On Baier’s Sort of Maximal Lyndon Substrings

Frantisek Franek\textsuperscript{1}, Michael Liut\textsuperscript{1}, and W. F. Smyth\textsuperscript{1,2}

\textsuperscript{1} Department of Computing and Software
McMaster University, Hamilton, Canada
\{franek/liutm/smyth\}@mcmaster.ca

\textsuperscript{2} School of Engineering and Information Technology
Murdoch University, Perth, Australia

Abstract. We describe and analyze in terms of Lyndon words an elementary sort of maximal Lyndon factors of a string and prove formally its correctness. Since the sort is based on the first phase of Baier’s algorithm for sorting of the suffixes of a string, we refer to it as Baier’s sort.

Keywords: string, suffix, suffix array, Lyndon array, Lyndon string, maximal Lyndon substring

1 Introduction

The computation of the maximal Lyndon substrings of a string has been actively researched since Bannai et al. presented a linear-time algorithm for computing runs \cite{3}. Their algorithm relies on knowledge of all maximal Lyndon factors with respect to an order of the alphabet, and knowledge of all maximal Lyndon factors with respect to the inverse order. The maximal Lyndon factors of a string $x = x[1..n]$ are represented in the Lyndon array $L[1..n]$, where $L[i] =$ the length of the maximal Lyndon factor starting at position $i$: see \cite{7,11} and references therein. Other linear-time algorithms for computing all the runs rely on Lempel-Ziv factorization, and so the comparative efficiency of the runs algorithms is determined by the comparison of computing the Lyndon array and computing the Lempel-Ziv factorization: see \cite{6,4} and references therein.

There is only one known linear-time algorithm for computing the Lyndon array and it relies on suffix sorting: compute the suffix array, then the inverse suffix array, and then employ NSV (Next Smaller Value) algorithm to compute the Lyndon array from the inverse suffix array \cite{12}. All other algorithms have $O(n \log n)$ or $O(n^2)$ worst case complexity: see \cite{9,14} and references therein. The Lempel-Ziv factorization can be efficiently computed in linear time. Thus, our research focuses on linear computation of the Lyndon array without the need to build an unrelated global data structure such as the suffix array.

Baier \cite{12} in 2016 presented an elementary though elaborate algorithm for suffix sorting. The algorithm works in two phases. Soon afterwards, Cristoph Diegelmann in
reaction to [9] was first to note that phase I of the Baier’s suffix sorting algorithm in fact identifies and sorts the maximal Lyndon factors in linear time [10]. The performance of the algorithm was not great: Baier himself noted that his implementation of the suffix sort was about four times slower than the best linear algorithms for suffix sorting. Our analysis indicated that the first phase was the main culprit. We analyzed phase II in detail and presented a different implementation in [10].

There are three main goals for our study of Baier’s sort. The first is to describe and formalize Baier’s sort in terms of Lyndon words. The second is to provide a more formal and more detailed proof of the correctness of the sort in the framework of Lyndon words. And finally, the third goal is a detailed analysis of the method essential for a more efficient programming implementation in order to speed up the execution and lower the memory required. An additional goal is to see whether knowledge of the Lyndon array for a string can be utilized to speed up Baier’s sort of the maximal Lyndon factors. This paper concerns the first two goals.

Lyndon words admit the standard factorization and so often are built from these two components in what is in essence a bottom-up approach [3,16]. However, maximal Lyndon factors of a string can be built in a top-down fashion not related to the standard factorization. In Baier’s sort, the maximal Lyndon factors are determined using what we call the water draining method\(^1\) since the process can be best visualized as if the string represented a bunch of hills in a water tank and we slowly drain the water from the tank. The maximal Lyndon factors are then built from the hills that the draining reveals. The water draining method has three steps:

(a) lower the water level by one
(b) extend the existing Lyndon factors
    \(\text{the revealed letters are used to extend the existing Lyndon factors where possible,}
    \text{or became Lyndon factors of length 1 otherwise;}\)
(c) consolidate the new Lyndon factors
    \(\text{processed from the right, if several Lyndon factors are adjacent and can be joined}
    \text{to a longer Lyndon factor, they are joined.}\)

In Fig. 1, we illustrate the process:

(1) We start with the string \(\text{abcededbcdba}\) and a full tank of water.
(2) We drain one level, only \(e\) is revealed, nothing to extend, nothing to consolidate.
(3) We drain one more level and three \(d\)’s are revealed, the first \(d\) extends \(e\) to \(de\) and the remaining two \(d\)’s form Lyndon factors \(d\) of length 1, nothing to consolidate.
(4) We drain one more level and two \(c\)’s are revealed, the first extends \(de\) to \(cde\) and the second extends \(d\) to \(cd\), the consolidation then joins \(cde\) and \(d\) to \(cded\) (5).

\(^1\) This is not a pun on “graindraining” string used by Baier for illustration of his method.
(6) We drain one more level and three b’s are revealed, the first extends cded to bced, the second extends cd to ced, the third is a Lyndon factor b of length 1, nothing to consolidate.

(7) We drain one more level and two a’s are revealed, the first extends bced to abced and the second becomes a Lyndon factor a of length 1, in (8) abced and ced are joined to abcedbcd, and we continue the consolidation in (9) where abcedbcd and b are joined to abcedbcd, and the consolidation is complete.

So, during the process the following maximal Lyndon factors were identified: e at position 5, de at position 4, d at positions 6, 9, cded at position 3, cd at position 8, bced at position 2, ced at position 7, b at position 10, abcedbcd at position 1, and a at position 11. Note that all positions are accounted for, we really got all maximal Lyndon factors of the string abcedbcd. For a depiction of all maximal Lyndon factors of abcedbcd, see Fig. 2.

![Figure 1. The water draining method for abcedbcd](image)

The true ingenuity of Baier was to realize the water draining method by a simple linear mechanism based on the $\text{prev}()$ operator.

2 Background notation, notions and facts

In this section, we provide the description and definitions of the basic string notions and notation we are using in this paper, and the relevant basic facts. A *string* $x$ is a sequence $x[1..n]$ of *letters* $x[i], 1 \leq i \leq n$, each drawn from a set $\mathcal{A}$ called the *alphabet*. The *length* of the string is $n$. $\mathcal{A}$ is a totally ordered set.
The symbol ε denotes the empty string.
- A concatenation of two strings x and y is denoted by xy. A concatenation of k copies of u is denoted as uk, for an integer k ≥ 2.
- If a string x = uvw, then u is a prefix of x, v is a factor or substring of x, and w is a suffix of x. A prefix (resp. suffix) is trivial if it is empty, and is proper if it is not the whole x.
- The fact that u is a prefix of x is denoted by u ⊄ x, while the fact that u is a proper prefix of x is denoted by u ⊳ x.
- A string w is a border of a string x, if w is both a proper prefix and a proper suffix of x. An empty string is a primitive border. If x has only a primitive border, it is unbordered.
- A string x is primitive if there are no integer k ≥ 2 and string u so that x = uk.
- The expression x ≺ y denotes the fact that x is lexicographically smaller than y; that is, either x is a proper prefix of y or there is i ≤ min{|x|, |y|} such that x[1..i−1] = y[1..i−1] and x[i] ≺ y[i].
- The expression x ≼ y denotes the fact that x ≺ y or x = y.
- The expression x ≺ · y denotes the fact that x ≺ y, but x is not a prefix of y.
- For a string x = wu, uw is called a rotation of x, if either u or w is empty, then the rotation is trivial.
- A string x is Lyndon if either |x| = 1, the so-called trivial Lyndon string, or x is lexicographically strictly smaller than any of its non-trivial rotations. Such a string is often called a Lyndon word [5].
- A Lyndon factor u = x[i..j] of a string x = x[1..n] is maximal if either j = n or for any j < ℓ ≤ n, x[i..ℓ] is not Lyndon.
- For a string x, Ax denotes the alphabet of the string; that is, the set of letters occurring in x.
- For two factors x[i1..j1] and x[i2..j2] a string x = x[1..n] with i1 ≤ i2, we say that these two factors

- are disjoint, if j1 < i2

  \[
  \begin{array}{c}
  \text{i}_1 \\
  \text{j}_1 \\
  \end{array} \\
  \begin{array}{c}
  \text{i}_2 \\
  \text{j}_2 \\
  \end{array}
  \]

- overlap, if i2 ≤ j1

- intersect, if i2 ≤ j1 ≤ j2

  \[
  \begin{array}{c}
  \text{i}_1 \\
  \text{j}_1 \\
  \end{array} \\
  \begin{array}{c}
  \text{i}_2 \\
  \text{j}_2 \\
  \end{array}
  \]

- the first includes the second (or, the second is included in the first), if j2 ≤ j1

  \[
  \begin{array}{c}
  \text{i}_1 \\
  \text{j}_1 \\
  \end{array} \\
  \begin{array}{c}
  \text{i}_2 \\
  \text{j}_2 \\
  \end{array}
  \]

Note that in our terminology
\((x \text{ and } y \text{ overlap}) \iff ((x \text{ and } y \text{ intersect}) \text{ or (one includes the other)})\)
- A family of factors of a string x has the Monge property, if each two distinct factors from the family are either disjoint or one includes the other, or, equivalently, no two factors intersect.
Facts 1. Basic facts of Lyndon strings ([8][12][13][15])

(i) For $u$ of length $> 1$,
   - $u$ is Lyndon $\Rightarrow$ $u$ is unbordered $\Rightarrow$ $u$ is primitive
   - $u$ is Lyndon $\iff u < u_2$ for any $u = u_1u_2$
   - $u$ is Lyndon $\iff u_1 < u_2$ for any $u = u_1u_2$
   - $u$ is Lyndon $\Rightarrow$ there are Lyndon words $u_1, u_2$ so that $u_1 < u_2$ and $u = u_1u_2$. If $u_2$ is the largest possible such suffix, it is called the standard factorization of $u$.

(ii) Let $u = wcv$ be Lyndon and let $c < d$. Then $uwd = wcwvd$ is Lyndon.

(iii) Let $x = vu$ for non-empty $v$ and $u$ and let $v$ be a Lyndon factor of $x$. Then $v$ is a maximal Lyndon factor of $x$ $\iff u < vu$.

(iv) Let $u = u_1vu_2$ be Lyndon and let $v$ be a maximal Lyndon factor of $u$. Then $u < v$.

3 Basic Definitions and Notions

In this section we present the basic definitions and notions for the description and analysis of Baier’s sort.

Definition 2. Let $x = x[1..n]$ be a string.

- A **group** with a **context** $u$, denoted as $G_u$, is an ascending sequence of indices $i_1 < \cdots < i_k$ such that $u$ is a prefix of $x[i_\ell..n]$ for every $\ell \in 1..k$.
- A sequence $\mathcal{C} = [G_{u(m)}, G_{u(m-1)}, \ldots, G_{u(1)}]$ is a **group configuration** for $x$ if
  - (i) for any $\ell \in 1..m$, $G_{u(\ell)}$ is a group with the context $u(\ell)$, and
  - (ii) $u(m) < u(m-1) < \cdots < u(1)$, and
  - (iii) for any $\ell \in 1..m$, $u(\ell)$ is a Lyndon substring of $x$, and
  - (iv) all the groups are pairwise mutually disjoint, and
  - (v) $\bigcup_{i=1}^{m} G_{u(\ell)} = 1..n$, and
  - (vi) the family $\{ \mathcal{C}(i) \mid 1 \leq i \leq n \}$ has the Monge property, where $\mathcal{C}(i) = x[i..i+|u(\ell)|-1]$ for the unique $\ell$ such that $i \in G_{u(\ell)}$, note that $\mathcal{C}(i) = u(\ell)$.

- For $i, j \in G_{u(\ell)}$, $i \otimes j$ iff $|i - j| = |u(\ell)|$. The relation $\sim$ is defined as a transitive closure of $\otimes$. Thus, $i \sim j$ is an equivalence relation on $G_{u(\ell)}$. The symbol $[i]_{\sim}$ denotes the class of equivalence $\sim$ to which $i$ belongs.
- For $i \in G_{u(\ell)}$, the **valence** of $i$ is defined as $val_{\mathcal{C}}(i) = |[i]_{\sim}|$.
- $gr_{\mathcal{C}}(i)$ denotes the unique group of $\mathcal{C}$ the index $i$ belongs to, i.e. $gr_{\mathcal{C}}(i) = G_{u(\ell)}$ iff $i \in G_{u(\ell)}$.
- A group $G_{u(\ell)}$ is **complete** if
  - for any $i \in G_{u(\ell)}$, $\mathcal{C}(i)$ is a maximal Lyndon factor of $x$, and
  - if $u(\ell) = x[j..j+|u|−1]$ is a maximal Lyndon factor of $x$, then $j \in G_{u(\ell)}$. 

Proceedings of the Prague Stringology Conference 2018

The operator \( \text{prev}_r(i) = \max\{j < i \mid \text{con}_r(gr_r(j)) < \text{con}_r(gr_r(i))\} \), if such a \( j \) exists, \( \text{nil} \) otherwise, where \( \text{con}_r(G) \) denotes the context of group \( G \) with respect to the configuration \( r \).

Note that it is possible for an \( i \in G_u \), that not only \( x[i..i+|u|-1] = u \), but also \( x[i+|u|..i+2|u|-1] = u \) and so forth. This is captured by the notion of the valence: if \( i \in G_u \) and \( \text{val}(i) = k \), the occurrence of \( u \) at position \( i \) is a part of a maximal repetition \( u^k \).

It is easy to see, that for \( j \in [i]_r \), \( \text{prev}_r(i) = \text{prev}_r(j) \). Let \( i, j \in G_u \). WLOG assume that \( j = i+p|u| \) and so \( x[i..j+|u|-1] = u^{p+1} \). Clearly, \( \text{prev}_r(i) \leq \text{prev}_r(j) \). If \( \text{prev}_r(j) > \text{prev}_r(i) \), it would indicate that a suffix of \( u \) must be lexicographically smaller than \( u \), which contradicts the Lyndon property of \( u \).

Together, (iv) and (v) assure that the groups of a group configuration form a disjoint partitioning of the string’s index range \( 1..n \).

Since Definition 2 is quite involved, we present an illustrative example on a string \( x = x[1..11] = abcedbcedba \).

\[
\begin{array}{cccccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\
a & b & c & d & e & d & b & c & d & b & a \\
\end{array}
\]

\[
G_a = \{11\}, G_{abcedbcedba} = \{1\}, G_b = \{10\}, G_{bcde} = \{7\}, G_{bcded} = \{2\}, \\
G_{cd} = \{8\}, G_{cded} = \{3\}, G_d = \{6, 9\}, G_{de} = \{4\}, G_e = \{5\}.
\]

Take for instance the group \( G_d = \{6, 9\} \), the context of the group is a string of length 1, \( d \), and indeed, at the positions 6 and 9 we have substrings \( d \) starting. Moreover, \( G_d \) is complete, as \( x[6] \) and \( x[9] \) are both maximal Lyndon, and there is no other occurrence of a maximal Lyndon substring \( d \). Similarly for \( G_{abcedbcedba} = \{1\} \), \( x[1..10] = abcedbcedba \), so it is the context, \( abcedbcedba \) is Lyndon, and \( x[1..10] = abcedbcedba \) is maximal, and there is no other occurrence of maximal \( abcedbcedba \), so \( G_{abcedbcedba} \) is complete. It is easy to see, that the whole set of the groups in the order given above is a group configuration: \( a < abcedbcedba < b < bcd < bced < cde < d < de < e \) and \( G_a \cup G_{abcedbcedba} \cup G_b \cup G_{bcde} \cup G_{bcded} \cup G_{cded} \cup G_d \cup G_{de} \cup G_e = 1..11 \).

Denote \( r = [G_a, G_{abcedbcedba}, G_b, G_{bcde}, G_{bcded}, G_{cded}, G_d, G_{de}, G_e] \). Consider the system of factors \{ \( \mathcal{C}(i) \mid 1 \leq i \leq n \} \), let us demonstrate that it indeed has the Monge property:

\[
\begin{align*}
\mathcal{C}(1) &= x[1..10] = abcedbcedba, \\
\mathcal{C}(2) &= x[2..6] = bcded, \\
\mathcal{C}(3) &= x[3..6] = cded, \\
\mathcal{C}(4) &= x[4..5] = de, \\
\mathcal{C}(5) &= x[5] = e, \\
\mathcal{C}(6) &= x[6] = d, \\
\mathcal{C}(7) &= x[7..9] = bcd,
\end{align*}
\]
Lemma 3. Let $G_u$ be a group with a primitive context $u$ for a string $x$. Let $i \in G_u$. Then $[i]_\sim = \{ \ell \in G_u \mid \min [i]_\sim \leq \ell \leq \max [i]_\sim \}$.

Proof. It suffices to prove that if $j - i = |u|$, then there is no $\ell \in G_u$ so that $i < \ell < j$. Assume there is such an $\ell$. Then $x[i \ldots i + |u| - 1] = x[j \ldots j + |u| - 1] = u$ and since $j = i + |u|$, we have $x[i \ldots i + 2|u| - 1] = 2u$. Moreover, $x[\ell \ldots \ell + |u| - 1] = u$ and $i < \ell < 2i$. Since $u$ is primitive, this contradicts the synchronization principle for primitive strings.

To illustrate Lemma 3 consider $abbbabb$: $G_b = \{2, 3, 4, 5, 7, 8\}$. Consider $[3]_\sim = \{2, 3, 4, 5\}$, you can see that $[3]_\sim = \{i \in G_b \mid 2 \leq i \leq 5\}$. Similarly, $[7]_\sim = \{7, 8\} = \{i \in G_b \mid 7 \leq i \leq 8\}$.

Definition 4. A group configuration $\mathcal{C} = [G_u(m), G_u(m-1), \ldots, G_u(1)]$ for a string $x$ is $r$-proper for $1 \leq r \leq m$ if

\begin{enumerate}
  \item[(vii)] all the groups $G_u(r), \ldots, G_u(1)$ are complete; and
  \item[(viii)] all the groups $G_u(r-1), \ldots, G_u(1)$ have been processed; and
  \item[(ix)] for any $i \in G_u(r)$, if $j = \text{prev}_\mathcal{C}(i)$, then $\mathcal{C}(j)(\mathcal{C}(i))^t$ where $t = \text{val}_\mathcal{C}(i)$, is a prefix of $x[j \ldots n]$.
\end{enumerate}

To process the group $G_u(r)$ entails:

\begin{enumerate}
  \item[(a)] Compute $P = \{\text{prev}_\mathcal{C}(i) \mid i \in G_u(r) \text{ and } \text{prev}_\mathcal{C}(i) \neq \text{nil}\}$;
  \item[(b)] Let $\approx$ be an equivalence on $P$ defined by $i_1 \approx i_2$ iff $\text{gr}_{\mathcal{C}}(i_1) = \text{gr}_{\mathcal{C}}(i_2)$. Compute the disjoint partitioning of $P = P_1 \cup P_2 \cup \cdots \cup P_k$ into the classes of equivalence $\approx$. Let $P_1 \subseteq G_u(n_1), \ldots, P_k \subseteq G_u(n_k)$;
  \item[(c)] For each $j \in 1..k$, let $\approx_j$ be an equivalence on $P_j$ defined by $i_1 \approx_j i_2$ iff $i_1 = \text{prev}_\mathcal{C}(\ell_1) \& i_2 = \text{prev}_\mathcal{C}(\ell_2)$ & $\text{val}(\ell_1) = \text{val}(\ell_2)$. Let $[i_1]_{\approx_j}$ denote the class of equivalence of $\approx_j$ containing $i_1$. Define $\text{val}_\mathcal{C}(i_1)_{\approx_j} = \text{val}(\ell_1)$ so that $i_1 = \text{prev}_\mathcal{C}(\ell_1)$. Compute the disjoint partitioning $P_j = P_{j,1} \cup P_{j,2} \cup \cdots \cup P_{j,t_j}$ into the classes of equivalence $\approx_j$. Moreover, let $\text{val}(P_{j,1}) > \text{val}(P_{j,2}) > \cdots > \text{val}(P_{j,t_j})$.
  \item[(d)] for each $j \in 1..k$, for each $\ell \in 1..t_j$, move all indices of $P_j,\ell$ from $G_u(n_j)$ to a new group $G'$. The group $G'$ is placed directly after $G_u(n_j)$ which is removed if it becomes empty. The context of $G'$ is set to $u_{n_j}(u_{\ell})^{\text{val}(P_{j,\ell})}$.
Remark: Definition 4 may not seem sound, for the definition of process is only provided after it had been used in (viii). However, the definition is in fact recursive; that is for a 1-proper configuration, no group needs to be processed, then we process the last group to obtain a 2-proper configuration, and so on.

We will use a simple string $aabbabb$ to illustrate the notions of the equivalence $\sim$, and the valence:

\begin{align*}
1 & 2 & 3 & 4 & 5 & 6 & 7 \\
a & a & b & b & a & b & b
\end{align*}

First consider the configuration $\mathcal{C}_1 = [G_a, G_b]$ where $G_a = \{1, 2, 5\}$ and $G_b = \{3, 4, 6, 7\}$. Then $\text{val}_{\mathcal{C}_1}(1) = \text{val}_{\mathcal{C}_1}(2) = \text{val}_{\mathcal{C}_1}(3) = \text{val}_{\mathcal{C}_1}(4) = \text{val}_{\mathcal{C}_1}(6) = \text{val}_{\mathcal{C}_1}(7) = 2$ and $\text{val}_{\mathcal{C}_1}(5) = 1$. It means, that for the processing of $G_b$ we do not consider 4 and 7, only 3 and 6 and their valences of 2. Thus $P = \{2, 5\}$ as $2 = \text{prev}_{\mathcal{C}_1}(3) = \text{prev}_{\mathcal{C}_1}(4)$ and $5 = \text{prev}_{\mathcal{C}_1}(6) = \text{prev}_{\mathcal{C}_1}(7)$. Since $\text{gr}_{\mathcal{C}_1}(2) = \text{gr}_{\mathcal{C}_1}(5) = G_a$, 2 and 5 are both in the same $P_j$. So, after the refinement of $G_a$ we get a new configuration $\mathcal{C}_2 = [G_a, G_{abb}, G_b]$ where $G_a = \{1\}$, $G_{abb} = \{2, 5\}$, and $G_b = \{3, 4, 6, 7\}$. Then $\text{val}_{\mathcal{C}_2}(2) = \text{val}_{\mathcal{C}_2}(5) = 2$, so $P = \{1\}$, as $1 = \text{prev}_{\mathcal{C}_2}(2) = \text{prev}_{\mathcal{C}_2}(5)$, and so we get a new and final configuration $\mathcal{C}_3 = [G_{aabbabb}, G_{abb}, G_b]$ where $G_{aabbabb} = \{1\}$, $G_{abb} = \{2, 5\}$, and $G_b = \{3, 4, 6, 7\}$.

For illustration, let us perform complete Baier’s sort of $abcdedbcdba$, the arrows represent the $\text{prev}$ operator. We start with the initial group configuration where each group contains all indices that start with the same letter. Processing from right, take the first unprocessed group from right, compute the $\text{prev}$ values for indices in that group, and perform context concatenation wherever it points, partitioning the groups in the process. Then move to the next unprocessed group, until all groups except the very first one are processed.

\begin{align*}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\
a & b & c & d & e & d & b & c & d & b & a
\end{align*}

\begin{align*}
G_a &= \{1, 11\} & G_b &= \{2, 7, 10\} & G_c &= \{3, 8\} & G_d &= \{4, 6, 9\} & G_e &= \{5\} \\
G_a &= \{1, 11\} & G_b &= \{2, 7, 10\} & G_c &= \{3, 8\} & G_d &= \{6, 9\} & G_{de} &= \{4\} \cdots \\
G_a &= \{1, 11\} & G_b &= \{2, 7, 10\} & G_c &= \{8\} & G_{cde} &= \{3\} & G_d &= \{6, 9\} \cdots
\end{align*}
The processed (in bold) classes represent not only all maximal Lyndon factors of the string, but they are also lexicographically ordered:

\[ G_a = \{1, 11\}, G_{bcdde} = \{1\}, G_b = \{10\}, G_{bcd} = \{7\} \]
\[ G_{cd} = \{8\}, G_{cdde} = \{3\} \] · · ·

Compare it to Fig. 2 to see that all maximal Lyndon factors are accounted for.

\[ \begin{array}{cccccccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\
\hline
a & b & c & d & e & d & b & c & d & b & a \\
\end{array} \]

\textbf{Figure 2.} Maximal Lyndon Factors of \textit{abcdedbeda}

4 Properties of the group refinement

In this section we deal with the basic arrangement of group configurations referred in the text as (B) given below:

Given a string \( x[1..n] \) with the alphabet \( \{a_1, \ldots, a_k\} \). For \( \hat{r} \geq 1 \),

\[ \mathcal{C}_1 = [G_{u(1, m_1)}, G_{u(1, m_1-1)}, \ldots, G_{u(1, 1)}] \]
\[ \ldots \]
\[ \mathcal{C}_{\hat{r}} = [G_{u(\hat{r}, m_{\hat{r}})}, G_{u(\hat{r}, m_{\hat{r}}-1)}, \ldots, G_{u(\hat{r}, 1)}] \]

where \( \mathcal{C}_1 = [G_{u(1, m_1)}, G_{u(1, m_1-1)}, \ldots, G_{u(1, 1)}] = [G_{a_1}, \ldots, G_{a_k}] \);
for each \( 1 \leq r < \hat{r} \), \( \mathcal{C}_r \) is an \( r \)-proper group configuration; and the
configuration $\mathcal{G}_{r+1} = [G_{u(r+1,m_{r+1}+1)}, G_{u(r+1,m_{r+1}+1-1)}, \ldots, G_{u(r+1,1)}]$ is produced by processing of the group $G_{u(r,r)}$.

First, a few fundamental observations of the nature of “refinement” of the groups during processing.

**Observation 5.** Referring to (B), for any $2 \leq r \leq \hat{r}$ and any $1 \leq i, j \leq n$,

(i) If $\mathcal{C}_r(i)$ is a proper suffix of $\mathcal{C}_r(j)$, then $\mathcal{C}_r(i) = u(\xi, \rho)$ and $\mathcal{C}_r(j) = \mathcal{C}_\xi(j)(u(\rho, \xi))^\gamma$ where $\xi < r$ and $t = \text{val}_{\mathcal{C}_\xi}(i)$.

(ii) Either $\mathcal{C}_r(i) = \mathcal{C}_{r-1}(\xi)$ or $\mathcal{C}_r(i) = \mathcal{C}_{r-1}(\xi)(u(\rho, \xi, \rho))^\gamma$ where $t = \text{val}_{\mathcal{C}_{r-1}}(i)$; consequently, $\mathcal{C}_{r-1}(\xi)$ is a prefix of $\mathcal{C}_r(i)$, while $\mathcal{C}_r(i)$ cannot be a prefix of $\mathcal{C}_{r-1}(\xi)$.

(iii) If $|\mathcal{C}_{r+1}(i)| > 1$, then there exist $\xi \leq r$ and $1 \leq \rho \leq m_\xi$, so that $\mathcal{C}_{r+1}(i) = u(\xi, \rho)(u(\xi, \rho))^\gamma$ where $t = \text{val}_{\mathcal{C}_\xi}(i + |u(\xi, \rho)|)$.

Lemma 6 shows how the Monge property of the system of all occurrences of all the group contexts propagates through a group configuration arrangement.

**Lemma 6.** Referring to (B), then for any $1 \leq r \leq \hat{r}$, the system $\{ \mathcal{C}_r(i) \mid 1 \leq i \leq n \}$ has the Monge property.

**Proof.** We are going to prove it by induction over $r$. First consider the case when $r = 1$.

Then each $u(1, \ell) = c$ for some letter $c$ of $x$. Thus each $\mathcal{C}_1(i) = c$ for some letter $c$ of $x$, and hence either $\mathcal{C}_1(i) = \mathcal{C}_1(j)$ or $\mathcal{C}_1(i) \cap \mathcal{C}_1(j) = \emptyset$ for any $i \neq j$.

Induction hypothesis: $\{ \mathcal{C}_{r-1}(i) \mid 1 \leq i \leq n \}$ has the Monge property.

Consider $\mathcal{C}_r(i)$ and $\mathcal{C}_r(j)$ for $i \neq j$. WLOG assume $i < j$.

- Case $\mathcal{C}_r(i) = \mathcal{C}_{r-1}(i)$ and $\mathcal{C}_r(j) = \mathcal{C}_{r-1}(j)$.
  
  By the induction hypothesis, either $\mathcal{C}_{r-1}(i) \cap \mathcal{C}_{r-1}(j) = \emptyset$ or $\mathcal{C}_{r-1}(i)$ includes $\mathcal{C}_{r-1}(j)$.

- Case $\mathcal{C}_r(i) = \mathcal{C}_{r-1}(i)(\mathcal{C}_{r-1}(\ell))^\rho$, $\text{prev}_{\mathcal{C}_{r-1}}(\ell) = i$, $\rho = \text{val}_{\mathcal{C}_{r-1}}(\ell)$, $\mathcal{C}_{r-1}(\ell) = u(r-1, r-1)$, and $\mathcal{C}_r(j) = \mathcal{C}_{r-1}(j)$. Since for all involved components, no two can intersect, the possible subcases are:
    - $\mathcal{C}_{r-1}(j)$ is disjoint from $\mathcal{C}_{r-1}(i)(\mathcal{C}_{r-1}(\ell))^\rho$.
    - $\mathcal{C}_{r-1}(j)$ is included in a copy of $\mathcal{C}_{r-1}(\ell)$.
    - $\mathcal{C}_{r-1}(j)$ is included in $\mathcal{C}_{r-1}(i)$.

- Case $\mathcal{C}_r(i) = \mathcal{C}_{r-1}(i)$ and $\mathcal{C}_r(j) = \mathcal{C}_{r-1}(j)(\mathcal{C}_{r-1}(\ell))^\rho$, $\text{prev}_{\mathcal{C}_{r-1}}(\ell) = j$, $\rho = \text{val}_{\mathcal{C}_{r-1}}(\ell)$, and $\mathcal{C}_{r-1}(\ell) = u(r-1, r-1)$. Since for all involved components, no two can intersect, the possible subcases are:
    - $\mathcal{C}_{r-1}(i)$ and $\mathcal{C}_{r-1}(j)$ are disjoint.
    - $\mathcal{C}_{r-1}(i)$ includes $\mathcal{C}_{r-1}(j)$ as a proper suffix. By Obs. 5(i), it means that $\mathcal{C}_{r-1}(j) = u(r-2, r-2)$. Since $\text{prev}_{\mathcal{C}_{r-1}}(\ell) = j$, it means $u(r-2, r-2) < u(r-1, r-1)$, which is a contradiction as $u(r-1, r-1) < u(r-1, r-2) = u(r-2, r-2)$.
If arguing by contradiction assume to have it for some $C$ configurations arrangement. Lemma 8 shows that.

- $C_{r-1}(i)$ includes $C_{r-1}(j)(C_{r-1}(\ell))^\xi$ for some $\xi \leq \rho$ as a proper suffix. Then $C_{r-1}(\ell)$ is a suffix of $C_{r-1}(i)$ and by (B), $C_{r-1}(\ell) = u_{r-2, r-2}$, i.e. $u_{r-1, r-1} = u_{r-2, r-2}$, a contradiction.
- $C_{r-1}(i)$ includes $C_{r-1}(j)(C_{r-1}(\ell))^\rho$ but not as a suffix.
- Case $C_r(i) = C_{r-1}(i)(C_{r-1}(\ell))^\rho$, $\text{prev}_{C_{r-1}}(\ell) = i$, $\rho = \text{val}_{C_{r-1}}(\ell)$, $C_{r-1}(\ell) = u_{r-1, r-1}$, and $C_r(j) = C_{r-1}(j)(C_{r-1}(k))^\xi$, $\text{prev}_{C_{r-1}}(k) = j$, $\xi = \text{val}_{C_{r-1}}(k)$, $C_{r-1}(k) = u_{r-1, r-1}$. Since for all involved components, no two can intersect, the possible subcases are:
  - $C_{r-1}(i)(u_{r-1, r-1})^\rho$ is disjoint from $C_{r-1}(j)(u_{r-1, r-1})^\xi$.
  - $C_{r-1}(i)$ includes $C_{r-1}(j)(u_{r-1, r-1})^\xi$ but not as a suffix, which is fine.
  - $C_{r-1}(i)$ includes $C_{r-1}(j)(u_{r-1, r-1})^\xi$ as a suffix, which is a contradiction as the $\xi$ copies of $u_{r-1, r-1}$ are immediately followed by another $\rho$ copies, so $C_r(i)$ should equal to $C_{r-1}(j)(C_{r-1}(k))^\xi+\rho$.
  - $C_{r-1}(i)$ includes $C_{r-1}(j)(u_{r-1, r-1})^\xi$, $\tau < \xi$ as a proper suffix. But then $\xi - \tau = \rho$ and $\ell = k + \tau|u_{r-1, r-1}|$. Since $\text{prev}_{C_{r-1}}(\ell) = i$ since $C_r(i) = C_{r-1}(i)(u_{r-1, r-1})^\rho$ and $\text{prev}_{C_{r-1}}(\ell) = j$ since $C_r(j) = C_{r-1}(j)(u_{r-1, r-1})^\xi$, a contradiction.
  - $C_{r-1}(j)$ is a suffix of $C_{r-1}(i)$ and $\rho = \xi$. Then $\ell = k$ and $\text{prev}_{C_{r-1}}(\ell) = i$ since $C_r(i) = C_{r-1}(j)(u_{r-1, r-1})^\rho$ and $\text{prev}_{C_{r-1}}(\ell) = j$ since $C_r(j) = C_{r-1}(j)(u_{r-1, r-1})^\xi$, a contradiction.

The process of refinement of the groups has a very particular property, namely $u_{r,k}$ is never followed immediately by $u_{r-\xi, r-\xi}$ for any $r-\xi \geq 1$ if $u_{r,k} < u_{r-\xi, r-\xi}$.

**Lemma 7.** Referring to (B), let $1 \leq r-\xi < r < \hat{r}$, let $u_{r,k} < u_{r-\xi, r-\xi}$, and let $j = i + |u_{r,k}|$. Then it is impossible to have $C_r(i) = u_{r,k}$ and $C_r(j) = u_{r-\xi, r-\xi}$.

**Proof.** Arguing by contradiction assume to have it for some $i$, $j$, $\xi$, $r$, and $k$. Then $C_{r-\xi}(i) \preceq C_r(i) = u_{r,k} < u_{r, r-\xi} = u_{r-\xi, r-\xi}$, and so $\text{prev}_{C_{r-\xi}}(j) \geq i$. If $\text{prev}_{C_{r-\xi}}(j) = i$, then $C_r(i) \prec C_{r-\xi+1}(i) \preceq C_r(i)$, a contradiction. Thus, $j_1 = \text{prev}_{C_{r-\xi}}(j) > i$. So, $C_{r-\xi+1}(j_1)$ and $C_r(i)$ intersect, and so $C_r(j_1)$ and $C_r(i)$ intersect as $C_{r-\xi+1}(j_1) \preceq C_r(j_1)$. But that contradicts Lemma 8.

We need to ascertain that the definition of the processing of $Gu_{r,r}$ can be carried out as defined in Def. 4, i.e. that the property of prev propagates through a group configurations arrangement. Lemma 8 shows that.
Lemma 8. Referring to (B), for any $1 \leq r \leq \hat{r}$, for any $i \in G_{(r, r)}$, if $j = \text{prev}_r(i)$, then $x[j..n]$ has $C_r(j)\langle u(r, r)\rangle^t$ where $t = \text{val}_r(i)$, as a prefix.

Proof. It is clear that for $r = 1$ it is true: Let $c$ be the largest letter in $x$, then $G_{u(1, 1)} = G_c$. Let $i \in G_c$ and let $j = \text{prev}_r(i)$. Let $i_1 = \min \{\ell \leq i \mid x[\ell] = c\}$, then $\text{val}_r(i) = \text{val}_r(i_1) = t$ and $\text{prev}_r(i_1) = \text{prev}_r(i) = j$. Then $j = i_1 - 1$, $C_1(j) = b$ for some $b < c$, and $C_1(i_1) = c'$.

We are assuming it holds true for $r$.

We shall prove it for $r+1$, but first we need to prove three simple claims.

Though a bit stronger, in essence, the first claim states that the configuration illustrated below cannot happen.

$$
\begin{array}{c|c|c}
\hat{i} & j & u(r, k) = C_r(i) \\
\hline
u(r, r) = C_r(j) & &
\end{array}
$$

Claim 1: If $1 \leq r \leq \hat{r}$, $i, j \in 1..n$, $i < j$, and $\xi \geq 0$, then $C_r(i)$ cannot include $C_{r+\xi}(j) = u(r+\xi, r+\xi)$.

Arguing by contradiction, take the minimal $r$ such that for some $i < j$, and some $\xi$, $C_r(i)$ includes $C_{r+\xi}(j) = u(r+\xi, r+\xi)$. Let $C_r(i) = u(r, k)$ for some $k$. Since $C_r(i)$ includes $C_{r+\xi}(j)$ and $C_r(i) \neq C_{r+\xi}(j)$ as $i < j$, $|C_r(i)| \geq 2$ and so $r > 1$. By Obs. (iii), there are $r', k'$, and $t \geq 1$ so that $u(r, k) = C_r(i) = u(r', k')(u(r', r'))^t$. For $0 \leq h < t$, define $\ell_h = i + |u(r', k')| + h|u(r', r')|$. Then $C_r(i) = u(r', k')$ and each $C_r(\ell_h) = u(r', r')$. For any $0 \leq h < t$, by Lemma \[ \text{u}_{r+\xi, r+\xi} = C_{r+\xi}(j) \] must be either disjoint from $u(r', r') = u(r+\xi, r') = C_{r+\xi}(\ell_h)$, or one must include the other.

Let $\xi' = r-r'+\xi$, then $\xi' \geq 0$ and $r+\xi = r'+\xi'$.

- If $C_{r+\xi}(j)$ is disjoint from every $C_{r+\xi}(\ell_h)$, then $u(r'+\xi', r'+\xi') = u(r+\xi', r+\xi)$ must be included in $u(r', k') = C_r(i)$. So $u(r'+\xi', r'+\xi')$ is included in $u(r', k') = C_r(i)$, which contradicts the minimality of $r$.

Thus, $C_{r+\xi}(j)$ must be included in or include $C_{r+\xi}(\ell_h)$ for some $h \in 0..t-1$. If $C_{r+\xi}(\ell_h) = u(r', r')$ included $C_{r+\xi}(j) = u(r'+\xi', r'+\xi')$, we would have a contradiction with the minimality of $r$. Thus $C_{r+\xi}(j) = u(r'+\xi', r'+\xi')$ must include $C_{r+\xi}(\ell_h) = u(r', r')$, and so $j = \ell_h$, $u(r'+\xi', r'+\xi') = u(r', r')$, and so $r' = r'+\xi'$, a contradiction.

This concludes the proof of Claim 1.

Claim 2: Let $j = \text{prev}_r(i)$, $C_r(i) = u(r, r)$, and for any $j < i'$ < $i$, $C_r(i') \neq u(r, r)$. Then $C_r(j)$ and $C_r(i)$ are adjacent.

By Claim 1, $C_r(j)$ cannot include $C_r(i)$, and so, by Lemma \[ C_r(j)$ and $C_r(i)$ are disjoint. By contradiction, we shall see that they must be adjacent. So assume that they are not adjacent,
F. Franek, M. Liut, W. F. Smyth: On Baier’s Sort of Maximal Lyndon Substrings

\[ u_{(r, k)} = \mathcal{C}_r(j) \]

\[ u_{(r, r)} = \mathcal{C}_r(i) \]

\[ j = \text{prev}_{\mathcal{C}_r}(i) \]

i.e. \( \mathcal{C}_r(l) \vartriangleright u_{(r, r)} \), therefore \( \mathcal{C}_r(l) = u_{(r, r)} = u(\rho, \rho) \) for some \( \rho \leq r \). But since there is no occurrence of \( u_{(r, r)} \) between \( j \) and \( i, \rho < r \). But this is not possible by Lemma[7]

This concludes the proof of Claim 2.

**Claim 3:** Let \( j = \text{prev}_{\mathcal{C}_r}(i_1) = \text{prev}_{\mathcal{C}_r}(i_2) \) and let \( \mathcal{C}_r(i_1) = \mathcal{C}_r(i_2) = u_{(r, r)} \), and let \( i_2 > i_1 + |u_{(r, r)}| \). Then \( \mathcal{C}_r(i_1 + |u_{(r, r)}|) = u_{(r, r)} \).

Let \( i_3 = i_1 + |u_{(r, r)}| \). Since \( j = \text{prev}_{\mathcal{C}_r}(i_2) \), it follows that \( i_3 \in \bigcup_{\xi = 1}^{r} G_{u(r, \xi)} \). If \( i_3 \in \bigcup_{\xi = 1}^{r-1} G_{u(r, \xi)} \), then \( \mathcal{C}_r(i_3) = u(r, \xi) = u(\xi, \xi) \) for some \( \xi < r \). Since \( \mathcal{C}_r(i_1) \in \bigcup_{\xi = 1}^{r-1} G_{u(r, \xi)} \), it follows that \( i_4 = \text{prev}_{\mathcal{C}_r}(i_3) \geq i_1 \). If \( \text{prev}_{\mathcal{C}_r}(i_3) = i_1 \), then for some \( t \geq 1 \), \( \mathcal{C}_{r+t}(i_1) = u(r, r)(u(\xi, \xi))^t \leq \mathcal{C}_r(i_1) = u_{(r, r)} \), a contradiction. Thus \( i_4 > i_1 \) and so \( \mathcal{C}_{r+t}(i_4) = x[i_4..i_3-1](u(\xi, \xi))^t \leq \mathcal{C}_r(i_4) \) and so \( \mathcal{C}_r(i_1) \) and \( \mathcal{C}_r(i_4) \) intersect, a contradiction with Lemma[3]. Thus, \( i_3 \in G_{u(r, r)} \).

Now we can prove the induction step. Let \( j = \text{prev}_{\mathcal{C}_r}(i) \), let \( p = |u_{(r, r)}| \), and let \( \mathcal{C}_r(i) = u_{(r, r)} \). Let \( \ell \) be the smallest \( i \) such that \( j = \text{prev}_{\mathcal{C}_r}(i) \). Then by Claim 3, \( \mathcal{C}_r(j) = u_{(r, r)} \) and \( \mathcal{C}_r(i) = u_{(r, r)} \). Then by Claim 3, \( x[i_1..i_1+p-1] = (u_{(r, r)})^t \) where \( t = \text{val}_{\mathcal{C}_r}(\ell) \), moreover \( x[\ell-p..\ell-1] \neq u_{(r, r)} \) and \( x[i+tp..i+(t+1)p-1] \neq u_{(r, r)} \). By Claim 2, \( \mathcal{C}_r(j) \) and \( \mathcal{C}_r(\ell) \) are adjacent, and so \( x[j..n] \) has \( \mathcal{C}_r(j)(u_{(r, r)})^t \) as a prefix.

Note that for \( r = 1 \), \( G_{u(1, 1)} = G_{a_k} \) where \( a_k \) is the largest letter of \( x \), and so \( G_{a_k} \) is complete. Lemma[9] shows how the processing of \( G_{u(r, r)} \) makes the group \( G_{u(r+1, r+1)} \) complete, i.e. how it propagates through the refinement process.
Lemma 9. Referring to (B), for any $1 \leq r < \hat{r}$, after processing the group $G_{u(r+1,r)}$, the group $G_{u(r+1,r+1)}$ is complete.

Proof. Assume to have $i \in G_{u(r+1,r+1)}$ so that $u(r+1,r+1) = \mathcal{C}_{r+1}(i) = \mathcal{C}_r(i) = u(r,r+1)$ is not maximal, i.e. $u(r+1,r+1) \preceq x[j .. n]$, where $j = i + |u(r+1,r+1)|$. There are two cases.

- $G_{u(r+1,r+1)} = G_{u(r,r+1)}$.

  Then for some $t \geq 1$, and some $\xi$, $x[j .. n]$ has $(u(r,r+1))^t u(r,\xi)$ as a prefix and $u(r,\xi)$ is not a prefix of $u(r,r+1)$, and $\mathcal{C}_{r+1}(i) = u(r+1,r+1) = u(r,r+1) \preceq (u(r,r+1))^t u(r,\xi)$, and so $u(r+1,r+1) \prec u(r,\xi)$, and so $u(r,r) \preceq u(r,\xi)$ and $\xi \leq r$. Then we have $u(r+1,r+1)$ immediately followed by $u(\xi,\xi)$, $\xi < r+1$, and $u(r+1,r+1) \prec u(\xi,\xi)$, which contradicts Lemma 7.

- $G_{u(r+1,r+1)}$ was created from $G_{u(r,r+1)}$ and so $\mathcal{C}_{r+1}(i) = u(r,r+1)(u(r,r))^t$ for $t = val_{\alpha}(j)$. Let $\ell = i + t |u(r,r+1)|$. There are three subcases:

  (a) For some $p \geq 1$, some $t_1 > t$, $x[j .. n]$ has as a prefix $(u(r,r+1)(u(r,r))^t)^p u(r,r+1)(u(r,r))^t_1$. But then there is a group that has $u(r,r+1)(u(r,r))^t_2$, $t_2 \geq t_1$, as the context, which means that the group with the context $u(r,r+1)(u(r,r))^t$ cannot be $G_{u(r+1,r+1)}$, a contradiction.

  (b) For some $p \geq 1$, some $t_1 < t$, $x[j .. n]$ has as a prefix $(u(r,r+1)(u(r,r))^t)^p u(r,r+1)(u(r,r))^t_1 u(r,\xi)$ and $u(r,\xi) \neq u(r,r)$. Since $u(r,r+1) \prec u(r,\xi)$, it follows that $\xi \leq r$, but since $u(r,\xi) \neq u(r,r)$, we have $\xi < r$. So $u(r,\xi) = u(\xi,\xi)$. But then we have $u(r,r)$ immediately followed by $u(\xi,\xi)$ with $u(r,r) \prec u(\xi,\xi)$, which contradicts Lemma 7.

  (c) For some $p \geq 1$ and some $\xi$, $x[j .. n]$ has $(u(r,r+1)(u(r,r))^t)^p u(r,\xi)$ as a prefix and so that $u(r,r+1) \prec u(r,\xi)$. It follows that $\xi \leq r$. If $u(r,\xi) = u(r,r)$, then this would be case (a) as $x[j .. n]$ would have $(u(r,r+1)(u(r,r))^t)^{p-1} u(r,r+1)(u(r,r))^t_1$ as a prefix. Thus we can assume that $\xi < r$ and so $u(r,\xi) = u(\xi,\xi)$. But then we have $u(r,r)$ immediately followed by $u(\xi,\xi)$ with $u(r,r) \prec u(\xi,\xi)$, which contradicts Lemma 7.

Thus, we showed that for any $i \in G_{u(r+1,r+1)}$, $\mathcal{C}_{r+1}(i)$ is a maximal Lyndon factor, i.e. $G_{u(r+1,r+1)}$ is complete. \qed
Theorem 10. Referring to (B), \( \mathcal{C}_\hat{r} = [G_{u(\hat{r}, m_{\hat{r}})}, G_{u(\hat{r}, m_{\hat{r}} - 1)}, \ldots, G_{u(\hat{r}, 1)}] \) is an \( \hat{r} \)-proper configuration.

Proof. It is straightforward to see that \( \mathcal{C}_\hat{r} \) satisfies Def. 2(i)-(v). That it satisfies Def. 2(vi) follows from Lemma 6. Thus, \( \mathcal{C}_\hat{r} \) is a group configuration. That Def. 4(vii) is satisfied follows from Lemma 9. That Def. 4(viii) is satisfied follows from the fact that \( \mathcal{C}_\hat{r} \) was obtained by processing of \( G_{u(\hat{r} - 1, \hat{r} - 1)} \). Finally, the fact that Def. 4(ix) is satisfied follows from Lemma 8. \( \square \)

Theorem 11. For a string \( x \), Baier’s sort identifies and sorts all maximal Lyndon factors of \( x \).

Proof. Consider (B) when the process of refinement stops. Consider an index \( i \). There is a unique \( 1 \leq \ell \leq m_{\hat{r}} \) so that \( i \in G_{u(\hat{r}, \ell)} \). Thus \( x[i .. n] \) has \( u(\hat{r}, \ell) \) as a prefix. Since \( G_{u(\hat{r}, \ell)} \) is complete, \( x[i .. i + |u(\hat{r}, \ell)| - 1] = u(\hat{r}, \ell) \) is a maximal Lyndon factor. Hence all maximal Lyndon factors of \( x \) are accounted for. \( \square \)

5 Conclusion and future work

We showed how the process of refinement of the initial configuration propagates and is sustained as long as there are some \( i \)'s in the group being processed such that \( \text{prev}(i) \neq \text{nil} \). We showed that this process will identify and present in a sorted way all maximal Lyndon factors of a given string.

An algorithmic analysis of the process and implementation details are a necessary further step. Baier’s implementation of phase I of his algorithm is linear in time, due to the use of the \text{prev} operator. The \text{prev}() values can be computed for the initial configuration in \( \mathcal{O}(n) \) steps using a stack of size \( n \), where \( n \) is the length of the input string, and then during the processing they can be updated in \( \mathcal{O}(k) \) steps to work properly for the next level of refinement, where \( k \) is the the size of the group being processed. Our introduction of the \textit{valence} makes the treatment of repetitions of the context of the group being processed more mathematically sound and straightforward and it lends itself to a simpler algorithmic treatment allowing to process the indices of the group being processed simply from the largest to the smallest with very little overhead. Our C++ implementation uses all these mathematical insights and thus only uses memory of \( 13n \) integers, a significant improvement over Baier’s implementation. Moreover, our implementation uses a static approach, so all required memory of \( 13n \) integers is allocated once and no further dynamic memory allocation is required, significantly speeding up the execution. Of course, rigorous comparison testing of the performances of the two implementations is needed before any conclusion can be drawn. The code can be obtained from

http://www.cas.mcmaster.ca/~franek/research/bls.cpp
References


