba\cdots$
(4) $bbabba\cdots$
(5) $aabaab\cdots$
(6) $babbab\cdots$
(7) $abbabb\cdots$
(8) $bbbbbb\cdots$

$\overline{\mathrm{STree}(W)}$

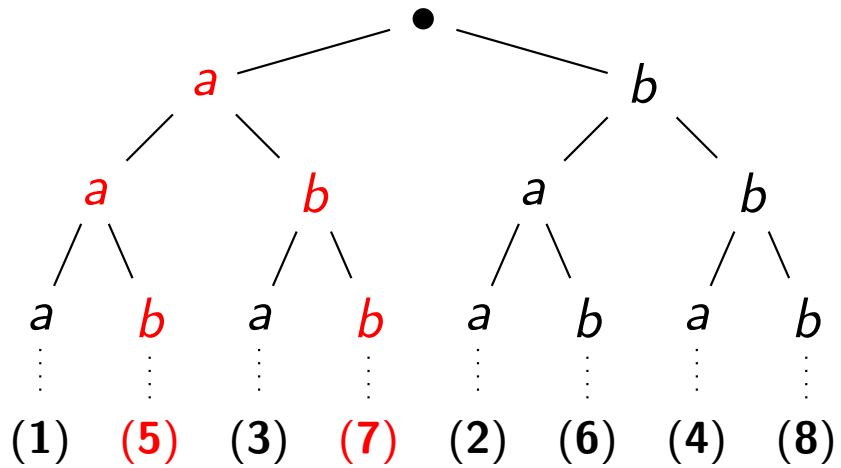


Reverse suffixes

$W = \{\!\{a, aab, abb, b\}\!\}$

(1) $aaaaaa\cdots$
(2) $baabaa\cdots$
(3) $abaaba\cdots$
(4) $bbabba\cdots$
(5) $aabaab\cdots$
(6) $babbab\cdots$
(7) $abbabb\cdots$
(8) $bbbbbb\cdots$

$\overline{\mathrm{STree}(W)}$



Dispersal pair with respect to $\leqslant$

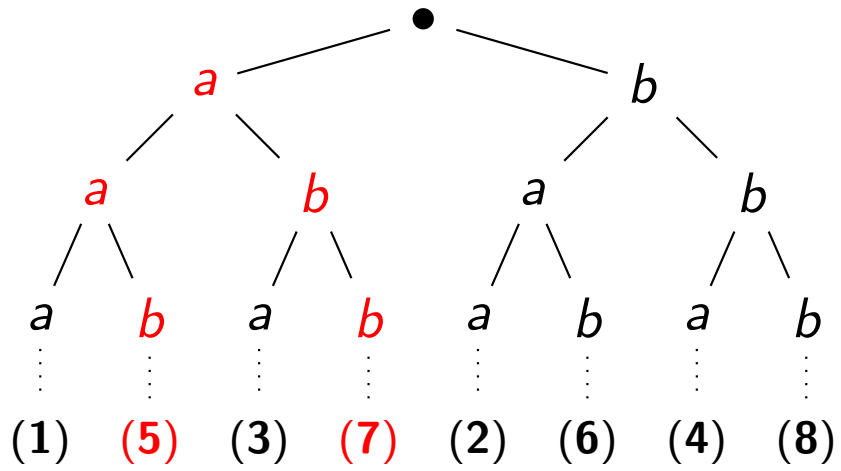
The pair of leaves (u, v) of an ordered tree is called a dispersal pair if $u < v$ the subtree rooted at their nearest common ancestor contains no leaf w such that $u < w < v$.

Reverse suffixes

$W = \{\!\{a, aab, abb, b\}\!\}$

(1) $aaaaaa\cdots$
(2) $baabaa\cdots$
(3) $abaaba\cdots$
(4) $bbabba\cdots$
(5) $aabaab\cdots$
(6) $babbab\cdots$
(7) $abbabb\cdots$
(8) $bbbbbb\cdots$

$\overline{\mathrm{STree}(W)}$



Lemma

$$D\left(\overline{\mathrm{STree}}(W), \leqslant_W\right) = \Sigma\mathrm{idp}(W)$$

$D(T, \leqslant)$ – the number of dispersal pairs in T

$n \log n$ upper bound

$$d(1) = 0$$
$$d(n) = \max_{i \in [1..\lfloor n/2 \rfloor]} d(n, i) \quad \text{when } n > 1$$
$$d(n, k) = d(k) + d(n - k) + \min\{2k, n - 1\}$$

where $n > 0$ and $k \in [0..\lfloor n/2 \rfloor]$.

Lemma

$d(n) = \max\{D(T, \leqslant)\}$, where the maximum is taken over any rooted tree T with n leaves and any total order $\leqslant$ on its leaves.

$n \log n$ upper bound

Lemma

$$d(n) = n\lceil \lg n \rceil - 2^{\lceil \lg n \rceil} + 1$$

Theorem

For any multiset W of words of total length $n > 0$, we have

$$\Sigma\mathrm{ilcp}(W) \leqslant \Sigma\mathrm{idp}(W) \leqslant d(n) \leqslant n\lceil \lg n \rceil - 2^{\lceil \lg n \rceil} + 1$$

$n \log r$ upper bound

Lemma

If $\mathrm{BWT}(W)$ has r runs, then $\left|D_u\left(\overline{\mathrm{STree}}(W), \leqslant_W\right)\right| < r$ for every vertex u in $\overline{\mathrm{STree}}(W)$.

$n \log r$ upper bound

$$d_r(1) = 0$$
$$d_r(n) = \max_{i \in [1..\lfloor n/2 \rfloor]} d_r(n, i) \quad \text{when } n > 1$$
$$d_r(n, k) = d_r(k) + d_r(n - k) + \min\{2k, n - 1, r - 1\}$$

where $r > 0$, $n > 0$ and $k \in [0..\lfloor n/2 \rfloor]$.

Lemma

$d_r(n) = \max\{d(T, \leqslant)\}$, where the maximum is taken over any rooted tree T with n leaves and any total order $\leqslant$ on its leaves s.t. $|D_u(T, \leqslant)| < r$ for every vertex $u \in T$.

$n \log r$ upper bound

Lemma

For any $2 \leqslant r \leqslant n$ we have $d_r(n) \leqslant n\lceil \lg r \rceil - 2^{\lceil \lg r \rceil} + 1$.

Theorem

For any multiset W of words of total length $n > 0$ such that $\mathrm{BWT}(W)$ has r runs, we have

$$\Sigma\mathrm{ilcp}(W) + r - 1 \;=\; \Sigma\mathrm{idp}(W) \;\leqslant\; d_r(n) \;\leqslant\; n\lceil \lg r \rceil - 2^{\lceil \lg r \rceil} + 1$$

De Bruijn set

A set of words $W = \{w_i\}_{i=1}^{t}$ over the alphabet $\mathcal{A}$ is a de Bruijn set of order k if $\sum_{i=1}^{t} |w_i| = 2^k$ and every word $v \in \mathcal{A}^k$ is a prefix of exactly one word in $\mathrm{suf}(W)$.

$$W = \{a,\ aab,\ abb,\ b\} \qquad \sum_{i=1}^{4} |w_i| = 2^3$$

$$\begin{aligned}
W_{\langle 1,1\rangle} &= \mathbf{a}\,\mathrm{a}\,\mathrm{a}\,\mathrm{a}\,\mathrm{a}\,\mathrm{a}\cdots \\
W_{\langle 2,1\rangle} &= \mathbf{a}\,\mathbf{a}\,\mathbf{b}\,\mathrm{a}\,\mathrm{a}\,\mathrm{b}\cdots \\
W_{\langle 2,3\rangle} &= \mathbf{a}\,\mathbf{b}\,\mathbf{a}\,\mathrm{a}\,\mathrm{b}\,\mathrm{a}\cdots \\
W_{\langle 3,1\rangle} &= \mathbf{a}\,\mathbf{b}\,\mathbf{b}\,\mathrm{a}\,\mathrm{b}\,\mathrm{b}\cdots
\end{aligned}$$

$$\begin{aligned}
W_{\langle 2,3\rangle} &= \mathbf{b}\,\mathbf{a}\,\mathbf{a}\,\mathrm{b}\,\mathrm{a}\,\mathrm{a}\cdots \\
W_{\langle 3,2\rangle} &= \mathbf{b}\,\mathbf{a}\,\mathbf{b}\,\mathrm{b}\,\mathrm{a}\,\mathrm{b}\cdots \\
W_{\langle 4,3\rangle} &= \mathbf{b}\,\mathbf{b}\,\mathbf{a}\,\mathrm{b}\,\mathrm{b}\,\mathrm{a}\cdots \\
W_{\langle 4,1\rangle} &= \mathbf{b}\,\mathrm{b}\,\mathrm{b}\,\mathrm{b}\,\mathrm{b}\,\mathrm{b}\cdots
\end{aligned}$$

$n \log n$ lower bound for Σidp

De Bruijn set

A set of words $W = \{w_i\}_{i=1}^{t}$ over the alphabet $\mathcal{A}$ is a de Bruijn set of order k if $\sum_{i=1}^{t} |w_i| = 2^k$ and every word $v \in \mathcal{A}^k$ is a prefix of exactly one word in $\mathrm{suf}(W)$.

$$W = \{a,\ aab,\ abb,\ b\} \qquad \sum_{i=1}^{4} |w_i| = 2^3$$

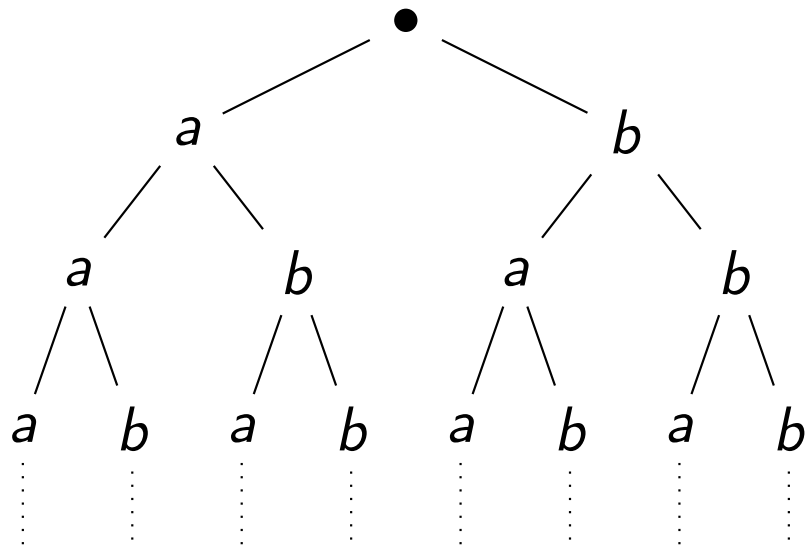
$$\begin{aligned}
W_{\langle 1,1\rangle} &= \mathbf{a}\,\mathrm{a}\,\mathrm{a}\,\mathrm{a}\,\mathrm{a}\,\mathrm{a}\cdots & W_{\langle 2,3\rangle} &= \mathbf{b}\,\mathbf{a}\,\mathbf{a}\,\mathrm{b}\,\mathrm{a}\,\mathrm{a}\cdots \\
W_{\langle 2,1\rangle} &= \mathbf{a}\,\mathbf{a}\,\mathbf{b}\,\mathrm{a}\,\mathrm{a}\,\mathrm{b}\cdots & W_{\langle 3,2\rangle} &= \mathbf{b}\,\mathbf{a}\,\mathbf{b}\,\mathrm{b}\,\mathrm{a}\,\mathrm{b}\cdots \\
W_{\langle 2,3\rangle} &= \mathbf{a}\,\mathbf{b}\,\mathbf{a}\,\mathrm{a}\,\mathrm{b}\,\mathrm{a}\cdots & W_{\langle 4,3\rangle} &= \mathbf{b}\,\mathbf{b}\,\mathbf{a}\,\mathrm{b}\,\mathrm{b}\,\mathrm{a}\cdots \\
W_{\langle 3,1\rangle} &= \mathbf{a}\,\mathbf{b}\,\mathbf{b}\,\mathrm{a}\,\mathrm{b}\,\mathrm{b}\cdots & W_{\langle 4,1\rangle} &= \mathbf{b}\,\mathrm{b}\,\mathrm{b}\,\mathrm{b}\,\mathrm{b}\,\mathrm{b}\cdots
\end{aligned}$$

Lemma (Higgins – 2012)

For $k \geqslant 1$, and any $u \in U_k = \{ab, ba\}^{2^{k-1}}$, $W = \mathrm{IBWT}(u)$ is a de Bruijn set of order k.

$n \log n$ lower bound for Σidp

$$W_k = \mathrm{IBWT}\left((ab)^{k-1}\right)$$



$$W_3 = \{a,\ aab,\ abb,\ b\}$$

Theorem

$$\Sigma\mathrm{ilcp}(W_k) = \sum_{i=1}^{k-1} i2^i = k \cdot 2^k - 2^{k+1} + 2$$

$n \log n$ lower bound for $\Sigma\mathrm{idp}$

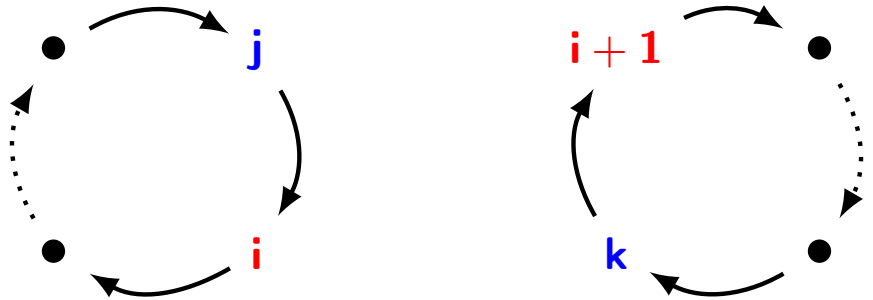
Theorem

For any $k \geqslant 1$, there exists a word w of length $n = 2^k$ such that $\Sigma\mathrm{idp}(w) = n \log n - O(n)$.

$n \log n$ lower bound for $\Sigma\mathrm{idp}$

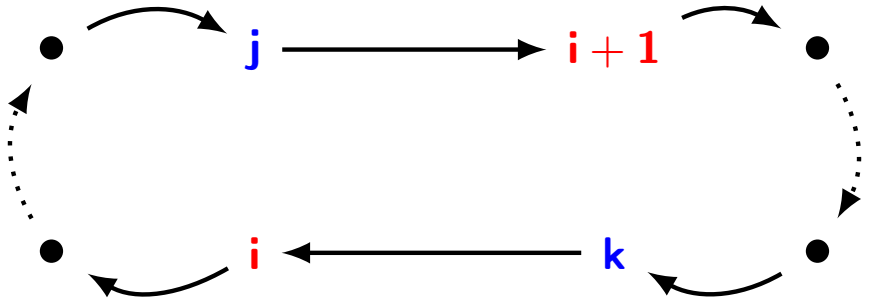
Theorem

For any $k \geqslant 1$, there exists a word w of length $n = 2^k$ such that $\Sigma\mathrm{idp}(w) = n \log n - O(n)$.



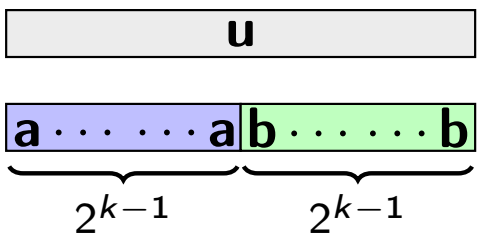
Theorem

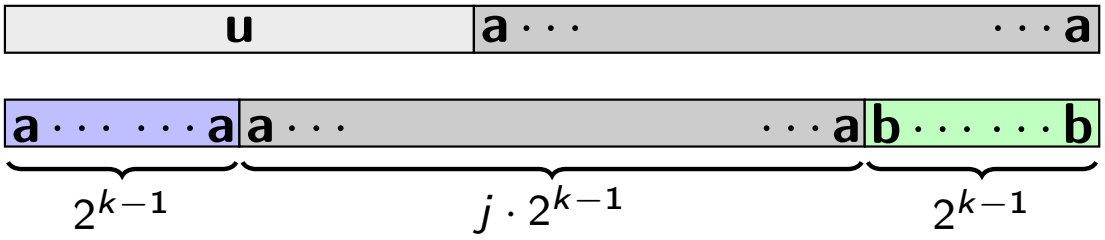
For any $k \geqslant 1$, there exists a word w of length $n = 2^k$ such that $\Sigma\mathrm{idp}(w) = n \log n - O(n)$.



$n \log r$ lower bound for Σidp

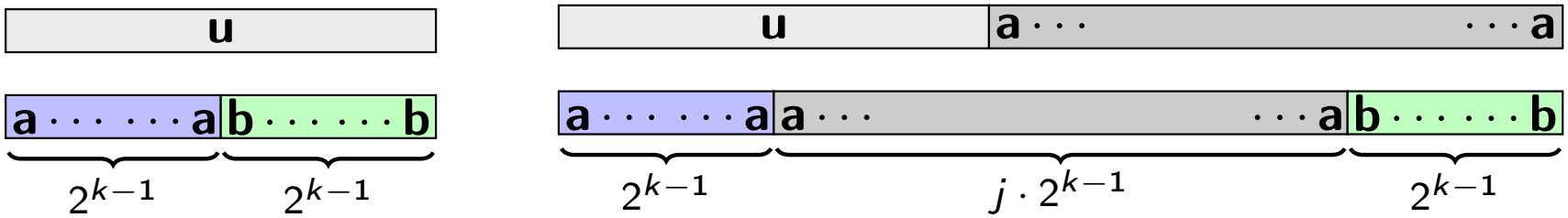
$$W_{k,j} = \mathrm{IBWT}\Big(\underbrace{(ab)^{2^{k-1}}}_{u} a^{j2^{k-1}} \Big)$$





$n \log r$ lower bound for Σidp

$$W_{k,j} = \mathrm{IBWT}\Big(\underbrace{(ab)^{2^{k-1}}}_{u} a^{j2^{k-1}} \Big)$$



$$W_{k,j} = \bigcup_{i=1}^{j+1} S_{i,k,j}$$

$$S_{i,k,j} = \begin{cases} a^i b a^j \{a, b a^j\}^{k-1} & : \ i \leqslant j \\ a^{j+1} \{a, b a^j\}^{k-1} & : \ i > j \end{cases}$$

Lemma

For $j \geqslant 1$ we have $\Sigma\mathrm{ilcp}(W_{k,j}) = (j+2)k2^{k-1} - 2^{k+1} + j + 1$.

$n \log r$ lower bound for $\Sigma\mathrm{idp}$

Theorem

For any $r = 2^k + 1$, $k \geqslant 1$, and $n \geqslant r$ such that $2^{k-1} | n$, there exists a word w of length n such that $\mathrm{BWT}(w)$ contains $r - o(r)$ runs and $\Sigma\mathrm{idp}(w) = n \log r - O(n)$.

Concluding remarks

- New upper bound for $\Sigma\mathrm{ilcp}$ and $\Sigma\mathrm{idp}$
- New upper bound for $\Sigma\mathrm{ilcp}$ and $\Sigma\mathrm{idp}$ related to the number of runs in BWT
- Lower bounds for $\Sigma\mathrm{ilcp}$ and $\Sigma\mathrm{idp}$ matching upper bounds
- Lower bounds for $\Sigma\mathrm{ilcp}$ and $\Sigma\mathrm{idp}$ – single word case

THANK YOU

KIITOS

DZIĘKUJĘ

GRAZIE