An Extension of the Periodicity Lemma to Longer Periods

Aviezri S. Fraenkel
Department of Computer Science and Applied Mathematics
The Weizmann Institute of Science
Rehovot 76100, Israel
fraenkel@wisdom.weizmann.ac.il
http://www.wisdom.weizmann.ac.il/~fraenkel

Jamie Simpson
Department of Mathematics and Statistics, Curtin University
Perth WA 6001, Australia
simpson@cs.curtin.edu.au
http://www.cs.curtin.edu.au/~simpson

February 15, 2002

Abstract

The well-known periodicity lemma of Fine and Wilf states that if the word x of length n has periods p, q satisfying $p+q-d \leq n$, then x has also period d, where $d = \gcd(p,q)$. Here we study the structure of *long* periods, namely $p+q-d > n$. In particular, we construct recursively a sequence of integers $p = p_1 > p_2 > \ldots > p_{t-1} \geq 2$, and show that such x is covered with periods p_i by a basic *kernel* factor. We further compute the maximum alphabet size $|A| = p+q-n$ of A over which a word with long periods can exist, and compute the subword complexity of x over this A.

1 Introduction

We consider finite words over a finite alphabet, not necessarily binary. If x is such a word and p is an integer in $\{1, \ldots, |x|-1\}$, we say that x has *period* p, if $x[i] = x[i+p]$ for all $i \in \{1, \ldots, |x|-p\}$. Here and below, $x[i]$ denotes the i-th letter of x, beginning with $x[1]$ from the left.

Definition 1. Let x be a word of length $|x| = n$ having integer periods p and q. Throughout we put $d = \gcd(p,q)$, $k = \lfloor q/p \rfloor$, and assume

$$0 < p < q < n. \tag{1}$$

In particular, we disregard the period of length n of x. If

$$p + q - d > n, \tag{2}$$

then we say that (p, q) is a *long pair of periods* of x. If p, q, d satisfy $p+q-d \leq n$, then (p, q) is said to be a *short pair of periods* of x.

The following *Periodicity Lemma* of Fine and Wilf [5] applies to words with a short pair of periods.[1]

Lemma 1. *If x has a short pair (p, q) of periods, then x has also period d.*

Notes.

- There are infinite families of words showing that the bound in the lemma is tight, i.e., if $p + q - d = |x| + 1$, then x doesn't necessarily have period d. See e.g., [3], § 8.1.2.
- Many words have both short and long pairs of periods.

Example 1. The periods $< n = 9$ of $x_1 = 110111011$ are $4, 7, 8$. Of these, $(p, q) = (4, 8)$ is a short pair (Lemma 1 gives no new information for this case); and $(4, 7)$ and $(7, 8)$ are long pairs. The pair $(4, 7)$ illustrates the tightness of the bound of Lemma 1, since $4 + 7 - 1 = 10 = n + 1$, and x_1 doesn't have periodicity gcd(7,8)=1.

Can we say anything about the structure of a word x in terms of its long pairs of periods? We are not aware that this question has been addressed previously. Perhaps the fact that the Fine-Wilf Lemma is best possible has discouraged attempts in this direction. In this connection we point out, however, that in view of the natural requirement (1), a long pair of periods isn't all that long compared to a short pair of periods.

We show that indeed an interesting structure hides under the wings of a long pair of periods. Roughly speaking, our discussion involves an inductive argument relating to two types of long pairs of periods which we call Type I and Type II. We'll see that if x has Type II pairs of periods, then it can be described in terms of a prefix of x with long pairs of periods, and that prefix, possibly, in terms of a still shorter prefix of x with long pairs of periods. The process continues while we deal with prefixes with Type II pairs of periods, and terminates when we get a prefix with a Type I pair of periods, also called the *kernel* of x.

We further compute the maximum alphabet size $|A| = p + q - n$ of A over which a word with long periods can exist, and compute the subword complexity of x over this A.

[1] The usual statement of Lemma 1 permits $q = n$ in (1). But then $d = p$, for which case the result is trivial.

2 Structure of Words with Long Pairs of Periods

For a word x of length n with a long pair (p,q) of periods with $p<q$, put

$$p_1 = p,\ q_1 = q,\ k_1 = \lfloor q_1/p_1 \rfloor,\ d_1 = d,\ n_1 = n,\ x_1 = x. \tag{3}$$

For $i \geq 1$ apply the Euclidean Algorithm for creating a simple continued fraction expansion of q_1/p_1, to construct recursively the following items,

$$\begin{aligned} q_i = k_i p_i + p_{i+1},\ 0 \leq p_{i+1} < p_i,\ \text{so } k_i = \lfloor q_i/p_i \rfloor;\ \text{also put } q_{i+1} = p_i, \\ d_{i+1} = \gcd(p_{i+1}, q_{i+1}),\ n_i = k_i p_i + n_{i+1},\ x_{i+1} = x_1[1 \ldots n_{i+1}]. \end{aligned} \tag{4}$$

Definition 2. For $i \geq 1$, the long pair of periods (p_i, q_i) of the word x_i of length n_i is Type I if

$$n_i \leq (k_i+1)p_i, \quad \text{which is equivalent to} \quad n_{i+1} \leq q_{i+1}, \tag{5}$$

and Type II if

$$n_i > (k_i+1)p_i, \quad \text{which is equivalent to} \quad n_{i+1} > q_{i+1}. \tag{6}$$

Example 2. For Example 1, the pair of periods $(p_1, q_1) = (4,7)$ is Type II, and $(7,8)$ is Type I.

Lemma 2. *Let (p_1, q_1) with $p_1 < q_1$ be a long pair of periods of a word x_1 of length n_1. Let $i \geq 1$.*
(i) *We have $d_i = d_1$ for all i. If p_i, q_i, n_i satisfy* (1) *and* (2), *then $p_i \not| q_i$ (p_i does not divide q_i), $p_i \geq 3$, and*

$$q_i - p_i + 1 \leq k_i p_i \leq q_i - 1. \tag{7}$$

(ii) *$p_i + q_i - n_i$ is a constant for all i.*
(iii) *If (p_i, q_i) is a Type* II *pair of periods of x_i, then the pair (p_{i+1}, q_{i+1}) and n_{i+1} satisfy* (1) *and* (2).
(iv) *There exists a smallest $t \in \mathbb{Z}_{>0}$ for which $n_t \leq (k_t+1)p_t$, i.e., (p_t, q_t) is a Type* I *pair of periods of x_t.*
(v) *For all $i \in [1, t+1]$ we have, $n_i > n_{i+1}$, $p_i > p_{i+1}$, $q_i > q_{i+1}$ (t as in* (iv)).
(vi) *We have $1 \leq p_{t+1} < n_{t+1} \leq q_{t+1}$ (t as in* (iv)).

Proof. (i) By (4), $d_{i+1} = \gcd(p_{i+1}, q_{i+1}) = \gcd(q_i - k_i p_i, p_i) = \gcd(p_i, q_i)$. This implies $d_{i+1} = d_1$ for all $i \geq 1$. We have (1) and (2) $\Longrightarrow q_i < n_i < p_i + q_i - d_i \Longrightarrow d_i < p_i \Longrightarrow p_i \not| q_i$. Then (7) follows from the fact that q_i/p_i is not an integer. Also $p_i > 1$. If $p_i = 2$, then $d_i = 1$, so (1) and (2) imply $q_i < n_i < q_i + 1$, a contradiction. Hence $p_i \geq 3$.

(ii) By (3), (4), $p_{i+1} + q_{i+1} - n_{i+1} = p_i + q_i - n_i$.

(iii) From (4) and (6), $n_{i+1} = n_i - k_i p_i > p_i = q_{i+1}$, which is one part of (1). The hypothesis of (iii) implies that p_i, q_i satisfy (1) and (2), so (7) holds with i replaced by $i+1$. Hence $2 \leq p_{i+1} = q_i - k_i p_i < p_i = q_{i+1}$, which are the other two parts of (1).

(iv) By (4), $0 < n_{i+1} < n_i$; and $(k_i+1)p_i \geq 2$ for all i. The well-ordering principle then implies that t with the desired property exists.

(v) Follows directly from (4).

(vi) The right side is the right side of (5). For verifying the middle inequality, we use (4), (1) and (iii) to write $p_{t+1} = q_t - k_t p_t < n_t - k_t p_t = n_{t+1}$. Also $p_{t+1} = q_t - k_t p_t \geq 1$ by (7). ■

Lemma 3. *Definitions* 1 *and* 2 *are consistent.*

Proof. For a Type I pair of periods, (2) and (5) do not imply one another in either direction. For a Type II pair of periods, (2) and (6) imply $kp \leq q-d-2 \Longrightarrow q+1-p \leq q-d-2$ (by(7)) $\Longrightarrow d \leq p-3$. Now $d = p-2 \Longrightarrow d|2 \Longrightarrow d \in \{1,2\} \Longrightarrow p \in \{3,4\}$. Similarly, $d = p-1 \Longrightarrow p = 2$. Also $p = 1$ is excluded since $p \not| q$. Hence for a Type II pair of periods we need $p \geq 5$ or $p = 4$, $d = 1$. We show now that this requirement is satisfied "automatically".

Let there be a word with a long pair (p,q) of periods, $p = 4 < q$ and $d = 2$. Then $q = 4t+2$, $t \in \mathbb{Z}_{>0}$. By (1), (2), $q < n < q+2$. Thus $n = q+1$. Hence $n = 4t+3 < (\lfloor q/4 \rfloor + 1)4 = 4t+4$, so (p,q) is a Type I pair of periods. We can see similarly that $p \in \{2,3\}$ (whence $d = 3$) implies that (p,q) is a pair of Type I periods.

Note that (1) and (6) do not imply one another in either direction.

In conclusion, Definition 1 is consistent with a word with a long pair of periods to be either Type I or Type II; and Type I or Type II pairs of periods are consistent with Definition 1. ■

Lemma 4. *Let x_1 be a word of length n_1 with a pair (p,q) of long periods satisfying $q_1 > p_1$. Then for $g_1 \in \{0,\ldots,k_1\}$ we have,*

$$x_1[i] = x_1[i + g_1 p_1] \tag{8}$$

for $i = 1,\ldots,n_2$. Moreover, the prefix x_2 of x_1 has period p_2 (where k_1, x_2, p_2 are given by (3), (4)).

Proof. Since $i + g_1 p_1 \leq n_2 + k_1 p_1 = n_1$ and x_1 has period p_1, (8) follows. For verifying the second statement, we first note that (1) and the right hand side of (7) imply the inequalities:

$$n_2 = n_1 - k_1 p_1 \geq 2 \tag{9}$$

and $p_2 = q_1 - k_1 p_1 \geq 1$. Moreover, $p_2 < n_2$ by Lemma 2(iii) and (vi). We have to show: $x_1[i] = x_1[i+p_2]$ for $i = 1,\ldots,n_2 - p_2 = n_1 - q_1$. Indeed, since x_1 has period q_1 we have for $i \in [1, n_1 - q_1]$,

$$x_1[i] = x_1[i+q_1] = x[i + q_1 - k_1 p_1] = x[i + p_2],$$

where the second equality follows from the p_1-periodicity of x_1. ■

Corollary 1. *If x_1 has a Type* I *pair (p_1, q_1) of periods with $p_1 < q_1$, then $x_1 = (x_2 z)^{k_1} x_2$, where the border x_2 has period p_2, and $z = x_1[n_2 + 1 \ldots p_1]$.*

Proof. Inequality (9) implies $|x_2| > 0$. Evidently, $|x_2 z| = p_1$. The result of the first part now follows from the structure of the word (8), which has period p_1. Also x_2 has period p_2 by the second part of Lemma 4. Further, $|z| = p_1 - n_2 = (k_1 + 1)p_1 - n_1 \geq 0$ by (5). ■

Example 3. Let $x_1 = 1101110111$, with $n_1 = 10$. The periods < 10 are $4, 8, 9$, where $(4, 8)$ is a short pair; $(4, 9)$, $(8, 9)$ are long pairs, of Type I. Corollary 1 for $(4, 9)$ (with $k = 2$, $n_2 = 2$) states that $x_2 = 11$, $z = 01$, so $x_1 = (1101)^2 11$, where x_2 has period $p_2 = 1$.

We give two more examples for illustrating Corollary 1.

Example 4. $x_1 = 11011101$ ($n_1 = 8$) has a Type I pair of periods ($p_1 = 4, q_1 = 7$), so $k_1 = 1$, $p_2 = 3$, $n_2 = 4$ and z is empty. Thus $x_1 = (1101)^1 1101$.

Example 5. $x_1 = 101101110110$ ($n_1 = 12$) has a Type I pair of periods ($p_1 = 7, q_1 = 10$), so $k_1 = 1$, $p_2 = 3$, $n_2 = 5$, $x_2 = 10110$ and $z = 11$. Thus, $x_1 = ((10110)(11))^1 10110$.

Lemma 2(iii) suggests that there is an iterative procedure for expressing the structure of x_1 in terms of the structure of its prefixes x_i. The recursion terminates when the smallest t is reached for which $n_t \leq (k_t + 1)p_t$. That is, the recursion terminates when the triple (p_t, q_t, n_t) corresponds for the first time to a pair of Type I periods.

Lemma 5. *Let $s \in \{1, \ldots, t-1\}$, where t is as in Lemma* 2(iv)*, $t \geq 2$. For $g_\ell = 0, \ldots, k_\ell$ $(1 \leq \ell < t)$, put $E_s = g_s p_s + \ldots + g_{t-1} p_{t-1}$. Then for $i = 1, \ldots, n_t$, $E_s + i$ assumes, at least once, every value in the integer interval* $[1, n_s]$.

Proof. Descent on s. For $s = t-1$, $E_{t-1} + i = i + g_{t-1} p_{t-1}$. For fixed values of $g_{t-1} \in \{0, \ldots, k_{t-1}\}$, we let i range over $[1, n_t]$. This produces the intervals $I_0 = [1, n_t], I_1 = [1 + p_{t-1}, n_t + p_{t-1}], \ldots, I_{k_{t-1}} = [1 + k_{t-1} p_{t-1}, n_t + k_{t-1} p_{t-1}]$. Now I_0 begins with 1. Since $n_t + k_{t-1} p_{t-1} = n_{t-1}$, $I_{k_{t-1}}$ ends with n_{t-1}. Moreover, every two consecutive intervals overlap: I_j ends with $n_t + j p_{t-1}$, and I_{j+1} begins with $1 + (j+1) p_{t-1}$ and we have $1 + (j+1) p_{t-1} \leq n_t + j p_{t-1} = n_{t-1} - k_{t-1} p_{t-1} + t p_{j-1}$ by (6). Thus the union of these intervals is $[1, n_{t-1}]$, as required.

Suppose that we have already shown that for $s > 1$, $E_s + i$ assumes all the values in $[1, n_s]$. Now $E_{s-1} + i = i + g_{s-1} p_{s-1} + \ldots + g_{t-1} p_{t-1} = i + g_{s-1} p_{s-1} + E_s$.

For fixed values of $g_{s-1} \in \{0, \ldots, k_{s-1}\}$, we let i range again over $[1, n_t]$. This produces the intervals $[1+E_s, n_t+E_s], [1+p_{s-1}+E_s, n_t+p_{s-1}+E_s], \ldots, [1+k_{s-1}p_{s-1}+E_s, n_t+k_{s-1}p_{s-1}+E_s]$. Let I denote the union of these intervals. By the induction hypothesis,

$$I = [1, n_s] \cup [1+p_{s-1}, n_s+p_{s-1}] \cup \ldots \cup [1+k_{s-1}p_{s-1}, n_s+k_{s-1}p_{s-1}].$$

The first of the intervals begins with 1. Since $n_s + k_{s-1}p_{s-1} = n_{s-1}$, the last ends with n_{s-1}. Moreover, every two consecutive intervals overlap: we have $1+(j+1)p_{s-1} \le n_s + jp_{s-1}$ by (6), proving the assertion. ■

Corollary 2. (i) $n_t + k_1p_1 + \ldots + k_{t-1}p_{t-1} = n_1$,

(ii) $n_t + k_2p_2 + \ldots + k_{t-1}p_{t-1} = n_2$,

(iii) *For* $i \in [1, n_1]$, $g_\ell \in \{0, \ldots, k_\ell\}$ $(1 \le \ell < t)$, $E_1 + i = i + g_1p_1 + \ldots + g_{t-1}p_{t-1}$ *assumes all the values in* $[1, n_1]$.

Proof. By (3), (4), $n_t + k_1p_1 + \ldots + k_{t-1}p_{t-1} = n_{t-1} + k_1p_1 + \ldots + k_{t-2}p_{t-2} = \ldots = n_2 + k_1p_1 = n_1$, proving the first identity. The second identity is proved similarly (or by substituting $n_1 = n_2 + k_1p_1$ in the first). The third part is the case $s = 1$ of Lemma 5. ■

The following is our main result.

Theorem 1. *Let x_1 be a word with a Type* II *pair (p_1, q_1) of periods satisfying $p_1 < q_1$; t as defined in Lemma* 2(iv). *Then the prefix $x_t = x_1[1 \ldots n_t]$ of x_1 is a word with a Type* I *pair (p_t, q_t) of periods satisfying $1 < p_t < q_t < n_t$. Moreover, for $i = 1, \ldots, n_t$ we have,*

$$x_1[i] = x_1[i + g_1p_1 + \ldots + g_{t-1}p_{t-1}], \tag{10}$$

for all choices of $g_\ell \in \{0, \ldots, k_\ell\}$, $\ell = 1, \ldots, t-1$. In addition, the prefixes $x_\ell = x_1[1 \ldots n_\ell]$ have periods p_ℓ, namely, $x_\ell[i] = x_1[i+p_\ell]$ for all $i \in \{1, \ldots, n_\ell - p_\ell\}$ and all $\ell \in \{1, \ldots, t+1\}$.

Proof. Note that $i + g_1p_1 + \ldots + g_{t-1}p_{t-1} \le n_t + k_1p_1 + \ldots + k_{t-1}p_{t-1} = n_1$ by Corollary 2(i), so the middle expression of (10) is well-defined. The right-hand expression follows from the middle expression since $q_{i+1} = p_i$ for $i \ge 1$.

Since x_1 has a Type II pair (p_1, q_1) of periods with $q_1 > p_1$, Lemma 2(iii) implies that x_2 has a long pair (p_2, q_2) of periods satisfying $n_2 > q_2 > p_2 > 0$. We proceed by induction on t. For $t = 2$ (i.e., (p_2, q_2) is a Type I pair of periods of x_2), (8) is (10) and we are done.

For $t > 2$ we may apply the induction hypothesis to x_2, to conclude that for $i = 1, \ldots, n_2$, $x_2[i] = x_2[i + g_2p_2 + \ldots + g_{t-1}p_{t-1}]$. We have $i + g_2p_2 + \ldots + g_{t-1}p_{t-1} \le n_t + k_2p_2 + \ldots + k_{t-1}p_{t-1} = n_2$ by Corollary 2(ii), so the equation is well-defined. The prefixes $x_2[i]$ are clearly identical to the prefixes of $x_1[i]$ for $i = 1, \ldots, n_t$, and $n_t < n_2$. Hence for $i = 1, \ldots, n_t$,

$$x_1[i] = x_1[i + g_2p_2 + \ldots + g_{t-1}p_{t-1}]. \tag{11}$$

Since x_1 has period p_1, we can add g_1p_1 to i on both sides of (11), to get, for $i = 1, \ldots, n_t$,

$$x_1[i + g_1p_1] = x_1[i + g_1p_1 + g_2p_2 + \ldots + g_{t-1}p_{t-1}].$$

Relation (10) now follows from this and from (8), since (8) holds a fortiori for $i = 1, \ldots, n_t$.

Finally, the last sentence of the theorem follows from Lemma 2(iii) and Lemma 4. ■

Definition 3. Let t be as defined in Lemma 2(iv), $t \geq 1$. The prefix $x_t = x_1[1 \ldots n_t]$ of x_1 is called the *kernel* of x_1.

Note. Theorem 1 asserts that a word x_1 with a Type II pair (p_1, q_1) of periods with $p_1 < q_1$ is covered, in general, by many, possibly overlapping, copies of the kernel, in a periodic fashion. In particular, there are $\prod_{i=1}^{t-1}(k_i + 1)$ locations L in x_1, not necessarily distinct, at each of which a copy of the kernel begins. The value t, which determines the *number* of locations L, depends on the relative sizes of p_1, q_1, n_1. We point out that x_1 is covered by the kernel *quasiperiodically* in the sense of [2]. See also [6], [7].

Example 6. Let $x_1 = 1101110110111011$ of length $n_1 = 16$. It has periods $7, 11, 14, 15$. Note that $(7, 11)$ is a Type II pair. Then $(n_1, p_1, q_1, k_1) = (16, 7, 11, 1)$. We have $(n_2, p_2, q_2, k_2) = (9, 4, 7, 1)$, with $(4, 7)$ being a Type II pair. Then $(n_3, p_3, q_3, k_3) = (5, 3, 4, 1)$, where $(3, 4)$ is a Type I pair. Thus $t = 3$, so by Theorem 1, $x_1[i] = x_1[i + 7g_1 + 4g_2]$, $g_1, g_2 \in \{0, 1\}$ for $i = 1, \ldots, 5$. Also, $(n_4, p_4, q_4, k_4) = (2, 1, 3, 3)$, and so by the last part of Theorem 1, $x_\ell[i] = x_1[i + p_\ell]$ for all $i \in \{1, \ldots, n_\ell - p_\ell\}$ and all $\ell \in \{1, \ldots, t + 1 = 4\}$

3 Maximum Subword Complexity

A word x of length n with a long pair (p, q) of periods can exist over various alphabets A. In this section we determine, for any given x with a long pair of periods, the maximum alphabet size $|A|$ such that x exists over A and every letter of A appears in x. An alphabet A which is maximum in this sense and every letter of A does appear in x will be called a *proper alphabet with respect to p, q, n*; *proper alphabet* for short. In Theorem 2 we determine the size of a proper alphabet, and in Theorem 3, the subword complexity of x over a proper alphabet with respect to (p, q).

Theorem 2. *The size of a proper alphabet for a word x_1 of length n_1 with a long pair (p_1, q_1) of periods satisfying $p_1 < q_1$, is $p_1 + q_1 - n_1$.*

Proof. We first deal with the prefix x_t where t is as defined in Lemma 2(iv), and then with the entire word x_1.

(I) Given p_t, q_t, n_t satisfying (1), (2) and (5), we construct x_t of size n_t with a Type I pair (p_1, q_1) of periods with $p_t < q_t$ over the alphabet $A = \{a_1, \ldots, a_{p_t+q_t-n_t}\}$, where the a_i are its distinct letters. Specifically, we put $x_1[1 \ldots p_{t+1}] = a_1 \ldots a_{p_{t+1}}$ which is possible since by (5), $p_{t+1} \leq p_t + q_t - n_t$. By Lemma 2(vi), $p_{t+1} < n_{t+1}$. Put $x_1[p_{t+1} + i] = x_1[i]$ for $i \leq n_{t+1} - p_{t+1}$ ($= n_t - q_t$). Then $x_{t+1} = x_1[1 \ldots n_{t+1}]$ is periodic with period p_{t+1}, consistent with Corollary 1 (with 1 replaced by t). Set $x_1[n_{t+1} + 1 \ldots p_t] = z = a_{p_{t+1}} + 1 \ldots a_{p_t+q_t-n_t}$.

Let $x_t = (x_{t+1}z)^{k_t} x_{t+1}$. Since $|x_{t+1}z| = p_t$, we have $|x_t| = k_t p_t + n_{t+1} = n_t$. Moreover, x_t is periodic with period p_t. To show that x_t is a word with a Type I pair (p_t, q_t) of periods with $p_t < q_t$, it suffices to show that it has period q_t. By the p_{t+1}-periodicity of x_{t+1}, $x_1[i] = x_1[i + p_{t+1}]$ for $i \in [1, n_{t+1} - p_{t+1}] = [1, n_t - q_t]$. Also $x_1[i + p_{t+1}] = x_1[i + q_t - k_t p_t] = x_1[i + q_t]$ by the p_t-periodicity of x_t. Thus $x_1[i] = x_1[i + q_t]$ for $i \in [1, n_t - q_t]$.

The alphabet A has size $p_t + q_t - n_t = p_1 + q_1 - n_1$ by Lemma 2(ii). In view of the p_{t+1}-periodicity of x_{t+1} and the p_t-periodicity of x_t, A cannot be any larger. It is, in fact, proper.

(II) Given $t \geq 2$, $p_i < q_i, n_i$ for $1 \leq i \leq t$, we construct a word x_1 with a Type II pair (p_i, q_i) of periods satisfying $p_i < q_i$ $(1 \leq i < t)$ and a Type I pair (p_t, q_t) of periods. For $1 \leq i \leq t$, we compute k_i by (4). Construct the first prefix $x_t = x_1[1 \ldots n_t]$ of x_1 over the same alphabet A as in (I), using the same construction. Longer prefixes $x_i = [1 \ldots n_i]$ are constructed iteratively by descent. Suppose that for some $\ell \in \{2, \ldots, t\}$ we have already constructed $x_\ell = x_1[1 \ldots n_\ell]$ with a long pair of periods (p_ℓ, q_ℓ) with $p_\ell < q_\ell$. To construct the longer $x_{\ell-1}$ with periods $p_{\ell-1} < q_{\ell-1}$, put $x_{\ell-1}[i] = x_1[i]$ for $1 \leq i \leq n_\ell$ and $x_{\ell-1}[i] = x_1[i + p_{\ell-1}]$ for $i = n_\ell - p_{\ell-1} + 1, \ldots, n_{\ell-1} - p_{\ell-1}$. (Note that $n_\ell - p_{\ell-1} > 0$ by (6)). Since the prefix x_ℓ of $x_{\ell-1}$ has period $q_\ell = p_{\ell-1}$ by the induction hypothesis, the entire factor $x_{\ell-1}$ has also period $p_{\ell-1}$. We now show that it has also period $q_{\ell-1}$.

Let $i \in [1, n_{\ell-1} - q_{\ell-1}]$. Since $x_{\ell-1}$ has period $p_{\ell-1}$, we have $x_1[i + q_{\ell-1}] = x_1[i + q_{\ell-1} - k_{\ell-1} p_{\ell-1}] = x_1[i + p_\ell]$. Now $i + p_\ell \leq n_{\ell-1} - q_{\ell-1} + p_\ell = n_\ell$, so $x_1[i + p_\ell]$ lies in $x_\ell = x_1[1 \ldots n_\ell]$ for $i \in [1, n_{\ell-1} - q_{\ell-1}]$. Since x_ℓ has period p_ℓ by the induction hypothesis, we get $x_1[i + q_{\ell-1}] = x_1[i]$, so $x_{\ell-1}$ has periods $p_{\ell-1} < q_{\ell-1}$ as required.

Our construction in (II) produced a word over the alphabet A. If A is not proper, then x_1 with the parameters $t \geq 2$, $p_i < q_i, n_i$ for $1 \leq i \leq t$, exists over a proper alphabet B with $|B| > |A|$. Let $b \in B \setminus A$, $L(b)$ the location of b in $[1, n_1]$. By Lemma 5, $L(b) = i + g_1 p_1 + \ldots + g_{t-1} p_{t-1}$ for some values $i \in \{1, \ldots, n_t\}$, $g_\ell \in \{0, \ldots, k_\ell\}$ $(1 \leq \ell < t)$. This is equivalent to $b = x_1[i + g_1 p_1 + \ldots + g_{t-1} p_{t-1}]$. Then by Theorem 1, $b = x_1[i]$. Thus b appears in $x_t = x_1[1 \ldots n_t]$, contradicting the fact, established in part (I), that A is proper for x_t. This shows that $|B| = |A|$, so A is proper for x_1. ■

Corollary 3. *Given any integers p_1, q_1, n_1 satisfying* (1), (2). *Then there exists a word x_1 with a long pair (p_1, q_1) of periods over an alphabet A if and only $|A| \in [1, p_1 + q_1 - n_1]$.*

Proof. By the *proof* of Theorem 2, x_1 of length n_1 with a long pair (p_1, q_1) of periods cannot exist over an alphabet A of size $|A| > p_1 + q_1 - n_1$, since it would upset the periodicity in x_t $(t \geq 1)$. By the *statement* of Theorem 2, there is a word x_1 with a long pair of periods (p_1, q_1) over A with $|A| = p_1 + q_1 - n_1$, say $A = \{1, \dots, p_1 + q_1 - n_1\}$. Suppose we have already shown that x_1 with a long pair of periods (p_1, q_1) exists over the alphabet $A' = \{0, \dots, i\}$ for some $i \in \{1, \dots, p_1 + q_1 - n_1 - 1\}$. By coalescing the letter i into $i - 1$, i.e., replacing every appearance of the letter i with $i - 1$, we have produced x_1 over A'' with $|A''| = |A'| - 1$, and evidently this transformation preserves the long pair (p_1, q_1) of periods of x_1 and its length n_1. ■

Corollary 4. *Any word x of length n with a long pair (p, q) of periods satisfying $1 < p < q < n$ over a proper alphabet $A = \{1, \dots, p + q - n\}$ is uniquely determined up to a permutation of A.*

Proof. Define the equivalence relation: $x[h] \sim x[k]$ if there exists a sequence $h = i_1, i_2, \dots, i_\ell = k$ such that $|i_{t+1} - i_t| \in \{0, p, q\}$ for all t satisfying $i_t, i_{t+1} \in [1, n]$. In view of the p- and q-periodicity of x, every member of any given equivalence class must be the same letter. If A is proper, then $|A|$ is the number of distinct equivalence classes. If it's not proper, then we have two equivalence classes labeled by the same letter of the alphabet. ■

Corollary 5. *Any word x_1 of length n_1 with a long pair (p_1, q_1) of periods satisfying $1 < p_1 < q_1 < n_1$ over a proper alphabet $A = \{1, \dots, p_1 + q_1 - n_1\}$ has distinct letters in $x_1[1 \dots p_{t+1}] =$ and in $z = x_1[n_{t+1} + 1 \dots p_t]$, where t is as in Lemma* 2(iv).

Proof. By the proof of Theorem 2, x_1 with the specified letter distribution exists. By Corollary 4, any other word of length n_1 with a long pair (p_1, q_1) of periods over the same proper alphabet, has letters which are a permutation of the letters of x_1. ■

The subword complexity of a word x_1 with a long pair (p_1, q_1) of periods with $p_1 < q_1$ over a proper alphabet A will be called *maximum subword complexity* with respect to p_1, q_1, n_1. Though this notion is generally used for infinite words only, in the present case it is of interest to compute it even for a finite word, in order to gauge the subword complexity of a word with a long pair of periods. We shall now begin to compute the maximum subword complexity. As is usual, the subword complexity computation is not easily accessible.

Lemma 6. *Let $x[1 \dots n]$ be a word with a long pair (p, q) of periods satisfying $p < q$ over a proper alphabet A. Let a, b be any integers satisfying $1 \le a < b \le n$. Then for every $m \in \{1, \dots, n\}$, the number of distinct factors of length m in $x[a \dots b]$ is the same as the number of distinct factors of length m in $x[n - b + 1 \dots n - a + 1]$.*

Proof. Let $y = R(x) = x[n \dots 1]$ be the *reverse* of $x[1 \dots n]$. The definition of the periodicity of x implies that y is also periodic with periods p, q. Thus y is also a word of length n with a long pair (p, q) of periods over the proper alphabet A. By Corollary 4 y is the same as x, up to a permutation of the alphabet A, and the result follows. ■

Remark. The distinct factors in the two intervals of Lemma 6 are not normally identical. It's just the *number* of them that's the same.

Lemma 7. *Let $x[1 \dots n]$, a, b be as in Lemma 6, $m \in \mathbb{Z}_{>0}$. Then the number of factors of length m of $x[a \dots b]$ which begin at different locations of x_1 is* $\max\{0, b - a - m + 2\}$.

Proof. For $m \le b - a + 1$, there are precisely $y + 1$ factors of length m in $x[a \dots b]$, beginning at $a, a + 1, \dots, a + y$, where $a + y + m - 1 = b$. Thus $y + 1 = b - a - m + 2$. ■

Remark. The $b - a - m + 2$ factors of length m of $x[a \dots b]$ are not necessarily distinct.

Notation. For the purposes of the next results, it is convenient to introduce the following notation for every $i \in \{1, \dots, t\}$ and $m_i \in \{1, \dots, n_i - q_i + 1\}$: $r_i = n_i - q_i - m_i + 2$, $s_i = p_i + m_i - 1$, $u_i = n_i - p_i - m_i + 2$, $v_i = q_i + m_i - 1$. Further, we put $w_i = s_i - r_i - m_i + 2 = v_i - u_i - m_i + 2 = p_i + q_i - n_i + m_i - 1$. In the sequel r_i and s_i will play the role of a and b in Lemma 6, u_i and v_i the role of $n - b + 1$ and $n - a + 1$ and w_i will play the role of $b - a - n + 2$ in Lemma 7.

Lemma 8. (i) *For every $i \in \{1, \dots, t\}$ and $m_i \in \{1, \dots, n_i - q_i + 1\}$, we have $[r_i, s_i] \subseteq [1, n_i - 1]$ and $[u_i, v_i] \subseteq [2, n_i]$.* (ii) *For $t \ge 2$ and every $i \in \{1, \dots, t - 1\}$, $n_{i+1} - q_{i+1} < n_i - q_i$.*

Proof. (i) Follows directly from the definitions of r_i, s_i, u_i, v_i. (ii) Follows from (4) and (7). ■

Lemma 9. *Let x_1 of length n_1 with a pair (p_1, q_1) of Type* I *periods and $p_1 < q_1$, be a word over a proper alphabet. Let $m_1 \in \{1, \dots, n_1 - q_1 + 1\}$. Then every factor of length m_1 of x_1 is a copy of a factor of length m_1 of $x_1[r_1 \dots s_1]$, and the w_1 factors of length m_1 of $x_1[r_1 \dots s_1]$ are distinct. Further, the w_1 factors of length m_1 of $x_1[u_1 \dots v_1]$ are distinct, and also constitute the full set of factors of length m_1 appearing in x_1.*

Proof. By Lemma 8(i), both $x_1[r_1 \ldots s_1]$ and $x_1[u_1 \ldots v_1]$ are factors of x_1 for every $m_1 \in \{1, \ldots, n_1 - q_1 + 1\}$.

Lemma 7 implies that $x_1[r_1 \ldots s_1]$ has w_1 factors beginning at different locations. The same holds for $x_1[u_1 \ldots v_1]$. By Corollary 1, $x_1 = (x_2 z)^{k_1} x_2$, where x_2 has period p_2. First, consider the factors of length m_1 in $x_1[r_1 \ldots n_2]$. (Note that (7) implies $r_1 \leq n_2$.) By Lemma 7, there are $n_2 - r_1 - m_1 + 2 = p_2$ of these, and each lies strictly to the left of z. By Corollary 5 (with $t = 1$) we may assume, without loss of generality, that these factors use precisely all the letters $1, \ldots, p_2$, since the left of z is $x_1[1 \ldots p_2 \ldots n_2]$, where $x_1[1 \ldots p_2] = 1 \ldots p_2$, and $x_1[p_2 + i] = x_1[i]$ for $1 \leq i \leq n_2 - p_2$. By construction, x_2 has no period $< p_2$. It follows that any p_2 consecutive factors of length m_1 in $x_1[1 \ldots p_2 \ldots n_2]$ are distinct. In particular, the p_2 factors $x_1[r_1 \ldots r_1 + m_1 - 1], \ldots, x_1[n_2 - m_1 + 1 \ldots n_2]$ are all distinct for every $m_1 \in \{1, \ldots, n_1 - q_1 + 1\}$, and there are no other factors of length m_1 in x_2.

The other factors of length m_1 of $x_1[r_1 \ldots s_1]$, namely those in $x_1[n_2 - m_1 + 2 \ldots s_1]$, all intersect $z = x_1[n_2 + 1 \ldots p_1]$, since the rightmost factor of length m_1 in $x_1[n_2 - m_1 + 2 \ldots s_1]$ begins at $s_1 - m_1 + 1 = p_1$. Moreover, since the letters appearing in z are distinct and different from those in x_2, these factors are distinct and different from those already counted. Together they yield precisely all the w_1 distinct factors of length m_1, as claimed.

Since x_1 has period p_1, any factor of length m_1 in $x_1[s_1 - m_1 + 2 \ldots n_1] = x_1[p_1 + 1 \ldots n_1]$ is a copy of one beginning in $x_1[1 \ldots s_1 - m_1 + 1 = p_1]$, so any factor of length m_1 of x_1 is a copy of a factor of $x_1[r_1 \ldots s_1]$.

Note that $u_1 = n_1 - s_1 + 1$, $v_1 = n_1 - r_1 + 1$. Lemma 6 thus implies that the number of distinct factors of length m_1 in $x_1[r_1 \ldots s_1]$ is the same as the number of factors in $x_1[u_1 \ldots v_1]$. As pointed out in the remark following Lemma 6, the distinct factors in the two intervals specified by Lemma 6 are not identical, in general. However, in the first part of the proof of this lemma, it was shown that *all* factors of length m_1 appear in $x_1[r_1 \ldots s_1]$. Hence also all of them appear in $x_1[u_1 \ldots v_1]$. ■

Lemma 10. *If $x[1...n]$ is a two period word over a proper alphabet with periods p and q then $x[p] \neq x[q]$.*

Proof. Suppose that $x[p] = x[q]$. Then we construct the word $x[0...n]$ where $x[0] = x[p]$. This has periods p and q, length $n+1$ and alphabet of size $p+q-n$, which contradicts Corollary 4. ■

Lemma 11. *Let x_1 of length n_1 with a pair (p_1, q_1) of long periods and $p_1 < q_1$, be a word over a proper alphabet. Let $m_1 \in \{1, \ldots, n_1 - q_1 + 1\}$. Then the w_1 factors of length m_1 of $x_1[r_1 \ldots s_1]$ are distinct, and every other factor of length m_1 of x_1 is a copy of one of these.*

Proof. By Lemma 8(i), both $x_1[r_i \ldots s_i]$ and $x_1[u_i \ldots v_i]$ are factors of x_i for all $m_i \in \{1, \ldots, n_i - q_i + 1\}$ and all $i \in \{1, \ldots, t\}$.

The result follows from Lemma 9 if $t = 1$. We may thus assume that $t \geq 2$. We use descent on the index i of p_i, q_i, n_i. For $1 \leq i < t$, assume that for $m_{i+1} \in \{1, \ldots, n_{i+1} - q_{i+1} + 1\}$, the w_{i+1} factors of length m_{i+1} of $x_1[r_{i+1} \ldots s_{i+1}]$ are distinct, and every other factor of length m_{i+1} of x_{i+1} is a copy of one of these. This assumption holds for $i = t-1$ by Lemma 9. We show that it also holds when $i+1$ is replaced by i. We consider two cases.

(I) $1 \leq m_i \leq n_{i+1} - q_{i+1} + 1$. The p_i-periodicity of x_i implies that every factor of length m_i of x_i is a copy of one whose first letter is in $x_1[1 \ldots p_i]$. Every such factor is thus a factor of $x_1[1 \ldots p_i + m_i - 1]$. Now $p_i + m_i - 1 \leq p_i + n_{i+1} - q_{i+1} = n_{i+1}$, where the equality follows from (4). So every factor of length m_i of x_i is a copy of one in x_{i+1}. Since $m_i \leq n_{i+1} - q_{i+1} + 1$, the descent hypothesis implies that every such factor occurs in $x_1[r_{i+1} \ldots s_{i+1}]$, and all factors of length m_i in this interval are distinct. Now $n_{i+1} - s_{i+1} + 1 = u_{i+1}$, and $n_{i+1} - r_{i+1} + 1 = v_{i+1}$. Hence by Lemma 6, the w_{i+1} distinct factors of length m_i in $x_1[u_{i+1} \ldots v_{i+1}]$ is the same as the number of distinct factors of length m_i in $x_1[r_{i+1} \ldots s_{i+1}]$. Further, by (4), $x_1[u_{i+1} \ldots v_{i+1}] = x_1[n_i - q_i - m_{i+1} + 2 \ldots p_i + m_{i+1} - 1]$. Now $1 \leq m_i \leq n_{i+1} - q_{i+1} + 1$, the same inequality as satisfied by m_{i+1}. Hence we may replace m_{i+1} by m_i in the last equality. In other words, $x_1[u_{i+1} \ldots v_{i+1}] = x_1[r_i \ldots s_i]$. By the same argument, $w_i = w_{i+1}$. It follows that the w_i factors of length m_i of $x_1[r_i \ldots s_i]$ are distinct, and every other factor of length m_i of the prefix x_i of x_1 is a copy of one of these.

(II) $n_{i+1} - q_{i+1} + 1 < m_i \leq n_i - q_i + 1$. We first show that every factor of length m_i of x_i has a copy which is a factor of $x_1[u_i \ldots v_i]$. For $1 \leq \ell \leq n_i - m_i + 1$, let $x_1[\ell \ldots \ell + m_i - 1]$ be any factor of $x_1[1 \ldots n_i]$ of length m_i. Let j be the smallest integer $\geq n_i - p_i + 1$, such that $x_1[j - m_i + 1 \ldots j] = x_1[\ell \ldots \ell + m_i - 1]$. Such j certainly exists by the p_i-periodicity of x_i. We show that $x_1[j - m_i + 1 \ldots j]$ is a factor of $x_1[u_i \ldots v_i]$. By construction, $j - m_i + 1 \geq u_i$, so it suffices to show that $j \leq v_i$. If this is not so, then $j \geq q_i + m_i$. By the q_i-periodicity of x_i, $x_1[j - m_i + 1 \ldots j] = x_1[j - m_i + 1 - q_i \ldots j - q_i]$. By (7), $k_i p_i < q_i$, so by the p_i-periodicity of x_i we have, $x_1[j - m_i + 1 - q_i + k_i p_i \ldots j - q_i + k_i p_i]$. Now $j - q_i + k_i p_i < j$, and

$$
\begin{aligned}
& j - q_i + k_i p_i \geq q_i + m_i - q_i + k_i p_i = m_i + k_i p_i > \\
& n_{i+1} - q_{i+1} + 1 + k_i p_i = n_i - k_i p_i - p_i + 1 + k_i p_i = \\
& n_i - p_i + 1 \geq u_i.
\end{aligned}
$$

Thus, $j - q_i + k_i p_i < j$, yet $j - q_i + k_i p_i \in [u_i, v_i]$, contradicting the minimality of j. Thus $j \leq v_i$ as claimed, so every factor of length m of x_i has a copy in $x_1[u_i \ldots v_i]$.

Secondly, we show that the w_i factors of length m_i of $x_1[u_i \ldots v_i]$ are distinct. By Lemma 6 it is sufficient to show that the w_i factors of $[r_i, s_i]$ are distinct. We prove this by induction on m_i. If $m_i = 1$ then it follows from (I) above that each factor of length 1 (that is, each letter) appearing in x_1 appears in $x_1[n - q_i + 1 ... p_i]$. Since there are $p_i + q_i - n_i$ such factors and we have seen in Theorem 2 that the size of the alphabet of such a word is $p_i + q_i - n_i$, these factors must be distinct. Thus the result holds for $m_i = 1$.

We now suppose it holds for $m_i = 1, ..., \mu$, where $\mu < n_i - q_i + 1$ and, for the sake of contradiction, that it does not hold for $m_i = \mu + 1$. So the length μ factors of $x_1[n_i - q_i - \mu + 2...p_i + \mu - 1]$ are distinct but two of the length $\mu + 1$ factors of $x_1[n_i - q_i - \mu + 1...p_i + \mu]$ are the same. These identical factors must have identical length μ prefixes, so only one of these prefixes can lie in $x_1[n_i - q_i - \mu + 2...p_i + \mu - 1]$. It follows that the two identical factors must be those at either end of $x_1[n_i - q_i - \mu + 1...p_i + \mu]$. That is

$$x_1[n_i - q_i - \mu + 1...n_i - q_i + 1] = x_1[p_i...p_i + \mu] \tag{12}$$

By Lemma 6, the length μ factors of $x_1[n_i - p_i - \mu + 2...q_i + \mu - 1]$ are distinct, but one of the length $\mu + 1$ factors of $x_1[n_i - p_i - \mu + 1...q_i + \mu]$ is repeated. By the argument used in the last paragraph we then have

$$x_1[n_i - p_i - \mu + 1...n_i - p_i + 1] = x_1[q_i...q_i + \mu]. \tag{13}$$

However, by the periodicities of $x_1[1 \ldots n_i]$,

$$\begin{aligned} x_1[n_i - q_i - \mu + 1] &= x_1[n_i - \mu + 1] \\ &= x_1[n_i - p_i - \mu + 1]. \end{aligned}$$

This with (12) and (13) implies that $x_1[p_i] = x_1[q_i]$, contradicting Lemma 10. This contradiction completes the proof. ■

Theorem 3. *Let x_1 of length n_1 be a word with a pair (p_1, q_1) of Type* I *periods satisfying $p_1 < q_1$ over a proper alphabet A with respect to p_1, q_1, n_1, and let $C(m)$ be its maximum subword complexity function with respect to p_1, q_1, n_1. Then*

$$\begin{aligned} C(m) &= p_1 + q_1 - n_1 + m - 1 \quad \textit{for} \quad 1 \le m \le n_1 - q_1 + 1, \\ C(m) &= p_1 \quad \textit{for} \quad n_1 - q_1 + 1 < m \le n_1 - p_1 + 1, \\ C(m) &= n_1 - m + 1 \quad \textit{for} \quad n_1 - p_1 + 1 < m \le n_1. \end{aligned}$$

Proof. Theorem 2 implies that $|A| = C(1) = p_1 + q_1 - n_1$. We consider three cases.

(I) $1 \le m \le n_1 - q_1$. Induction on m. We have just seen that the result holds for $m = 1$. Assume it holds for $m - 1$ ($2 \le m \le n_1 - q_1 + 1$), i.e., $C(m-1) = p_1 + q_1 - n_1 + m - 2$. Thus there are $p_1 + q_1 - n_1 + m - 2$ distinct factors of length $m - 1$ in x_1. Since x_1 has period p_1, each of these factors has a copy whose first letter is in $x_1[1 \ldots p_1]$. Each of these copies is a prefix of a factor of length m, so $C(m) \ge p_1 + q_1 - n_1 + m - 2$. If we had $C(m) = C(m-1)$, then x_1 would be periodic with period $p_1 + q_1 - n_1 + m - 2 \le p_1 - 1$, contradicting the assumption of computing the subword complexity for a proper alphabet with respect to $p_1 < q_1$. Therefore $C(m) \ge p_1 + q_1 - n_1 + m - 1$.

We now show that this is actually an equality. We do this by showing that every factor of length m has a copy in one of the $p_1 + q_1 - n_1 + m - 1$ factors which have their last letter in the interval $I = [n_1 - p_1 + 1, q_1 + m - 1]$.

Let $x_1[\ell \ldots \ell+m-1]$ be any factor of x_1 of length m. Let i be the smallest integer $\geq n_1 - p_1 + 1$ such that $x_1[i \ldots i+m-1] = x_1[\ell \ldots \ell+m-1]$. Since x_1 has period p_1, such i does exist. We wish to show that $i \in I$. Suppose not. Then $i \geq q_1 + m$, so $i - q_1 \geq m \geq 2$. By the q_1-periodicity of x_1 we then have $x_1[i \ldots i+m-1] = x_1[i-q_1 \ldots i+m-1-q_1]$.

By (7), $k_1p_1 < q_1$, hence $x_1[i \ldots i+m-1] = x_1[i-q_1+k_1p_1 \ldots i+m-1-q_1+k_1p_1]$. Now $i-(q_1-k_1p_1) < i$, and $i-(q_1-k_1p_1) \geq q_1+m-q_1+k_1p_1 \geq k_1p_1+2 > n_1-p_1+1$, since the pair (p_1, q_1) is Type I. Thus $i-(q_1-k_1p_1)$ is smaller than i, yet has the desired properties. This contradiction shows that $i \in I$, so $C(m) = p_1+q_1-n_1+m-1$.

(II) $n_1-q_1 < m \leq n_1-p_1+1$. From the case $m = n_1-q_1+1$ we see that the p_1 factors $x_1[1 \ldots n_1-q_1+1], \ldots, x_1[p_1 \ldots n_1-q_1+p_1]$ of length n_1-q_1+1 are all distinct. Therefore the same holds for the p_1 factors $x_1[1 \ldots m], \ldots, x_1[p_1 \ldots p_1+m-1]$ for every m satisfying $p_1+m-1 \leq n_1$. There can be no other factors of this length, since x_1 has period p_1.

(III) $n_1-p_1+1 < m \leq n_1$. As in the previous case, the n_1-m+1 factors $x[1 \ldots m], \ldots, x_1[n_1-m+1 \ldots n_1]$ of length m are all distinct. Thus $C(m) = n_1-m+1$. ■

Remarks.

• Theorem 3 asserts that the maximum subword complexity is linear. In fact, "on the average", the maximum subword complexity is only p_1, the smaller of the two periods of x_1. This low complexity is consistent with the periodic-type structure proclaimed in Theorem 1. For the special case $p_1+q_1-n_1 = 2$, when the word is necessarily binary, we get $c(m) = m+1$ for the beginning part $(1 \leq m \leq n_1-q_1+1)$, the same as for the beginning part of a rational number p/q. For the middle part $(n_1-q_1+1 < m \leq n_1-p_1+1)$, we get $m-(q_1-p_1)+1 \leq c(m) < m+1$. Since $c(m)$ is an integer, $m-(q_1-p_1)+1 \leq c(m) \leq m$. The subword complexity of the trailing part $(n_1-p_1+1 < m \leq n_1)$ is not very meaningful, as it is affected by the finiteness of the word. Thus, on the average, $c(m) \leq m$, less than that of a Sturmian sequence, which is $c(m) = m+1$ (see e.g., [1], [4]). For an *infinite* word, $c(m) \leq m$ implies that the word is ultimately periodic (see [4], Proposition 2).

• Of course for different pairs of long periods of the same word x we have, in general, a proper alphabet of different size, as well as a different subword complexity.

• Theorem 3 allows us to construct infinite words with complexity $C(m) = m+a$ for any positive integer a. First construct a Type I word $x[1 \ldots n]$ with periods p and q where $(p,q) = 1$ and $n = p+q-a-1$. This is easy—we can use, for instance, $p = a+2$, $q = a+3$ and $n = a+4$. By Theorem 3 this has complexity $C(m) = m+a$ for $1 \leq m \leq n-q+1$. We now choose a positive integer k and extend our word to become a Type II word $x[1 \ldots n+kq]$ with periods q and $p+kq$ as follows. For $i = n+1 \ldots n+kq$ set $x[i] = x[i-q]$. Then $x[1 \ldots n+kq]$ clearly has period q. Furthermore,

$$x[1 \ldots n-p] = x[1+p \ldots n] = x[1+p+kq..n+kq]$$

so the word also has period $p + kq$. Again by Theorem 3, it has complexity

$$C(m) = q + (p + kq) - (n + kq) + m - 1 = a + m$$

for $1 \leq m \leq (n+kq)-(p+kq)+1 = n-p+1$. Since $n-p+1 > n-q+1$ we have a longer word with the required complexity for a greater interval of m values. By iterating this process an infinite number of times, with any sequence of k values, we get an infinite word with complexity $C(m) = a + m$ for all values of m. Using $a = 2, p = 4, q = 5$ and $k = 1$ then 2 we get, using a proper alphabet,

$$x[1 \ldots 6] = 001200$$

$$x[1 \ldots 11] = 00120001200$$

$$x[1 \ldots 28] = 0012000120012000120012000 1200$$

the last of which has complexity $C(m) = 2 + m$ for $1 \leq m \leq 7$. When $a = 1$ this process produces a Sturmian word.

References

[1] J.-P. Allouche, Sur la complexité des suites infinies, [On the complexity of infinite sequences], *Belg. Math. Soc. Simon Stevin* **1** (1994) 133–143.

[2] A. Apostolico, M. Farach and C. S. Iliopoulos, Optimal superprimitivity testing for strings, *Inform. Process. Lett.* **39** (1991) 17-20.

[3] J. Berstel and D. Perrin, eds., *Algebraic Combinatorics on Words*, to appear.

[4] S. Ferenczi, Complexity of sequences and dynamical systems, *Discrete Math.* **206** (1999) 145–154.

[5] N. J. Fine and H. S. Wilf, Uniqueness theorems for periodic functions, *Proc. Amer. Math. Soc.* **16** (1965) 109–114.

[6] C. S. Iliopoulos and L. Mouchard, Quasiperiodicity and string covering, *Theoret. Comput. Sci.* **218** (1999) 205–216.

[7] C. S. Iliopoulos and L. Mouchard, Quasiperiodicity: from detection to normal forms, *J. Autom. Lang. Comb.* **4** (1999) 213–228.