About Duval's Conjecture

Tero Harju and Dirk Nowotka

Turku Centre for Computer Science (TUCS)
Department of Mathematics, University of Turku, Finland

Abstract. A word is called unbordered if it has no proper prefix which is also a suffix of that word. Let $\mu(w)$ denote the length of the longest unbordered factor of a word w. Let a word where the longest unbordered prefix equal to $\mu(w)$ be called Duval extension. A Duval extension is called trivial, if its longest unbordered factor is of the length of the period of that Duval extension. In 1982 it was shown by Duval that every Duval extension w longer than $3\mu(w) - 4$ is trivial. We improve that bound to $5\mu(w)/2-1$ in this paper, and with that, move closer to the bound $2\mu(w)$ conjectured by Duval. Our proof also contains a natural application of the Critical Factorization Theorem.

1 Introduction

The periodicity and the borderedness of words are two basic subjects of interest in the study (e.g., Chapter 8 in [10]) and application (e.g., [8]) of combinatorics on words. We investigate the relation between periodicity and the length of borders in this paper. In particular, we improve a result by Duval [4] which relates the length of the longest unbordered factor of a word to the period of that word.

Let us consider an arbitrary finite word w of length n. The period of w, denoted by $\partial(w)$, is the smallest positive integer p such that the ith letter equals the $(i+p)$th letter for all $1 \leq i \leq n-p$. Let $\mu(w)$ denote the length of the longest unbordered factor of w, where a word v is bordered, if it has a proper prefix u, which is neither empty nor v itself, such that u is also a suffix of v. Assume w has an unbordered prefix u of maximum length $\mu(w)$, then w is called *Duval extension* of u, and w is called trivial Duval extension if u is of length $\partial(w)$.

In 1979 Ehrenfeucht and Silberger started a line of research [5,1,4] exploring the relation between the length of the longest unbordered factor, $\mu(w)$, and the period, $\partial(w)$. In 1982 these efforts culminated in Duval's result: If $n \geq 4\mu(w)-6$ then $\partial(w) = \mu(w)$. However, it is believed that $\partial(w) = \mu(w)$ holds for $n \geq 3\mu(w)$ which follows if Duval's conjecture [4] holds true.

Conjecture 1 (Duval's conjecture). If $n \geq 2\mu(w)$ and w is a Duval extension, then $\partial(w) = \mu(w)$.

After that, no progress was recorded, to the best of our knowledge, for 20 years. However, the topic remained popular; see for example Problem 8.2.13 on page 308 in Chapter 8 of [10]. Only recently the theme was independently picked up again by Mignosi and Zamboni [11] and us [7]. However, these papers investigate not

Duval's conjecture but rather its opposite, that is: which words admit only trivial Duval extensions? It is shown in [11] that unbordered Sturmian words allow only trivial Duval extensions, in other words, if an unbordered Sturmian word of length $\mu(w)$ is a prefix of w, then $\partial(w) = \mu(w)$. That result was improved in [7] by showing that Lyndon words allow only trivial Duval extensions and the fact that every unbordered Sturmian word is a Lyndon word.

We show in this paper that, if w is a Duval extension and $n \geq 5\mu(w)/2$ then $\partial(w) = \mu(w)$. This is a first improvement of Duval's result which is given implicitly in [4], namely that, if w is a Duval extension and $n \geq 3\mu(w) - 3$ then $\partial(w) = \mu(w)$. Duval's conjecture, which is actually believed to hold even in the slightly stronger version where $n \geq 2\mu(w) - 1$ is required, may lead to an ultimate bound for the length of words that have $\partial(w) \neq \mu(w)$. Note, that the bound $n \geq 2\mu(w) - 1$ is sharp by the following example. Let $w = a^i ba^j bba^j ba^i$ with $1 \leq i < j$. Then $\mu(w) = i + j + 3$ and $\partial(w) = \mu(w) + (j - i) = 2j + 3$ and $|w| = 2\mu(w) - 2 = 2(i + j + 2)$.

Section 4 presents the proof of our main result, Theorem 2, which uses the notations from Section 2 and preliminary results from Section 3. We conclude with Section 5.

2 Notations

In this section we define the notations of this paper. We refer to [9,10] for more basic and general definitions.

We consider a finite alphabet A of letters. Let A^* denote the monoid of all finite words over A including the empty word, denoted by ε. A nonempty word u is called a *border* of a word w, if $w = uv = v'u$ for some suitable words v and v'. We call w *bordered* if it has a border that is shorter than w, otherwise w is called *unbordered*. Note, that every bordered word w has a minimum border u such that $w = uvu$, where clearly u is unbordered. Let $w = w_{(1)}w_{(2)}\cdots w_{(n)}$ where $w_{(i)}$ is a letter, for every $1 \leq i \leq n$. Then we denote the length n of w by $|w|$. An integer $1 \leq p \leq n$ is a *period* of w, if $w_{(i)} = w_{(i+p)}$ for all $1 \leq i \leq n - p$. The smallest period of w is called the *minimum period* of w. Let $w = uv$. Then u is called a *prefix* of w, denoted by $u \leq w$, and v is called a *suffix* of w, denoted by $v \preccurlyeq w$. Let $u, v \neq \varepsilon$. Then we say that u *overlaps* v *from the left* or *from the right*, if there is a word w such that $|w| < |uv|$, and $u \leq w$ and $v \preccurlyeq w$, or $v \leq w$ and $u \preccurlyeq w$, respectively. We say that u *overlaps* v, if either v is a factor of u, or u overlaps v from the left or right.

Let w be a nonempty word of length n. We call wu a *Duval extension* of w, if every factor of wu longer than n is bordered. A Duval extension wu of w is called *trivial*, if there exists a positive integer j such that $u \leq w^j$, that is, the minimum period of wu is n.

Example 1. Let $w = abaabb$ and $u = aaba$. Then

$$wu = abaabbaaba$$

is a nontrivial Duval extension of w where $\partial(wu) = 7$ and $\mu(wu) = |w| = 6$. Actually, wu is the longest possible nontrivial Duval extension of w. The word

$$w = aababb$$

is not a prefix of any nontrivial Duval extension wu where $\mu(wu) = |w| = 6$.

Let an integer p with $1 \leq p < |w|$ be called *position* or *point* in w. Intuitively, a position p denotes the place between $w_{(p)}$ and $w_{(p+1)}$ in w. A word $u \neq \varepsilon$ is called a *repetition word* at position p if $w = xy$ with $|x| = p$ and there exist x' and y' such that $u \preccurlyeq x'x$ and $u \leq yy'$. For a point p in w, let

$$\partial(w,p) = \min\big\{|u| \;\big|\; u \text{ is a repetition word at } p\big\}$$

denote the *local period* at point p in w. Note, that the repetition word of length $\partial(w,p)$ at point p is unbordered, and we have $\partial(w,p) \leq \partial(w)$. A factorization $w = uv$, with $u, v \neq \varepsilon$ and $|u| = p$, is called *critical* if $\partial(w,p) = \partial(w)$, and, if this holds, then p is called *critical point*.

3 Preliminary Results

We state some auxiliary and well-known results in this section which will be used to prove our main contribution, Theorem 2, in Section 4.

Lemma 1. *Let $zf = gzh$ where $f, g \neq \varepsilon$. Let az' be the maximum unbordered prefix of az. If az does not occur in zf, then agz' is unbordered.*

Proof. Assume agz' is bordered, and let be y its shortest border. In particular, y is unbordered. If $|z'| \geq |y|$ then y is a border of az' which is a contradiction. If $|az'| = |y|$ or $|az| < |y|$ then az occurs in zf which is again a contradiction. If $|az'| < |y| \leq |az|$ then az' is not maximum since y is unbordered; a contradiction. □

Lemma 2. *Let w be an unbordered word and $u \leq w$ and $v \preccurlyeq w$. Then uw and wv are unbordered.*

Proof. Obvious. □

The critical factorization theorem (CFT) was discovered by Césari and Vincent [2] and developed into its current form by Duval [3]. We refer to [6] for a short proof of the CFT.

Theorem 1 (CFT). *Every word w, with $|w| \geq 2$, has at least one critical factorization $w = uv$, with $u, v \neq \varepsilon$ and $|u| < \partial(w)$, i.e., $\partial(w, |u|) = \partial(w)$.*

Lemma 3. *Let $w = uv$ be unbordered and $|u|$ be a critical position of w. Then u and v do not overlap.*

Proof. Note, that $\partial(w,|u|) = \partial(w) = |w|$ since w is unbordered. Let $|u| \leq |v|$ without restriction of generality. Assume that u and v overlap. If $u = u's$ and $v = sv'$, then $\partial(w,|u|) \leq |s| < |w|$. If $u = su'$ and $v = v's$, then w is bordered with s. If $v = sut$ then $\partial(w,|u|) \leq |su| < |w|$. □

Lemma 4. *Let uv be unbordered and $|u|$ be a critical position of uv. Then either vxu is unbordered or has a minimum border g such that $|g| \geq |uv|$, for any word x.*

Proof. Follows directly from Lemma 3. □

The following Lemmas 5, 6 and 7 and Corollary 1 are given in [4]. Let $a_0, a_1 \in A$, with $a_0 \neq a_1$, and $t_0 \in A^*$. Let the sequences (a_i), (s_i), (s_i'), (s_i''), and (t_i), for $i \geq 1$, be defined by

- $a_i = a_{i \pmod 2}$, that is, $a_i = a_0$ or $a_i = a_1$ if i is even or odd, respectively,
- s_i such that $a_i s_i$ is the shortest border of $a_i t_{i-1}$,
- s_i' such that $a_{i+1} s_i'$ is the longest unbordered prefix of $a_{i+1} s_i$,
- s_i'' such that $s_i' s_i'' = s_i$,
- t_i such that $t_i s_i'' = t_{i-1}$.

For any parameters of the above definition, we have the following.

Lemma 5. *For any a_0, a_1, and t_0 there exists an $m \geq 0$ such that*

$$|s_1| < |s_2| < \cdots < |s_m| = |t_{m-1}| \leq \cdots \leq |t_1| \leq |t_0|$$

and $s_m = t_{m-1}$.

Lemma 6. *Let $z \leq t_0$ such that $a_0 z$ and $a_1 z$ do not occur in t_0. Let $a_0 z_0$ and $a_1 z_1$ be the longest unbordered prefixes of $a_0 z$ and $a_1 z$, respectively. Then*

1. *if $m = 1$ then $a_0 t_0$ is unbordered,*
2. *if $m > 1$ is odd, then $a_1 s_m$ is unbordered and $|t_0| \leq |s_m| + |z_0|$,*
3. *if $m > 1$ is even, then $a_0 s_m$ is unbordered and $|t_0| \leq |s_m| + |z_1|$.*

Lemma 7. *Let v be an unbordered factor of w of length $\mu(w)$. If v occurs twice in w, then $\mu(w) = \partial(w)$.*

Corollary 1. *Let wu be a Duval extension of w. If w occurs twice in wu, then wu is a trivial Duval extension.*

4 An Improved Bound for Duval Extensions

We present the main result of this paper. Note that the proof of Theorem 2 rests to a great extend on the application of the CFT by applying Lemma 4.

Theorem 2. *Let wu be a nontrivial Duval extension of w. Then $|u| < 3|w|/2$.*

Proof. Recall that every factor of wu which is longer than $|w|$ is bordered since wu is a Duval extension.

Let z be the longest suffix of w that occurs twice in zu. If $z = \varepsilon$ then $a \preccurlyeq w$ and $u = b^j$, where $a, b \in A$ and $a \neq b$ and $j \geq 1$, but now $|u| < |w|$ since ab^j is unbordered. So, $z \neq \varepsilon$. We have that $z \neq w$ since wu is otherwise trivial by Corollary 1. Let $a, b \in A$ such that

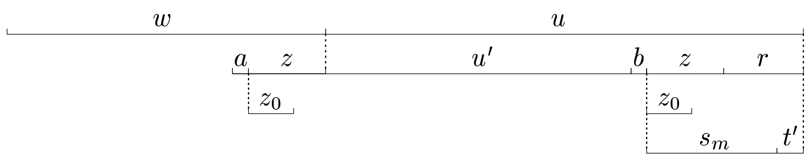
$$w = w'az \qquad \text{and} \qquad u = u'bzr$$

and z occurs in zr only once, that is, bz matches the rightmost occurrence of z in u. Note, that bz does not overlap az from the right, and therefore u' exists, by Lemma 2. Naturally, $a \neq b$ by the maximality of z, and $w' \neq \varepsilon$, otherwise $azu'bz \leq wu$ has either no border or w is bordered or az occurs in zu; a contradiction in any case. Let $a_0 = a$ and $a_1 = b$ and $t_0 = zr$, and let z_0 and z_1 and the sequences (a_i), (s_i), (s_i'), (s_i''), (t_i), and the integer m be defined as in Lemma 6. Let

$$t_0 = s_m t' \ .$$

Consider $azu'bz_0$. We have that az and $azu'bz_0$ are both prefixes of a_0zu, and bz_0 is a suffix of $azu'bz_0$ and az does not occur in $zu'bz_0$. From Lemma 1 it follows that $azu'bz_0$ is unbordered, and hence,

$$|azu'bz_0| \leq |w| \tag{1}$$



Case: Suppose that m is even. Then as_m $(= a_m s_m)$ is unbordered and $|t_0| \leq |s_m| + |z_1|$ by Lemma 6. From (1) it follows that $|z_1| \leq |z| \leq |w| - 2$. If $|s_m| \leq |z_0|$, then $|azu| \leq |w| + |z_1|$, and hence, $|u| \leq |w|$, since we have

$$|azu| = |azu'bz_0| - |z_0| + |t_0| \leq |azu'bz_0| - |z_0| + |s_m| + |z_1| \ .$$

Suppose then that $|s_m| > |z_0|$. We have that as_m is unbordered, and since az_0 is the longest unbordered prefix of az, we have $az \leq as_m$, and hence, $|z| \leq |s_m|$. Now, the word $azu'as_m$ is unbordered or otherwise its shortest border is longer than az, since no prefix of az is a suffix of as_m, and az occurs in u; a contradiction. So, $|azu'as_m| \leq |w|$ and $|u| < |w|$, since $|z_1| \leq |z|$.

Case: Suppose that m is odd. Then bs_m $(= a_m s_m)$ is unbordered and $|t_0| \leq |s_m| + |z_0|$ (see Lemma 6). If $s_m = \varepsilon$. Then $|t_0| \leq |z_0|$ and $t_0 = z_0$, since $z_0 \leq t_0$, and hence, $|azu| \leq |w|$, by (1). So, assume $s_m \neq \varepsilon$. If $|s_m| < |z|$, then $|u| < |w|$ since

$$|u| = |azu'bz_0| - |bz_0| + |bt_0| - |az|$$

and

$$|azu'bz_0| \leq |w| \qquad \text{and} \qquad |t_0| \leq |s_m| + |z_0| \ .$$

So, assume $|s_m| \geq |z|$. Since $|bs_m| \geq 2$ there exists a critical point p in bs_m such that $bs_m = v_0v_1$, where $|v_0| = p$, by the CFT.

Consider $wu'bz$ which is bordered and must have a shortest border longer than z, otherwise w is bordered since we have $z \preccurlyeq w$. So, bz occurs in w. Note, that $|az_0| \leq \partial(az)$ and that bz occurs left from az in w. If az and bz do not overlap in w then $|az_0| \leq |az| \leq |w|/2$. Also, if bz overlaps az from the left, then $|az_0| \leq \partial(az) \leq |w|/2$. It follows that

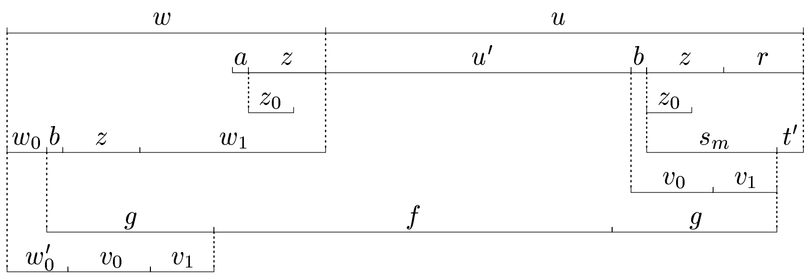
$$\text{if} \quad |u| < |w| + |z_0| \quad \text{then} \quad |u| < 3|w|/2 \ . \tag{2}$$

Let

$$w = w_0bzw_1$$

where bz occurs only once in w_0bz, that is, we consider the leftmost occurrence of bz in w. Consider the factor

$$f = bzw_1u'bs_m \ . \tag{3}$$

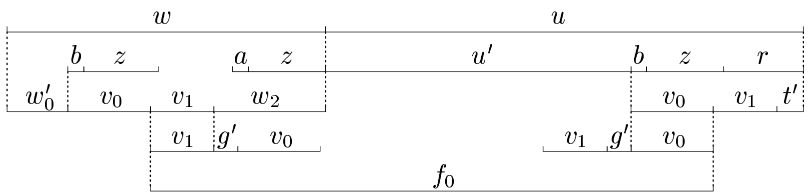


If f is unbordered then $|f| \leq |w|$, and hence, $|u| \leq |w| - 2$. Assume f is bordered, and let g be its shortest border. If $g \leq bz$ then bs_m is bordered which is a contradiction. If $|bz| < |g| \leq |s_m|$ then bz occurs in zr; a contradiction as well. Hence, $|bs_m| \leq |g|$. Note, that also $|g| < |bzw_1|$ since az does not occur in u. So, we have

$$w = w_0'bs_mw_2 = w_0'v_0v_1w_2$$

where $w_0 \leq w_0'$ and $w_2 \neq \varepsilon$. Consider

$$f_0 = v_1w_2u'v_0 \ .$$



If f_0 is unbordered, then $|f_0| \leq |w|$ and

$$|u| = |v_1 w_2 u' v_0| - |v_0| + |bt_0| - |v_1 w_2| \leq |w| + |z_0| - |w_2|$$

since $|bt_0| \leq |bs_m| + |z_0| = |v_0 v_1| + |z_0|$, and we have $|u| < 3|w|/2$ by (2). Assume that f_0 is bordered. Then its shortest border g_0 is longer than $|v_0 v_1|$ by Lemma 4. Let

$$g_0 = v_1 g' v_0 \ .$$

If $|g_0| > |v_1 w_2|$. Then $|v_1 w_2| < |az|$ otherwise az occurs in u; a contradiction. But, now v_1 occurs in z, and hence, overlaps with v_0 since $bz \leq v_0 v_1$ and $w_2 \neq \varepsilon$. This contradicts Lemma 3.

Assume that $|g_0| \leq |v_1 w_2|$, and let w_2' be such that $v_0 w_2' \preccurlyeq w_2$. Consider

$$f_1 = v_0 w_2' u' v_0 v_1 \ .$$

If f_1 is unbordered, then $|f_1| \leq |w|$ and $|u| < |f_1| + |t'|$. We have $|u| < |w| + |z_0|$, and hence, $|u| < 3|w|/2$ by (2). Assume now that g_1 is the shortest border of f_1. We have $|g_1| \geq |v_0 v_1|$ by Lemma 4, and therefore $v_0 v_1 \preccurlyeq g_1$. If $g_1 \leq v_0 w_2'$ then $v_0 v_1$ has two nonoverlapping occurrences in w and is therefore at most half as long as w. We have then

$$|u| \leq |azu'bz_0| - |az| - |bz_0| + |v_0 v_1| + |t'|$$

and from $|azu'bz_0| \leq |w|$ and $|v_0 v_1| \leq |w|/2$ and $|t'| \leq |z_0|$ and (2) it follows that $|u| \leq 3|w|/2 - 3$. So, assume $v_0 w_2' \leq g_1$. We have $|az| > |v_0 w_2'|$ since az does not occur in u. Certainly, $v_0 v_1$ occurs in u'. Indeed, it does not overlap with az from the right by Lemma 2 since $bz \leq v_0 v_1$. Let

$$x v_0 v_1 \leq u' \ .$$

Consider

$$f' = bzw_1 x v_0 v_1 \ .$$

If f' is unbordered than we have two nonoverlapping occurrences of $v_0 v_1$ in a factor that is at most as long as $|w|$ and $|u| < 3|w|/2 - 2$. Suppose, f' is bordered, then its shortest border is longer than $v_0 v_1$ since otherwise g is not the shortest border of f; see (3). In fact, the shortest border of f' is g since otherwise we have again two nonoverlapping occurrences of $v_0 v_1$ in a factor that is at most as long as $|w|$ and $|u| < 3|w|/2 - 2$. Moreover, $|azx| > |gv_0 v_1|$ otherwise az occurs in u (in the border of f). Actually, $|x| \geq |gv_0 v_1|$ since $bz \leq g$ and bz does not overlap w from the right by Lemma 2. Let now $x'bz \leq xbz$ where bz occurs only once in $x'bz$. Then the shortest border of $wx'bz$ is at least as long as $|w_0 bz|$ and at most as long as $|x'bz|$. So, $|w_0'| \leq |x|$. We have

$$w = w_0' v_0 v_1 g' v_0 w_2' \tag{4}$$

and

$$u = x v_0 v_1 u'' v_0 w_2' x b t_0 \ . \tag{5}$$

Moreover, $v_1 g' \preccurlyeq x v_0 v_1 u'' v_0 w_2' x$ by the border g_0 of f_0. Note, that

$$|g'| \leq |u'' v_0 w_2' x| \tag{6}$$

otherwise $|g'| \geq |v_0 v_1 u'' v_0 w_2' x|$ since v_0 and v_1 do not overlap by Lemma 4, but then $v_0 v_1$ has two nonoverlapping occurrences in w and $|v_0 v_1| \leq |w|/2 - 2$, where the constant 2 comes from the fact that v_0 has at least four nonoverlapping occurrences in w, and $|u| \leq 3|w|/2 - 3$. Now,

$$|azxv_0 v_1 u'' v_0 w_2' x b z_0| \leq |w|$$

by (1) and (5), and

$$|azxu'' x b z_0| \leq |w_0' g'|$$

by (4) and $|w_0'| \leq |x|$, and

$$|azbz_0| \leq |v_0 w_2'|$$

by (6), and

$$|az| < |v_0 w_2'|$$

and hence, az occurs in u; a contradiction. $\square$

5 Conclusions

We have lowered the bound of nontrivial Duval extentions from $3\mu(w) - 4$ in [4] to $5\mu(w)/2 - 1$ which brings us closer to the improved Duval's conjecture of $2\mu(w) - 2$. It should be noted that our result rests to a great part on the critical factorization theorem, which is new application in this context, and might help in finding a new approach to eventually solving the slightly improved version of Duval's conjecture by finding a sharp bound for the length of words which contain no unbordered factor of the length of of their period or longer.

References

1. R. Assous and M. Pouzet. Une caractérisation des mots périodiques. *Discrete Math.*, 25(1):1–5, 1979.
2. Y. Césari and M. Vincent. Une caractérisation des mots périodiques. *C. R. Acad. Sci. Paris Sér. A*, 286:1175–1177, 1978.
3. J.-P. Duval. Périodes et répétitions des mots de monoïde libre. *Theoret. Comput. Sci.*, 9(1):17–26, 1979.
4. 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.
5. A. Ehrenfeucht and D. M. Silberger. Periodicity and unbordered segments of words. *Discrete Math.*, 26(2):101–109, 1979.
6. T. Harju and D. Nowotka. Density of critical factorizations. *Theor. Inform. Appl.*, 36(3):315–327, 2002.
7. T. Harju and D. Nowotka. Duval's conjecture and Lyndon words. submitted to Discrete Math., 2002.

8. D. E. Knuth, J. H. Morris, and V. R. Pratt. Fast pattern matching in strings. *SIAM J. Comput.*, 6(2):323–350, 1977.
9. M. Lothaire. *Combinatorics on Words*, volume 17 of *Encyclopedia of Mathematics*. Addison-Wesley, Reading, Massachusetts, 1983.
10. M. Lothaire. *Algebraic Combinatorics on Words*, volume 90 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, United Kingdom, 2002.
11. F. Mignosi and L. Q. Zamboni. A note on a conjecture of Duval and Sturmian words. *Theor. Inform. Appl.*, 36(1):1–3, 2002.