Volker Diekert | Tero Harju | Dirk Nowotka

Weinbaum Factorizations of Primitive Words

Turku Centre *for* Computer Science

TUCS Technical Report
No 780, July 2006

Weinbaum Factorizations of Primitive Words

Volker Diekert
Institute of Formal Methods in Computer Science,
University of Stuttgart, D-70569 Stuttgart, Germany
volker.diekert@informatik.uni-stuttgart.de
Tero Harju
Turku Centre for Computer Science and
Department of Mathematics,
University of Turku, FIN-20014 Turku, Finland
harju@utu.fi
Dirk Nowotka
Institute of Formal Methods in Computer Science,
University of Stuttgart, D-70569 Stuttgart, Germany
dirk.nowotka@informatik.uni-stuttgart.de

TUCS Technical Report
No 780, July 2006

Abstract

C. M. Weinbaum [Proc. AMS, 109(3):615–619, 1990] showed the following: Let w be a primitive word and a be letter in w. Then a conjugate of w can be written as uv such that a is a prefix and a suffix of u, but v neither starts nor ends with a, and u and v have a unique position in w as cyclic factors. The latter condition means that there is exactly one conjugate of w having u as a prefix and there is exactly one conjugate of w having v as a prefix. It is this condition which makes the result non-trivial.

We give a simplified proof for Weinbaum's result. Guided by this proof we exhibit quite different, but still simple, proofs for more general statements. For this purpose we introduce the notion of *Weinbaum factor* and *Weinbaum factorization*.

Keywords: combinatorics on words, Weinbaum factorization, critical points, bordered word, primitive words

TUCS Laboratory
Discrete Mathematics for Information Technology

1 Introduction

Words w and w' are conjugated, if they can be written as $w = vu$ and $w' = uv$. Cyclic factors of a primitive word w are factors of conjugates of w. A cyclic factor u has a unique position in w, if there is exactly one conjugate of w having u as a prefix. Weinbaum showed in [5] that for each letter a occurring in a primitive word w there exists a conjugate uv of w such that: both factors u and v are uniquely positioned in w, the cyclic factor u begins and ends with a, but v neither begins nor ends with a.

In this paper we present variations and simple proofs of Weinbaum's result. Guided by these proofs we obtain in fact more general statements. This leads us to the notion of *Weinbaum factor* and *Weinbaum factorization* (or W-*factorization* for short). A Weinbaum factor of a primitive word w is a cyclic factor which satisfies some natural condition which is sufficient to prove an analogue of Weinbaum's result, if we replace the letter a by a factor f. Moreover, a Weinbaum factor provides us with another factor g, which we call a *complementary marker*, and in Theorem 10 we establish a W-factorization for the pair (f, g) instead of a single letter (or factor) as in Weinbaum's original result.

In Section 5 we prove the existence of W-factorizations by employing Lyndon words w.r.t. specific lexicographic orders. In Section 6 we consider W-factorizations without orderings of the alphabets. The main result there is Theorem 12, it states that a W-factorization of a word w for a Weinbaum factor f is found be iterating a given relation W at most $\log_\Phi(n)$ many times for $n = |w|$, where Φ is the golden ratio. In Section 7 we prove that the bound is tight by providing a lower bound for this number of iterations by a variant of Fibonacci words which correspond to so-called *singular factors* in the infinite Fibonacci sequence.

Finally, in Sections 9 and 8 we somewhat reverse our viewpoint. We start with a word f (or with a suitable pair of words f and g). Proposition 22 says that for all f with at least two letters there are primitive words having f as a factor, but without any W-factorization for f. On the other hand, the proportion of words with a W-factorization for f and g tends rapidly to 1, as soon as f and g satisfy some trivial necessary condition. Pairs satisfying this conditions are called *Weinbaum candidates*. Starting with such a pair (f, g), we show in Proposition 19 that all long enough random words have W-factorization for f and g. So, in general, we can hardly expect any structural property of triples (w, f, g) which is simultaneously necessary and sufficient in order to characterize the existence of a W-factorization of w for f and g.

2 Preliminaries

Let A be an alphabet, i.e., a finite set of letters, and let A^* be the free monoid over A. We denote the *empty word* by ε, and A^+ is the free semigroup over A. Thus, $A^+ = A^* \setminus \{\varepsilon\}$. The length of a word $w \in A^*$ is denoted by $|w|$. Let $w = uv$. Then u is called a *prefix* of w, denoted by $u \leq_{\mathrm{p}} w$, and v is called a *suffix* of w, denoted by $v \leq_{\mathrm{s}} w$. We also write $u \leq_{\mathrm{p}} wA^*$, if $u \leq_{\mathrm{p}} w'$ for some $w' \in wA^*$, analogously we write $v \leq_{\mathrm{s}} A^*w$.

A word $w \in A^*$ is called *primitive*, if it is not a power of a different word, i.e., $w = x^k$ implies $k = 1$. A word w' is a *conjugate* of w, if there is a word z such that $zw = w'z$. For words *conjugation* is equivalent to *transposition*; cf. page 7 in [2]. This means, w' is a *conjugate* of w, if there are words u and v such that $w = vu$ and $w' = uv$. A word f is a (*proper*) *factor* of w if $w = ufv$ (and $\varepsilon \neq uv \neq w$).

We say that f *occurs* in w if f is a factor of w. We say that a word f is a *cyclic factor* of w if both, $|f| \leq |w|$ and f is a factor of w^2. Note that all conjugates of a word w are factors of w^2. Hence, the cyclic factors of w are exactly the factors of conjugates of w. A cyclic factor f is *uniquely positioned* in w, if there exists a unique conjugate w' of w having f as a prefix.

Two words u and v are said to *intersect in* (the cyclic word) w, if w^3 has a factor xyz such that $u = xy$ and $v = yz$ or $u = yz$ and $v = xy$, where x, y and z are nonempty words, or if u is a factor of v, or if v is a factor of u. We say that a word u has a *self-intersecting* occurrence in w (or intersects itself) if it intersects itself in a non-trivial way in w. A word f is called a *marker* if f does not intersect itself in w. Note that f may have a non-trivial overlap with itself, but then the overlap must not occur in w. For example, *aba* is a marker in *ababb*, but not in *ababa*.

If f is a marker in w with $f \leq_{\mathrm{p}} w$, then w has a unique factorization of the form $w = fz_1fz_2 \cdots fz_k$ with $z_i \in A^*$ and $fz_if \notin A^+fA^+$.

3 Weinbaum's Theorem

Weinbaum's original result is the following theorem for the case $m = 1$.

Theorem 1 ([5]). *Let w be a primitive word and a^m be a cyclic factor of w with $m \geq 1$. Then some conjugate w' of w has a factorization $w' = uv$ where u and v are uniquely positioned in w with $u \in a^mA^* \cap A^*a^m$ and $v \notin aA^* \cup A^*a$.*

Proof. Let $f = a^m$. Let us choose some longest cyclic factor g of w with $g \notin aA^* \cup A^*a \cup A^*fA^*$, but where fgf is a cyclic factor of w^2. Such a factor g exists. Indeed, some conjugate of w begins with f followed by a letter which is not a. We start a factor g here and we stop just before we see in w^2 the factor f again. Then the last letter of g is not a neither. Hence

$g \notin aA^* \cup A^*a \cup A^*fA^*$. Since there is at least one such factor we can choose the longest one. The crucial observation is that every cyclic occurrence of g in w is preceded and followed by f, and, moreover, g is a marker.

Now, either g is a letter or we replace g by some new letter b. We obtain a (new) word $\overline{w}$. If we prove exactly the same statement for $\overline{w}$ and b, then we obtain (up to the ordering of the two factors) the desired factorization for w and f.

Indeed, let $\overline{w} = v'u'$ where v' and u' are uniquely positioned in $\overline{w}$ and v' begins and ends with b, but u' neither begins nor ends with b. Since we allow conjugation we may consider $\overline{w}' = u'v'$ as well. This corresponds to a desired factorizations $w = uv$, we just have to see that u and v are uniquely positioned in w. This is true because g has been a marker without intersection of a^m in w.

By induction on the length of w we may assume that $g = b$ is in fact a letter. Now the repetition of the process starting with b yields a new factor h which corresponds to g in the first round. Again, we can use either induction or h is a single letter. Hence we may assume that h is a letter, too. Since b is preceded and followed by f in w, this means $m = 1$ and $h = a$. But this implies $w \in (ab)^+ \cup (ba)^+$. Since w is primitive we obtain $w = ab$ or $w = ba$, and the result becomes trivial. □

4 Weinbaum Factorizations and Factors

Let w be a primitive word with a conjugate w', and let f be a word. Then $w' = uv$ is called a *Weinbaum factorization of w for f* (or W-*factorization* for short), if u and v are uniquely positioned in w, and $u \in fA^* \cap A^*f$ and $v \notin fA^* \cup A^*f$. This coincides with Weinbaum's original definition in the case where f is a single letter. Every letter a is a marker in w, but a first guess that a marker leads always to a W-factorization fails. The situation is more complicated.

Example 2. *Consider the word $w = abaaaba$. We observe the following.*

- *The conjugate $w' = (aaa) \cdot (baab)$ yields a* W-*factorization for a, aa, and aaa. The factors a and aaa are markers.*
- *However, aa is not a marker.*
- *The cyclic factor $f = aabaa$ is not a marker and there is no* W-*factorization of w for f, because a has no unique position in w.*
- *The factor aba is a marker, but again there is no* W-*factorization of w for aba, because neither a nor $aaba$ nor $abaa$ have a unique position in w.*

Before we continue, let us propose a stronger and more symmetric definition for two factors f and g.

Let w, f, g be words and w be a primitive. A factorization $w' = uv$ of a conjugate w' is called a W-*factorization of* w *for* f *and* g, if the following three conditions hold:

1. u and v are uniquely positioned in w,
2. $u \in (fA^* \cap A^*f) \setminus (gA^* \cup A^*g)$,
3. $v \in (gA^* \cap A^*g) \setminus (fA^* \cup A^*f)$.

Note that if uv is a W-factorization of w for f, then uv is a W-factorization of w for f and v, too.

Remark 3. *The following holds.*

$$(fA^* \cap A^*f) \setminus (gA^* \cup A^*g) \neq \emptyset \iff g \not\leq_{\mathrm{p}} f \text{ and } g \not\leq_{\mathrm{s}} f \ .$$

Example 4. *Consider the word*

$$w = bbaabacbaacbaaaca.$$

Then w *has a* W-*factorization for* $f = ab$ *and* $g = ac$*. Indeed, by shifting the suffix* a *to the beginning, we obtain:*

$$w' = abbaabacbaacbaaac = (fbaf) \cdot (gbagbaag)\,,$$

where both $fbaf$ *and* $gbagbaag$ *are uniquely positioned in* w*.*

In the following, let w be a primitive word. Let f be a proper factor of w. We define the set $G(f)$ of factors of w such that $g \in G(f)$ if and only if g is a cyclic factor of w that is preceded and followed by f in w^3, and g does not intersect with f. More precisely, we put

$$G(f) = \left\{ g \;\middle|\; |fg| \leq |w|,\ fgf \text{ is a cyclic factor of } w^2, \text{ and } fgf \notin A^+fA^+ \right\}.$$

Remark 5. *The set* $G(f)$ *can be empty even for short factors* f*. For instance, consider* $w = (aab)^k aaaba$ *with* $k \geq 2$*, and* $f = aabaa$*. Here each factor* fgf *has a third occurrence of* f*.*

We have constructive results, only if $G(f) \neq \emptyset$. In this case we are interested in maximal elements of $G(f)$. Therefore we let:

$$\max(G(f)) = \{\, g \in G(f) \mid g \text{ does not occur in any other element of } G(f) \,\}.$$

We define the subset $R(f)$ of the set of factors of w as follows:

$$R(f) = \{\, g \in \max(G(f)) \mid f \text{ and } g \text{ do not intersect in } w \,\}.$$

A word f is called a *Weinbaum factor* of w, if $R(f) \neq \emptyset$.

Note that a Weinbaum factor is not necessarily a marker and being a marker does not mean that $R(f) \neq \emptyset$. Here is such an example:

Example 6. *Let $w = abababb$. Then $f = aba$ is not a marker in w but $R(f) = \{bb\}$. However, ab is a marker in w, but $G(ab) = \{b\}$ and $R(ab) = \emptyset$.*

However, the crucial observation, stated in Lemma 8 is that every element g of $R(f)$ is a marker. Therefore an element of $R(f)$ can be called a *complementary marker*, because if $g \in R(f)$, then g is a marker of w. Moreover, f and g do not intersect in w by the very definition of $R(f)$. In particular, g is not a factor of f.

Remark 7. *For $|w| \geq 2$ and $a \in A$ a letter, g can be chosen to be a cyclic factor of maximum length between two occurrences of the letter a in w; and $g \in R(a)$ is a complementary marker. More general, let $f \in a^+$ and g be any longest cyclic factor of w with $g \notin aA^* \cup A^*a \cup A^*fA^*$, but where fgf is a cyclic factor of w^2. Then we have $g \in R(f)$; and in particular, $R(f) \neq \emptyset$*

Lemma 8. *Let f be a cyclic factor of w. Then each $g \in R(f)$ is a marker.*

Proof. We may assume that $\varepsilon \neq g \in R(f) \neq \emptyset$. Assume that g is not a marker. Then $h = xyz$ is a factor in w^2 where $xy = g = yz$ for some nonempty words x, y, and z. Since f and g do not intersect in w, neither do f and h. Hence, there exists $h' = phq \in G(f)$ (between two consecutive occurrences of f in w^2) and h is a factor of h'. This yields a contradiction, because now $g \notin \max(G(f))$. □

5 Weinbaum Factorizations by Lyndon Words

In this section we give a proof of a strengthened version of Weinbaum's Theorem [5] using Lyndon words.

Let $\trianglelefteq$ be a total order on the alphabet A. Then $\trianglelefteq$ can be extended to a *lexicographic* order on A^* by setting $u \trianglelefteq v$ if either $u \leq_{\mathrm{p}} v$ or $xa \leq_{\mathrm{p}} u$ and $xb \leq_{\mathrm{p}} v$ where $a \neq b$ and $a \trianglelefteq b$ and $x \in A^*$. A *Lyndon word* is a primitive word that is the minimum among its conjugates w.r.t. $\trianglelefteq$. Note that if w is a Lyndon word then it is *unbordered*: this means $w \notin fA^* \cap A^*f$ for every proper factor f of w. Indeed, assume the opposite and let f be a proper factor of minimum length such that $w \in fA^* \cap A^*f$. Clearly, $w = fxf$, but then $w = fxf \trianglelefteq ffx$ implies $xf \trianglelefteq fx$, and hence, $xff \trianglelefteq fxf = w$; a contradiction.

Lemma 9. *Let $w = uv$ be a Lyndon word w.r.t. a lexicographic order $\trianglelefteq$ such that v is the maximum suffix of w w.r.t. $\trianglelefteq$. Then both u and v are uniquely positioned in w. Moreover, if v' is a cyclic factor of w such that $v \trianglelefteq v'$, then $v \leq_{\mathrm{p}} v'$.*

Proof. First, observe that v is uniquely positioned. This is clear, because v is the maximum suffix and a Lyndon word is unbordered. Assume now that v' is a cyclic factor of w such that $v \trianglelefteq v'$, and let $w^2 = xv'y$ where $|x| < |w|$.

Suppose first that $w = xv'y'$. Since v is the maximum suffix, we obtain $v'y' \trianglelefteq v \trianglelefteq v'$. Hence $y' = \varepsilon$, and thus $u = x$ and $v = v'$.

In the other case we have $w = xv'_1 = v'_2y$, where $v' = v'_1v'_2$ with $v'_2 \neq \varepsilon$. Assume that $v'_1 \neq v$. If $|v| < |v'_1|$, then $v \trianglelefteq v'$ implies $v \trianglelefteq v'_1$ contradicting the maximality of v. If $|v'_1| < |v|$ then $v'_1 \trianglelefteq v \trianglelefteq v'_1v'_2 = v'$ implies that $v = v'_1v_2$ for some $v_2 \neq \varepsilon$ with $v_2 \trianglelefteq v'_2$. Thus, $v_2 \not\leq_{\mathrm{p}} v'_2$, for otherwise $w \in v_2A^* \cap A^*v_2$ would imply that w is not a Lyndon word. But now, $v_2uv'_1$ is a conjugate of w and $v_2uv'_1 \trianglelefteq v'_2y = w$ contradicts the assumption that w is a Lyndon word. Consequently, $v \trianglelefteq v'$ implies that $v \leq_{\mathrm{p}} v'$.

Finally, consider the occurrences of the prefix u. Let $w^2 = xuy$ where $0 < |x| \leq |w|$. Let $w^2 = xuv'y'$ where $|v'| = |v|$. We have $v \trianglelefteq v'$ because uv' is a conjugate of w and w is a Lyndon word. Now v' is a cyclic factor of w, and hence, by the above, $v = v'$ and v' is uniquely positioned in w. This means that $w = x$ and therefore also u is uniquely positioned in w. □

Theorem 10. *Let w be a primitive word and let f be a Weinbaum factor of w and $g \in R(f)$. Then w has a* W*-factorization for f and g.*

Proof. In this proof, we let $\bar{z}$ denote a letter corresponding to a word z. Since we consider conjugates of words, and g is a marker by Lemma 8 (or by an obvious reason in case $f = a$ and g any longest cyclic factor of w without a), we may assume that $w = gz_1gz_2 \cdots gz_k$ where $k \geq 1$ such that $z_i \in fA^* \cap A^*f$ and g is not a factor of any z_i. Let $B = \{\bar{g}, \bar{z}_i \mid i = 1, 2, \ldots, k\}$ be a new alphabet corresponding to the words g and z_i. We may assume that $x = \bar{g}\bar{z}_1\bar{g}\bar{z}_2 \cdots \bar{g}\bar{z}_k$ is a Lyndon word w.r.t. a lexicographic order $\trianglelefteq$ on B^* such that $\bar{g}$ is the minimum in B and, if z_i occurs in z_j, then $\bar{z}_i \trianglelefteq \bar{z}_j$ for all $1 \leq i, j \leq k$.

Let t is the maximum suffix of x w.r.t $\trianglelefteq$, say $x = st$. Then $s = \bar{g}\bar{z}_1 \cdots \bar{g}\bar{z}_{m-1}\bar{g}$ and $t = \bar{z}_m\bar{g} \cdots \bar{z}_{k-1}\bar{g}\bar{z}_k$, where $\bar{z}_m$ is the maximum element w.r.t. $\trianglelefteq$. By Lemma 9, the prefix s is uniquely positioned in x, and hence also the corresponding prefix $v = gz_1 \cdots gz_{m-1}g$ of w is uniquely positioned in w, since the factor g serves as a marker. Also, $v \in gA^* \cap A^*g$.

Again by Lemma 9, the word $t = \bar{z}_m\bar{g} \cdots \bar{z}_{k-1}\bar{g}\bar{z}_k$ is uniquely positioned in x, but now it is not so immediate that the position of $u = z_mg \cdots z_{k-1}gz_k$ is unique in w. The factor z_m corresponding to the maximum $\bar{z}_m$ serves as a marker, and thus there is a cyclic factor u' in w with $u' = z_mg \cdots z_{k-1}gz_\ell$ where $z_k \leq_{\mathrm{p}} z_\ell$. But then $t \trianglelefteq \bar{z}_m\bar{g} \cdots \bar{z}_{k-1}\bar{g}\bar{z}_\ell$. By Lemma 9, this implies $\bar{z}_k = \bar{z}_\ell$, and so $u = u'$ and $x = s\bar{z}_m\bar{g} \cdots \bar{z}_{k-1}\bar{g}\bar{z}_\ell$. This means $w = vu$ and u is uniquely positioned in w. Finally, we obtain from $z_m, z_k \in fA^* \cap A^*f$ that also $u \in fA^* \cap A^*f$. □

Weinbaum's original theorem, Theorem 1, a special instance of Theorem 10. Its proof is basically self-contained in this section due to the parenthesis in Theorem 10 or Remark 7.

6 An iterative construction

In this section we consider W-factorizations from a different point of view. We do not require orderings of the alphabets here. The main result of this section is Theorem 12, which shows that a W-factorization of a word w is found by iterating the relation W at most $\frac{3}{2}\log_2(n)$ many times where $n = |w|$.

Lemma 11. *Let f be a Weinbaum factor of w and $g \in R(f)$. Then we have*

1. *The marker g is a Weinbaum factor of w, i.e., $R(g) \neq \emptyset$.*
2. *Each $h \in R(g)$ is in $fA^* \cap A^*f$.*
3. *If f is a marker, then either $R(g) = \{f\}$ or $R(g) \subseteq fA^*f$.*

Proof. We know by Lemma 8 that g is a marker. By the maximality assumption, $g \in R(f)$ is not a proper factor of any $h \in G(f)$, and therefore every cyclic occurrence of g in w^2 must be preceded and followed by f.

We may assume that $w = gz_1gz_2 \cdots gz_k$ where $k \geq 1$ such that $z_i \in fA^* \cap A^*f$ and g is not a factor of any z_i. Now let h be some z_i of maximal length, then $h \in \max(G(g))$, and h and g do not intersect because f and g do not intersect. Hence $h \in R(g)$ and $R(g) \neq \emptyset$. In fact every $h \in R(g)$ is one of the z_i above, hence $h \in fA^* \cap A^*f$. But if f is a marker, then

$$\{\, h \mid h \text{ is a factor of } w \,\} \cap fA^* \cap A^*f \subseteq \{f\} \cup fA^*f.$$

Therefore, by the maximality condition, if $f \in R(g)$, then we must have $R(g) \cap fA^*f = \emptyset$ □

The basic idea in the following proof is that for every Weinbaum factor f of w and $g \in R(f)$, there exists an i such that $R^i(g) = R^{i+2}(g)$. (Here and in the following R^n means the i-th fold iteration of the relation w, therefore $R^i(g)$ is a set.) In fact, by Lemma 11 either $R^i(g) = R^{i+2}(g)$ or $R^{i+2}(g)$ contains a word at least twice the length of a word in $R^i(g)$. Therefore, we must reach a situation $R^i(g) = R^{i+2}(g)$ with $i \leq 2\log_2(n)$. However, with a little bit of effort we can be much more precise. We give an upper bound of the number of iterations in terms of Fibonacci numbers and this meets exactly the lower bound as we will see in the next section.

Recall that the sequence of Fibonacci numbers $\{F_i\}_{i\in\mathbb{N}}$ is given by the following conditions:

$$F_0 = 0, \quad F_1 = 1, \quad \text{and} \quad F_{i+1} = F_i + F_{i-1}$$

It is a well-known classical fact that Fibonacci grow exponentially fast, more precisely, we have for all $k \geq 0$:

$$F_k = \left[\frac{\Phi^k}{\sqrt{5}}\right]$$

where $\Phi = (1+\sqrt{5})/2$ is the golden ratio and $[x]$ denotes the nearest integer of x. (See any text book which says something non-trivial about Fibonacci numbers, e.g. [4].) Thus, if $F_k \leq n$, then $k \leq \lceil \log_\Phi(n) \rceil \leq \lceil \frac{3}{2} \log_2(n) \rceil$.

The following theorem gives our main result about W-factorizations.

Theorem 12. *Let w be a primitive word, and let f be a Weinbaum factor and $g \in R(f)$. Let $2\ell \geq \log_\Phi(n)$, then $R^{2\ell-1}(g) \neq \emptyset$, for every $u \in R^{2\ell-1}(g)$, the set $R(u)$ is a singleton; and for $R(u) = \{v\}$ we obtain $R(v) = \{u\}$ and a* W*-factorization of $w = uv$ for f and g.*

Proof. By Lemma 8 the word g is a marker, by Lemma 11 we have $R^i(g) \neq \emptyset$ for all $i \geq 0$.

Consider sequences $(f_0, f_1, f_2, \ldots, f_k)$ with $k \geq 2$ which satisfy the following conditions:

1. $f_0 = \varepsilon,\ f_1 = f,\ f_2 = g$,
2. $f_{i+1} \in R(f_i)$ for $1 \leq i < k$,
3. $f_{i+2} \neq f_i$ for $0 \leq i < k-1$.

Note that $(f_0, f_1, f_2) = (\varepsilon, f, g)$ is such a sequence, so they do exist.

We claim that $|f_i| \geq F_i$ for all $0 \leq i \leq k$. This is correct for $i = 0$, 1 and $i = 2$. Hence let $k \geq 3$ and consider first $i = 3$. We have $f_3 \neq f_1 = f \neq \varepsilon$, but f is a prefix of f_3, so it is a proper prefix and we obtain $|f_3| \geq 2 = F_3$. Now let $3 \leq i+1 < k$ and $|f_j| \geq F_j$ for all $0 \leq j \leq i+1$. We have to show $|f_{i+2}| \geq F_{i+2}$.

Every cyclic occurrence of f_{i+1} is followed by $f_i f_{i-1}$ and preceded by $f_{i-1} f_i$. Since $f_{i+2} \in R(f_{i+1})$ we obtain:

$$f_i \leq_\mathrm{p} f_{i+2} \leq_\mathrm{p} f_i f_{i-1} A^* \text{ and } f_i \leq_\mathrm{s} f_{i+2} \leq_\mathrm{s} A^* f_{i-1} f_i.$$

Since $i \geq 2$ the word f_i does not intersect in w neither f_i (because it is a marker) nor f_{i-1} (because $f_i \in R(f_{i-1})$). Since $f_{i+2} \neq f_i$, we obtain

$$f_{i+2} \in f_i(f_{i-1}A^* \cap A^* f_{i-1}) f_i.$$

This means by induction

$$|f_{i+2}| \geq 2\,|f_i| + |f_{i-1}| \geq 2F_i + F_{i-1} = F_{i+2}.$$

This implies $F_k \leq n$, and therefore $k \leq \lceil \log_\Phi(n) \rceil \leq \lceil \frac{3}{2} \log_2(n) \rceil$. Thus, we may assume that in the sequence $(f_0, f_1, f_2, \ldots, f_k)$ the value k is maximal, hence $R(f_{k+1}) \subseteq \{f_{k-1}\}$. Hence in fact, $R(f_{k+1}) = \{f_{k-1}\}$. This means that f_{k-1} is a marker, and every cyclic occurrence of the marker f_{k-1} is followed by the marker f_k, and every cyclic occurrence of the marker f_k is followed by the marker f_{k-1}. Thus, $w' \in (f_{k-1} f_k)^+$ for some conjugate w' of w. But w is

primitive, hence $w' = f_{k-1}f_k$. Since f_{k-1} and f_k are markers, their positions are uniquely defined. In particular, we get $R(f_{k-1}) = \{f_k\}$, too.

Now, let $2\ell \geq \log_\Phi(n)$, then every $u \in R^{2\ell-1}(g)$ is some f_{k-1} or f_k in a sequence as above. Thus, the set $R(u)$ is a singleton and for $R(u) = \{v\}$ we obtain $R(v) = \{u\}$, and a W-factorization of $w = uv$ for f and g. □

7 An example related to Fibonacci Words

The following example provides a sequence of words with a large number of iterations of w in order to find a W-factorization. This example is related to Fibonacci words.

Consider the following sequence $\{f_i\}_{i\geq 0}$ of words over the binary alphabet $\{a, b\}$:

$$f_0 = \varepsilon, \quad f_1 = a, \quad f_2 = b, \quad \text{and} \quad f_{i+1} = f_{i-1}f_{i-2}f_{i-1} \quad (i \geq 2)\,.$$

We have for example $f_3 = aa$, $f_4 = bab$, $f_5 = aabaa$, and so forth. Let

$$w_n = f_n f_{n-1}\,.$$

For instance $w_1 = a$, $w_2 = ba$, $w_3 = aab$, $w_4 = babaa$, and so on. We will show in the following that we need $\log_\Phi(|w_n|)$ many iterations of w to obtain a W-factorization of w_n for a. It is clear that

$$|f_n| = F_n \ \text{ and } \ |w_n| = |f_n f_{n-1}| = F_{n+1}\,.$$

Remark 13. *All words f_i are* palindromes*. This means they remain same when reading from right to left. This is obvious from the recursive definition.*

There is also a very close connection between the sequence $\{f_i\}_{i\geq 1}$ and the sequence $\{h_i\}_{i\geq 1}$ of Fibonacci words defined by

$$h_1 = b, \quad h_2 = a, \quad \textit{and} \ \ h_{i+1} = h_i h_{i-1} \quad (i \geq 2).$$

We have $f_{2i} = bh_{2i}^\bullet$ and $f_{2i-1} = ah_{2i-1}^\bullet$, for all $i \geq 1$, where $x^\bullet$ denotes x without its last letter.

In fact, the words f_i are known as the singular factors *of the infinite Fibonacci sequence, [6]. (Note that h_i is a prefix of h_{i+1} for $i \geq 1$, hence we can define the infinite Fibonacci sequence as the limit of the sequence $\{h_i\}_{i\geq 1}$.) The infinite Fibonacci sequence is a* Sturmian word*, so it has exactly $n + 1$ different factors of length n. Out of these $n + 1$ different factors of length n there are n conjugates of the Fibonacci word h_n and the missing one is called the* singular factor *of length n. It turns out that it is the word f_n we are considering here.*

Let us continue with some observations about $\{f_i\}_{i\geq 0}$.

Lemma 14. *The following holds for all $0 \le i \le n$ with $n \ge 2$.*

1. $f_{n-2}f_{n-3}\cdots f_0 \le_{\mathrm{p}} f_n$,
2. $f_i \le_{\mathrm{p}} f_n \iff i \equiv n \pmod 2$,
3. $|f_n| = |f_{n-2}f_{n-1}|$ *and* $f_n \ne f_{n-2}f_{n-1}$.

Proof. The proof is by induction on n in all cases. Clearly, $\varepsilon \le_{\mathrm{p}} b$ and $a \le_{\mathrm{p}} aa$ and $b \ne a$. Let $n > 2$ and assume that the claims hold for all $k < n$.

(1) We have $f_{n-4}f_{n-5}\cdots f_0 \le_{\mathrm{p}} f_{n-2}$ by the induction hypothesis, and hence, $f_{n-2}f_{n-3}f_{n-4}\cdots f_0 \le_{\mathrm{p}} f_{n-2}f_{n-3}f_{n-2} = f_n$.

(2) ($\Rightarrow$) From the definition of $\{f_i\}_{i\ge 0}$ follows immediately that $a \le_{\mathrm{p}} f_i$ if i is odd and $b \le_{\mathrm{p}} f_i$ otherwise. ($\Leftarrow$) Clearly, $f_i \le_{\mathrm{p}} f_n$ if $i = n$. Let $i < n$. Then $i \le n-2$ and $f_i \le_{\mathrm{p}} f_{n-2}f_{n-3}f_{n-2} = f_n$ by the induction hypothesis.

(3) The fact $|f_n| = |f_{n-2}f_{n-1}|$ can be easily seen from the definitions of $\{f_i\}_{i\ge 0}$ and $\{F_i\}_{i\ge 0}$. We have $f_{n-2} \ne f_{n-4}f_{n-3}$ by the induction hypothesis, and hence, $f_n = f_{n-2}f_{n-3}f_{n-2} \ne f_{n-2}f_{n-3}f_{n-4}f_{n-3} = f_{n-2}f_{n-1}$. □

The next lemma shows that every f_i in w_n with $i+1 < n$ is a marker.

Lemma 15. *f_i does not intersect itself in f_n for all $0 \le i < n$.*

Proof. We proceed by induction on n. The cases for $n \le 8$ can be easily checked.

If $n \ge i+4$, then we have that f_i has no self-intersecting occurrence in $f_{n-2} = f_{n-4}f_{n-5}f_{n-4}$ by the induction hypothesis. From

$$f_n = \underbrace{f_{n-4}f_{n-5}\overbrace{f_{n-4}}^{g}}_{f_{n-2}}f_{n-3}\underbrace{f_{n-4}f_{n-5}f_{n-4}}_{f_{n-2}}$$

it follows that a possible intersection of f_i with itself can occur only in

$$g = f_{n-4}f_{n-3}f_{n-4} = f_{n-4}\underbrace{f_{n-5}f_{n-6}f_{n-5}}_{f_{n-3}}f_{n-4}\,.$$

We have by the induction hypothesis that f_i has no self-intersecting occurrence in f_{n-3}, which means the only remaining part of f_n to consider are the prefix and suffix of g of length at least $2|f_i|$. Consider

$$h = f_{n-4}f_{n-5}f_{n-6}f_{n-7}f_{n-8} \le_{\mathrm{p}} f_{n-4}f_{n-5}f_{n-6}f_{n-5} \le_{\mathrm{p}} g$$

and, by Lemma 14(1), we have

$$h = f_{n-4}f_{n-5}f_{n-6}f_{n-7}f_{n-8} \le_{\mathrm{p}} f_{n-4}f_{n-5}f_{n-4} = f_{n-2}\,.$$

So, f_i has no self-intersecting occurrence in $h \leq_{\mathrm{p}} g$. Symmetrically, f_i has no self-intersecting occurrence in $f_{n-8}f_{n-7}f_{n-6}f_{n-5}f_{n-4} \leq_{\mathrm{s}} g$. We have $|h| \geq 2|f_{n-4}| \geq 2|f_i|$ because $|f_{n-4}| = 2|f_{n-6}| + |f_{n-7}| < |f_{n-5}f_{n-6}f_{n-7}|$ since $|f_{n-6}| < |f_{n-5}|$. That proves the case.

If $n = i + 3$, then f_{n-3} has no self-intersecting occurrence in f_{n-2} by the induction hypothesis. Hence, if f_{n-3} has a self-intersecting occurrence in f_n, then it intersects with the centered occurrence of f_{n-3} in $f_n = f_{n-2}f_{n-3}f_{n-2}$. Since by the induction hypothesis f_{n-5} has no self-intersecting occurrence in f_{n-4} we have that f_{n-3} could only intersect itself so that f_{n-5} is aligned in $f_n = f_{n-2}f_{n-5}f_{n-6}f_{n-5}f_{n-2}$, that is, either $f_{n-3} \leq_{\mathrm{p}} f_{n-5}f_{n-4} \leq_{\mathrm{p}} f_{n-5}f_{n-2}$ or, symmetrically, $f_{n-3} \leq_{\mathrm{s}} f_{n-4}f_{n-5} \leq_{\mathrm{s}} f_{n-2}f_{n-5}$ which contradicts Lemma 14(3).

If $n = i + 2$, then $f_{n-2} = f_{n-4}f_{n-5}f_{n-4}$ occurs in $f_n = f_{n-2}f_{n-3}f_{n-2} = f_{n-4}f_{n-5}f_{n-4}f_{n-3}f_{n-4}f_{n-5}f_{n-4}$ only so that f_{n-4} is aligned otherwise f_{n-4} has a self-intersecting occurrence in f_n contradicting previous case. Therefore, if f_{n-2} has a self-intersecting occurrence in f_n, then we have either $f_{n-2} \leq_{\mathrm{p}} f_{n-4}f_{n-3}$ or, symmetrically, $f_{n-2} \leq_{\mathrm{s}} f_{n-3}f_{n-4}$ which contradicts Lemma 14(3).

If $n = i + 1$, then, similarly to the previous case, $f_{n-1} = f_{n-3}f_{n-4}f_{n-3}$ occurs in $f_n = f_{n-2}f_{n-3}f_{n-2}$ only so that f_{n-3} is aligned otherwise f_{n-3} has a self-intersecting occurrence in f_n contradicting a previous case. Therefore, if f_{n-1} has a self-intersecting occurrence in f_n, then we have either $f_{n-1} \leq_{\mathrm{p}} f_{n-3}f_{n-2}$ or, symmetrically, $f_{n-1} \leq_{\mathrm{s}} f_{n-2}f_{n-3}$ which contradicts Lemma 14(3). □

The next two lemmas show that $G(f_i) = \{f_{i-1}, f_{i+1}\}$ for all $i + 1 < n$.

Lemma 16. *f_i does not occur in f_{i+1}.*

Proof. By Lemma 15, $f_i = f_{i-2}f_{i-3}f_{i-2}$ can only occur in $f_{i+1} = f_{i-1}f_{i-2}f_{i-1}$ such that f_{i-2} has no self-intersecting occurrence in f_{i+1}. Hence, we have either $f_{i-3}f_{i-2} \leq_{\mathrm{p}} f_{i-1}$ or $f_{i-2}f_{i-3} \leq_{\mathrm{s}} f_{i-1}$. Both of these cases contradict Lemma 14(3). □

Lemma 17. *If f_igf_i occurs in f_n such that f_i is not a factor of g, then we have $g \in \{\varepsilon, f_{i-1}, f_{i+1}\}$.*

Proof. We proceed by induction on n. Again the cases $n \leq 8$ can be easily checked. We have that the claim holds for all occurrences of f_igf_i in f_{n-2} and f_{n-3} by the induction hypothesis. Hence, we need to consider only those cases where f_igf_i intersects with both f_{n-2} and f_{n-3} in $f_n = f_{n-2}f_{n-3}f_{n-2}$. Assume that $i \equiv n \pmod 2$. The case with different parities is symmetric. It follows from Lemma 14(2) that $f_i \leq_{\mathrm{p}} f_{n-2}$ and $f_i \not\leq_{\mathrm{p}} f_{n-3}$ and, again by symmetry, $f_i \leq_{\mathrm{s}} f_{n-2}$ and $f_i \not\leq_{\mathrm{s}} f_{n-3}$. Consider the latter case. It follows from Lemma 15 that $gf_i \leq_{\mathrm{p}} f_{n-3}f_i$. Clearly, if $i+2 = n$ then $g = f_{n-3} = f_{i-1}$, and if $i+4 = n$ then $g = f_{n-3} = f_{i+1}$. Let $i+4 < n$. Then $f_{i+1}f_i \leq_{\mathrm{p}} f_{n-3}$ by

Lemma 14(1). The claim follows from the fact that f_i is not a factor of f_{i+1} which is shown by Lemma 15. □

Consider now w_n for some fixed n. Proposition 18 follows straightforwardly from Lemma 17.

Proposition 18. *$R(f_i) = \{f_{i+1}\}$, for all $1 \leq i < n$, and $R(f_n) = \{f_{n-1}\}$.*

We have $w_n = R^{n-1}(a)R^{n-2}(a)$. From $|w_n| = F_n = \left[\frac{\Phi^n}{\sqrt{5}}\right]$ follows that we need $\lceil \log_\Phi |w_n| \rceil$ many steps to reach a W-factorization of w_n for a.

8 Weinbaum Factorizations for random words

Consider alphabets of size at least two in the following. By definition,W-factorizations for given words f and g can exist only if $(fA^* \cap A^*f) \setminus (gA^* \cup A^*g) \neq \emptyset$ and $(gA^* \cap A^*g) \setminus (fA^* \cup A^*f) \neq \emptyset$. In the following we call a pair (f, g) satisfying these conditions a *Weinbaum candidate* for short. For all Weinbaum candidates there exist W-factorizations. In fact this is not a rare event at all, they exist in all long enough random words. This means it is a generic property. It is the gist of the next proposition.

Proposition 19. *Let $k, m \geq 1$ be constants and $\Pr_{k,m}(n)$ be the probability that a word w of length n (under the uniform distribution) is primitive and that w admits at least k different* W*-factorizations for all Weinbaum candidates (f, g) with $|fg| \leq m$. Then we have:*

$$1 - \Pr_{k,m}(n) \in 2^{-\Omega(n)}.$$

Proof. The proof relies on standard arguments as used e.g. in Kolmogorov complexity frequently. Therefore we give a sketch of the proof, only. We refer to [3] for details on Kolmogorov complexity.

In order to prove the claim it is enough to consider random words. More precisely, it suffices to show that almost all words w which are either not primitive or which have less than k different W-factorizations for some Weinbaum candidate (f, g) with $|fg| \leq m$ can be compressed by a fixed linear factor. The compression rate depends on m, k, and the alphabet A, only. So, it is independent of n and this will give the result.

A *compression* is here simply an injective function $\gamma : A^* \to A^*$. A subset $X \subseteq A^*$ is compressible (by γ) with compression rate $\epsilon > 0$, if $|\gamma(w)| < (1 - \epsilon)|w|$ for almost all words w in X. Clearly, for each compression γ and $\epsilon > 0$, the probability that a word w of length n belongs to X is in $2^{-\Omega(n)}$. Let w be a word of length n where $n > n(k, m, A)$ is large enough and assume that w is not compressible (by some fixed compression γ) by an ϵ-factor where $0 < \epsilon < \epsilon(k, m, A)$ is small. (The description of the compression γ and

possible values for $n(k,m,A)$ and $\epsilon(k,m,A)$ can be derived from the following considerations, we omit details.) Then w must be primitive, otherwise w were highly compressible. Write $w = w_1 \cdots w_{k+4}$ where each w_i has length at least $\frac{n}{k+4} - 1$. We may assume that this value is still huge since k is a constant. The position of each cyclic factor w_i must be unique in w, because otherwise a (say Lempel–Ziv like) encoding would lead to a compression by a linear factor. Note that this implies that the position of a cyclic factor u in w is unique, as soon as any w_i is a factor of u. Now we claim that all words v with $|v| = m$ are factors of all factors w_i. Indeed, by contradiction assume that some word v does not appear in some w_i. Then in a block code of length m not all letters of this block code are necessary to code w_i. This knowledge can be used to compress w_i, and hence w (because k is a constant) by a linear factor.

Now it is easy to exhibit at least k different W-factorizations for all Weinbaum candidates (f,g) with $|fg| \leq m$. For each candidate (f,g) we choose a factor fg in w_1 and for each $i = 3, \ldots, k+3$ we choose a factor gf in w_i. This is possible since all words of length m are factors of each w_i. This leads to k different conjugated words $u_i v_i$ of w with $v_i \in gA^*w_2A^*g$ and $u_i \in fA^*w_{k+4}A^*f$ such that each position of u_i and v_i is unique. Moreover, $(fA^* \cap A^*f) \setminus (gA^* \cup A^*g) \neq \emptyset$ and $(gA^* \cap A^*g) \setminus (fA^* \cup A^*f) \neq \emptyset$ (by definition of a Weinbaum candidate) implies that f (g resp.) is neither prefix nor suffix of g (f resp.). Therefore $u_i \notin gA^* \cup A^*g$ and $v_i \notin fA^* \cup A^*f$. □

9 Not all factors allow always Weinbaum Factorizations

Weinbaum's result, Theorem 1, states that every primitive word w admits a W-factorization for all letters a that occur in w. Moreover, we have seen that this is true for all factors of the form a^m with $m \geq 1$. Let us show that this is the best we can expect. For this we investigate some few cases for which a factor f of a primitive word w does or does not admit a W-factorization.

Proposition 20. *Let $f \in A^+$, $a \in A$. Let m be the maximum exponent such that a^m occurs in f. Then $w = fa^n$ admits a* W*-factorization for f if and only if both, $f \notin aA^* \cup A^*a$ and $n > m$.*

Proof. Clearly, if $f \notin aA^* \cup A^*a$ and $n > m$, then $w = (f) \cdot (a^n)$ is a W-factorization for f. For the converse, since w must be primitive and f must be a proper factor of w, there is a letter $b \neq a$ occurring in f. But then there is only one cyclic position for f in w and $w = (f) \cdot (a^n)$ must be the W-factorization for f. Since a^n has a unique position, we see $f \notin aA^* \cup A^*a$ and $n > m$. □

It is well-known that the following proposition follows from the Theorem of Fine and Wilf; see for example [1].

Proposition 21. *Let $f \in A^*$ be a word. Then there is at most one letter $a \in A$ such that fa is not primitive.*

Corollary 22. *Let $f \in A^*$ be a word where pairwise different letters $a_1, \ldots, a_k$ occur for $k \geq 2$. Then at least $k-1$ of the words fa_i are primitive, but none of them admits a* W*-factorization for f.*

10 Conclusion

The original idea to this paper has been modest, just provide a simple proof of Weinbaum's result, Theorem 1. But when playing around with the result we discovered some nice combinatorics on words which seems to be unexplored so far. We have however absolutely no application at all for our investigation which go beyond the original statement of Weinbaum.

We did not discuss algorithmic issues. How expensive is it to compute a W-factorization of w for f, if it exists? The reason for our silence is simple. We do not have any non-trivial result here. By a further exploration it seems however possible that some clever use of *stringology* might lead to fast algorithms.

Acknowledgement

We thank Jean Berstel and Julian Cassaigne for pointing out that the words f_i in Section 7 are the singular factors of the infinite Fibonacci sequence.

References

[1] T. Harju, V. Halava, and L. Ilie. Periods and binary words. *J. Combin. Theory Ser. A*, 89:298–303, 2000.

[2] M.A. Harrison. *Introduction to formal language theory.* Addison-Wesley Publishing Co., Reading, Mass., 1978.

[3] M. Li and P. Vitányi. *An introduction to Komolgorov complexity and its applications.* Texts and Monographs in Computer Science. Springer-Verlag, New York, 1993.

[4] M. Lothaire. *Combinatorics on Words*, volume 17 of *Encyclopedia of Mathematics.* Addison-Wesley, Reading, MA, 1983.

[5] C.M. Weinbaum. Unique subwords in nonperiodic words. *Proc. Amer. Math. Soc.*, 109(3):615–619, 1990.

[6] Zh.X. Wen and Zh.Y. Wen. Some properties of the singular words of the Fibonacci word. *European J. Combin.*, 15(6):587–598, 1994.

Turku
Centre *for*
Computer
Science

Lemminkäisenkatu 14 A, 20520 Turku, Finland | www.tucs.fi

University of Turku
- Department of Information Technology
- Department of Mathematics

Åbo Akademi University
- Department of Computer Science
- Institute for Advanced Management Systems Research

Turku School of Economics and Business Administration
- Institute of Information Systems Sciences

ISBN 952-12-1774-X
ISSN 1239-1891