A PROOF OF THE EXTENDED DUVAL'S CONJECTURE

ŠTĚPÁN HOLUB

Abstract. We give a short and elementary proof of the following stronger version of Duval's conjecture: Let u be an unbordered word, and v a word of length $|u|-1$, such that v is not a prefix of u. Then uv contains an unbordered word of length at least $|u|+1$.

Investigation of the relation between the length of a word and the length of its unbordered factors dates back to [ES79] and [AP79]. In recent years the topic was subject to research by Tero Harju and Dirk Nowotka in a series of papers ([HN02], [HN03b], [HNa], [HN03a], [HNb]). In the last one they offered a prove of the following statement.

Theorem. *Let u be an unbordered word, and v a word of length $|u|-1$, such that v is not a prefix of u. Then uv contains an unbordered word of length at least $|u|+1$.*

This is a slightly stronger version of the old conjecture formulated by J.–P. Duval in [Duv82].

Conjecture (Duval). *Let u and v be words such that $u \neq v$, $|u| = |v| = n$, and u is unbordered. Then uv contains an unbordered word of length at least $n+1$.*

A simple example from [HN03b] shows that the bound $|u|-1$ is optimal.

Example 1. *Consider the words*

$$u = a^i ba^j bb, \qquad\qquad v = a^j ba^i,$$

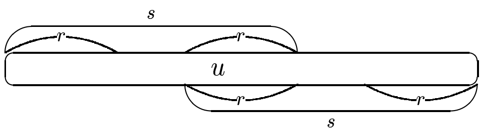
with $1 \leq i < j$. The word u is unbordered, v is not a prefix of u, and $|v| = |u| - 2$. It is easy to check that all factors of uv longer than $|u|$ are bordered.

Independently, the author of this paper presented in [Hol03] a short proof of the original Conjecture, based on the use of lexicographic orderings of words. The method has been inspired by the proof of the Critical Factorization Theorem given by M. Crochemore and D. Perrin in [CP91]. Here we employ the same method to obtain an alternative proof of the Theorem.

Preliminaries

We suppose that the reader is familiar with basic terminology as presented for example in [CK97]. The length of a word u is denoted by $|u|$. A word u is said to be *bordered* if and only if there exists a nonempty word r, $r \neq u$, which is both prefix and suffix of u. Any such r is called a *border* of u.

Remark. *It is easy to see that if u is bordered, it has a border of length at most $|u|/2$.*



The period of a word $s = l_1 l_2 \cdots l_{|s|}$ is the smallest positive integer $\delta = \delta(s)$, such that $l_i = l_{i+\delta}$, for each $1 \le i \le |s| - \delta$. Note that a word s is unbordered if and only if $\delta(s) = |s|$.

If $t = sr$, we write $s = tr^{-1}$.

We say that two lexicographic orderings $\lhd$ and $\blacktriangleleft$ are *mutually inverse* if

$$c \lhd d \Longleftrightarrow d \blacktriangleleft c,$$

for any two letters c and d from the domain alphabet

For a word $s = l_1 l_2 \cdots l_{|s|}$ denote by $\overline{s} = l_{|s|} l_{|s|-1} \cdots l_1$ its mirror image. We say that $\prec$ is a *mirror lexicographic ordering* if

$$s \prec t \Longleftrightarrow \overline{s} \lhd \overline{t},$$

for a lexicographic ordering $\lhd$. Informally, a mirror lexicographic ordering is a lexicographic ordering on words read from right to left.

Proof of the Theorem

The following proof is constructive. Claims 1–5 reveal how to find an unbordered factor of uv longer than $|u|$ in respective cases. In each case it is straightforward to verify that the factor indeed has the required length.

Put $n = |u|$. Let p denote the last letter of u.

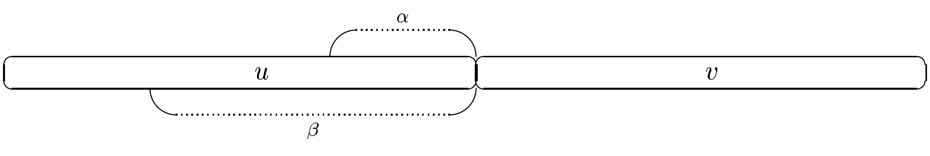
Claim 1. *Suppose that up^{-1} is a power of a single letter q. Then the Theorem holds.*

Proof. Since v is not a prefix of u, we have $v = v_1 q' v_2$ for a letter q' distinct from q. The Remark implies that the word uv_1q' is unbordered. □

We shall further suppose that up^{-1} contains (at least) two different letters.

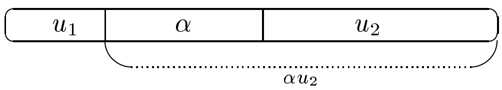
Consider two mutually inverse lexicographic orderings $\lhd$ and $\blacktriangleleft$ on factors of uv.

Let α (β, resp.) be the maximal suffix of u with respect to $\lhd$ ($\blacktriangleleft$, resp.). If two different letters in up^{-1} are chosen as maximal (and minimal) ones, we can suppose $1 < |\alpha| < |\beta| \le n$.



Lemma. *The factor α occurs just once in u.*

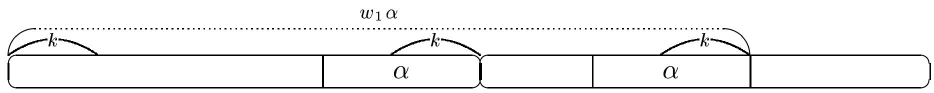
Proof. Let $u = u_1 \alpha u_2$, with nonempty u_2. Then $\alpha \lhd \alpha u_2$ yields a contradiction.



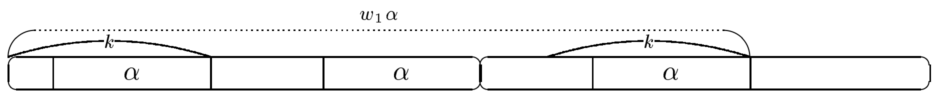
□

Claim 2. *Suppose that α has at least two occurrences in uv. Then the Theorem holds.*

Proof. Let $uv = w_1\alpha w_2$, with $|w_1\alpha| \neq n$. We show that the word $w_1\alpha$ is unbordered. The Lemma implies that $|w_1\alpha| > n$. Suppose for contradiction that k is the shortest border of $w_1\alpha$. Note that $|k| < n$, by the Remark. If $|k| < |\alpha|$, the word k is also a border of u, a contradiction.



If $|k| > |\alpha|$ then the word α is a suffix of k, and we obtain a contradiction with the Lemma.



□

For the rest of the paper we adopt the following

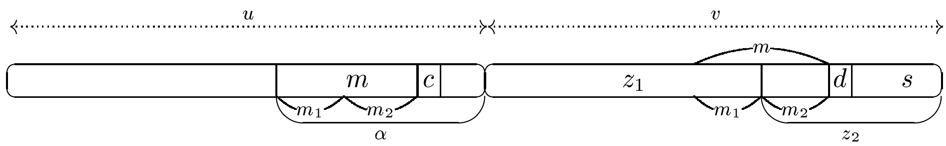
Assumption. *The word α has just one occurrence in uv.*

The remaining possibilities are divided in two cases.

Case 1. In the first case we suppose that αp^{-1} *is not* a suffix of v.

Let $v = z_1 z_2$ be a factorization of v such that $|z_2| = |\alpha| - 1$ and $|\alpha z_1| = n$. If the word αz_1 is bordered, its longest border denote by m_1. Otherwise, let m_1 be the empty word. Let m_2 be the longest prefix of z_2 such that $m = m_1 m_2$ is a prefix of αv (also m_2 can be empty).

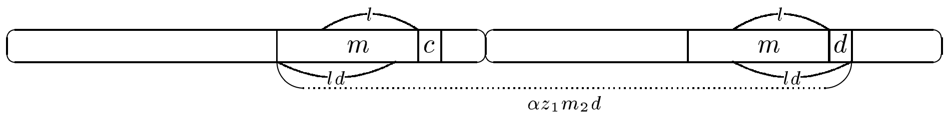
By the Assumption, the word m is shorter than α. Moreover, in the present case we suppose that $z_2 \neq \alpha p^{-1}$. Therefore, we have $z_2 = m_2 ds$, where d is a letter. The construction yields that mc is a prefix of α, for a letter c distinct from d.



We now indicate an unbordered factor of uv, which is longer than n. It will depend on the relation between d and e.

Claim 3. *If $d \lhd c$ then the word $\alpha z_1 m_2 d$ is unbordered.*

Proof. Suppose for contradiction that ld is a border of the word $\alpha z_1 m_2 d$. The definition of m, namely the maximality of both m_1 and m_2, implies that the word ld is a suffix of md. Therefore, l is a suffix of m. Since mc is a prefix of α, the word lc is a factor of α. But ld is a prefix of α, and $ld \lhd lc$ yields a contradiction with the definition of α.

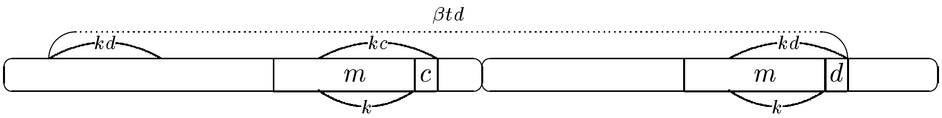


□

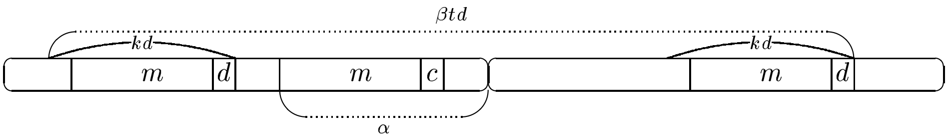
Claim 4. *If $d \blacktriangleleft c$ then the word βtd is unbordered.*

Proof. Suppose for a contradiction that kd is the shortest border of βtd.

If $|k| < |m|$ then k is a suffix of m, and kc a factor of α. Since kd is a prefix of β, the relation $kd \blacktriangleleft kc$ yields a contradiction with the definition of β.



Suppose, on the other hand, that $|k| \geq |m|$. By the Assumption, the word kd is shorter than β. Thus md, as a suffix of kd, is a factor of u. But $mc \lhd md$, a contradiction with the maximality of α.

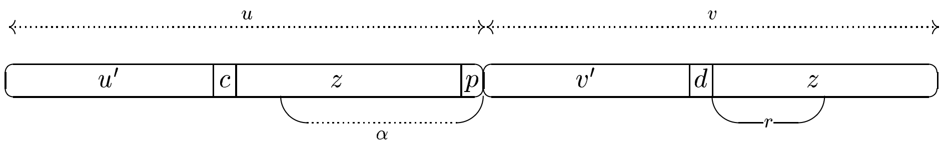


□

Case 2. In the second case we shall suppose that αp^{-1} *is* a suffix of v.

Let z be the maximal common suffix of v and up^{-1}. By the assumptions, $|\alpha| - 1 \leq |z| < n-1$. Let c and d be distinct letters, such that $u = u'czp$, and $v = v'dz$. Let $\prec$ be an arbitrary mirror lexicographic ordering satisfying $c \prec d$.

Let r be the prefix of z, such that $uv'dr$ is maximal with respect to the ordering $\prec$, i.e., for any prefix r' of z the relation $uv'dr' \prec uv'dr$ holds.

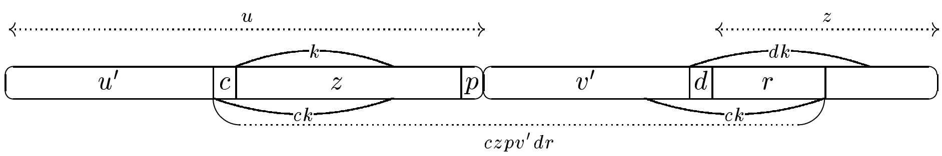


We are ready to point out the sought unbordered factor of this case.

Claim 5. *The word $czpv'dr$ is unbordered.*

Proof. Suppose for contradiction that ck is the shortest border of the word $czpv'dr$. Since α is a suffix of zp, the Assumption implies $|ck| \leq |cz|$. Therefore, k is a prefix of z.

Note that $uv'dk$ is a prefix of uv, and ck is a suffix of $uv'dr$. From $ck \prec dk$ we deduce $uv'dr \prec uv'dk$, a contradiction with the definition of r.



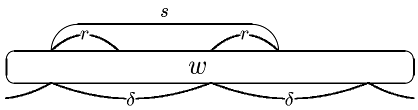
□

This completes the proof of the Theorem.

Open questions

As noted in the introduction, the Theorem is part of a broader question: How long a word w can be, provided that its longest unbordered factor is of length n?

It turns out immediately that the question is not very interesting if the word w is allowed to have the period n. Then it can be arbitrarily long, since each factor longer than the period of the word is clearly bordered.



Suppose, therefore, that the period of w is greater than n. In terms of the present paper the question can be formulated as follows:

Question 1. *Let $w = v_1uv_2$ be a word such that u is unbordered, the period of w is greater than $|u|$, and w does not contain any unbordered factor of length greater than $|u|$. What can be said about $|w|$?*

The Theorem, applied simultaneously on left hand and right hand extension of the word u, implies $|w| \leq 3n - 4$. In contrast to the bound of the Theorem, this bound is strongly believed not to be optimal. On the other hand the conjecture from [ES79] that $|w| < 2n$ was disproved in [AP79] by the following example.

Example 2. *Consider the words*

$$v_1 = a^i, \qquad u = ba^{i+1}ba^iba^{i+2}, \qquad v_2 = ba^iba^{i+1}ba^i.$$

The word u is unbordered, and the word $w = v_1uv_2$ does not contain any unbordered factor longer than u. For $i > 2$ we have $|w| = 7i + 10 > 2(3i + 6) = 2|u|$.

Note that in Example 2 the word v_1 is a suffix of u. That leads to the following question.

Question 2. *What can be said about $|w|$ if v_1 is not a suffix of u, and v_2 is not its prefix? In particular, can $|w| \geq 2|u|$?*

We conclude by an example of words satisfying the conditions of Question 2. Using the methods of this paper it turns out that short examples of this kind do not exist.

Example 3. *Consider the words*

$$v_1 = babb, \qquad u = abaabbababbaababbabaabbababbaabb, \qquad v_2 = abab.$$

The word u is unbordered, the word $w = v_1uv_2$ does not contain any unbordered factor longer than u, v_1 is not a suffix of u, and v_2 is not its prefix.

Acknowledgment

I am grateful to Dirk Nowotka for introducing me to the exciting Duval's conjecture. I also thank Petr Pitřinec, who wrote a computer program searching for examples and counter–examples.

References

[AP79] R. Assous and M. Pouzet. Une caractérisation des mots périodiques. *Discrete Math.*, 25(1):1–5, 1979.

[CK97] Christian Choffrut and Juhani Karhumäki. Combinatorics of words. In G.Rozenberg and A.Salomaa, editors, *Handbook of formal languages, Vol. 1*, pages 329–438. Springer, Berlin, 1997.

[CP91] Maxime Crochemore and Dominique Perrin. Two-way string-matching. *J. Assoc. Comput. Mach.*, 38(3):651–675, 1991.

[Duv82] J.-P. Duval. Relationship between the period of a finite word and the length of its unbordered segments. *Discrete Math.*, 40(1):31–44, 1982.

[ES79] Andrzej Ehrenfeucht and D. M. Silberger. Periodicity and unbordered segments of words. *Discrete Math.*, 26(2):101–109, 1979.

[HNa] Tero Harju and Dirk Nowotka. Minimal Duval extensions. *Internat. J. Found. Comput. Sci.* to appear.

[HNb] Tero Harju and Dirk Nowotka. Periodicity and unbordered words. *STACS (Montpellier 2004)*. to appear.

[HN02] Tero Harju and Dirk Nowotka. Duval's conjecture and Lyndon words. Technical Report 479, TUCS - Turku Centre for Computer Science, Oct 2002.

[HN03a] Tero Harju and Dirk Nowotka. About Duval extensions. In T. Harju and J. Karhumäki, editors, *WORDS 2003 (Turku)*, number 27 in TUCS General Publications, pages 392–397, Finland, Aug 2003.

[HN03b] Tero Harju and Dirk Nowotka. About Duval's conjecture. In Z. Esik and Z. Fülöp, editors, *Developments in Language Theory (Szeged 2003)*, volume 2710 of *Lecture Notes in Computer Science*, pages 316–324. Springer-Verlag, Jul 2003.

[Hol03] Štěpán Holub. A proof of Duval's conjecture. In T. Harju and J. Karhumäki, editors, *WORDS 2003 (Turku)*, number 27 in TUCS General Publications, pages 398–399, Finland, Aug 2003.