Theoretical Informatics and Applications
Theoret. Informatics Appl. **36** (2002) 97–127
DOI: 10.1051/ita:2002005

ON THE DISTRIBUTION OF CHARACTERISTIC PARAMETERS OF WORDS II*

Arturo Carpi[1] and Aldo de Luca[2]

Abstract. The characteristic parameters K_w and R_w of a word w over a finite alphabet are defined as follows: K_w is the minimal natural number such that w has no repeated suffix of length K_w and R_w is the minimal natural number such that w has no right special factor of length R_w. In a previous paper, published on this journal, we have studied the distributions of these parameters, as well as the distribution of the maximal length of a repetition, among the words of each length on a given alphabet. In this paper we give the exact values of these distributions in a special case. However, these values give upper bounds to the distributions in the general case. Moreover, we study the most frequent and the average values of the characteristic parameters and of the maximal length of a repetition over the set of all words of length n.

Mathematics Subject Classification. 68R15, 68R05.

Introduction

In a recent paper [4], which hereafter will be also referred to as CP, we have studied some properties of the distributions of two basic parameters which can be associated with any finite word w on a given alphabet A. These parameters, called *characteristic parameters* and denoted by K_w and R_w, are defined as follows: K_w is the length of the shortest unrepeated suffix of w and R_w is the minimal natural

Keywords and phrases: Special factor, characteristic parameter, repeated factor.

* *The work for this paper has been supported by the Italian Ministry of Education under Project COFIN 2001 – Linguaggi Formali e Automi: teoria ed applicazioni.*

[1] Dipartimento di Matematica e Informatica, Università di Perugia, via Vanvitelli 1, 06123 Perugia, Italy; e-mail: carpi@dipmat.unipg.it

[2] Dipartimento di Matematica dell'Università di Roma "La Sapienza", piazzale Aldo Moro 2, 00185 Roma, Italy and Centro Interdisciplinare "B. Segre", Accademia dei Lincei, via della Lungara 10, 00100 Roma, Italy; e-mail: deluca@mat.uniroma1.it

number such that w has no right special factor of length R_w. We recall that a factor u of a word w is (*right*) *special* if there exist two distinct letters a and b such that ua and ub are both factors of w.

As shown in a series of papers [1–5] characteristic parameters give a great amount of information about the structure of a word. For instance, the maximal length G_w of a repeated factor of a non-empty word w is given by

$$G_w = \max\{R_w, K_w\} - 1. \tag{1}$$

In CP we studied how the values of the characteristic parameters, as well as of some other related quantities, are distributed among the words of each length. More precisely, if A is a fixed d-letter alphabet, for any pair of natural numbers i and n, we denote by $D_R(i,n)$, $D_K(i,n)$, and $D_G(i,n)$ the number of words w of length n on the alphabet A such that, respectively, R_w, K_w, and G_w is equal to i. In the case of a binary alphabet, the values of $D_R(i,n)/2$, $D_K(i,n)/2$, and $D_G(i,n)/2$ for small values of i and n are reported in Tables 1–3 of CP. The following basic relation between D_R and D_G holds: for all $i, n > 0$ one has

$$D_R(i, n+1) = D_R(i,n) + (d-1)D_G(i-1,n). \tag{2}$$

Moreover, for $i, n > 0$ one has

$$D_G(i-1,n) \le D_R(i,n) + D_K(i,n),$$

where equality holds if and only if $i > n/2$. We also showed that when i is fixed and n grows, $D_R(i,n)$ and $D_K(i,n)$ are non-decreasing. This is not true for $D_G(i,n)$, because one has $D_G(i,n) \neq 0$ if and only if $i < n \le i + d^{i+1}$.

In CP we studied the "diagonal behaviour" of D_R, D_K, and D_G, *i.e.*, the behaviour of $D_R(i,n)$, $D_K(i,n)$, and $D_G(i,n)$ when variables i and n are simultaneously increased by 1. We showed that, for any $i, n \ge 0$,

$$D_K(i,n) \le D_K(i+1, n+1), \tag{3}$$

where equality holds if and only if $i > n/2$. In other terms, for any fixed $m \ge 0$, the values of D_K on the points of a diagonal line $(t, m+t)_{t \ge 0}$ are initially increasing and ultimately constant. A similar property holds for functions D^*_G and D^*_K where for all $i, n \ge 0$, $D^*_G(i,n)$ and $D^*_K(i,n)$ denote respectively the number of the words of length n such that $G_w \ge i$ and $K_w \ge i$. Moreover, for $t, m \ge 0$, one has

$$D_R(t, m+t) \le D_R(m, 2m), \tag{4}$$

where the "=" sign holds if and only if $t \ge m$. Similarly, for $m > 0$ and $t \ge 0$ one has

$$D_G(t, m+t) \le D_G(m, 2m), \tag{5}$$

where the "=" sign holds if and only if $t \geq m-1$. When $i \geq n/2$ some noteworthy relations hold. In particular, one has for $i \geq n/2 > 0$

$$D_R(i,n) = (d-1)d^{n-i}\sum_{t=1}^{n-i} d^{-t} D_K(t, 2t-1) \tag{6}$$

and, for $i \geq \lfloor n/2 \rfloor \geq 1$

$$D_R(i,n) = (d-1)D^*_G(i-1, n-1). \tag{7}$$

In this paper we continue the analysis of the distributions of characteristic parameters of words. In Section 2 we obtain explicit arithmetic expressions, involving the Möbius function, for $D_R(i,n)$, $D_K(i,n)$, $D_G(i,n)$, $D^*_K(i,n)$, and $D^*_G(i,n)$, at least when $i > n/2$. In view of the "diagonal behaviour", these expressions give upper bounds to the values of the preceding maps, in the general case.

Another result concerns the counting of repetitions. By *repetition* of length m in a word w we mean any unordered pair of distinct occurrences of the same factor of length m in w. We show that the total number of repetitions of length m in all the words of length n on a d-letter alphabet is given by

$$d^{n-m}\binom{n-m+1}{2}.$$

This and other related results are of interest for applications, since repetitions play an essential role in algorithms for text compression and sequence assembly [5,8,10].

In the last two sections we study the behaviour of $D_R(i,n)$, $D_K(i,n)$, and $D_G(i,n)$ when the length n is fixed and i varies. In Section 3 we are mainly interested in the average values $\langle R\rangle_n$, $\langle K\rangle_n$, $\langle G\rangle_n$ of R_w, K_w, and G_w on the words w of length n on a d-letter alphabet, with $d \geq 2$. We study the most frequent values of the characteristic parameters and of the maximal length of a repetition in the set of words of length n and use these results for evaluating the average values. We show that $\langle G\rangle_n$ and $\langle K\rangle_n$ are upperbounded, respectively, by $\lceil 2\log_d n\rceil - 1/2$ and $\lceil \log_d n\rceil + 2$ while $\langle R\rangle_n$ is lowerbounded by $\lfloor \log_d(n-1)\rfloor$. Moreover,

$$\lim_{n\to\infty} \frac{\langle K\rangle_n}{\log_d n} = \lim_{n\to\infty} (\langle R\rangle_n - \langle G\rangle_n) = 1$$

and

$$\lim_{n\to\infty} (\langle G\rangle_n - \langle G\rangle_{n-1}) = \lim_{n\to\infty} (\langle R\rangle_n - \langle R\rangle_{n-1}) = \lim_{n\to\infty} (\langle K\rangle_n - \langle K\rangle_{n-1}) = 0.$$

We also obtain upper bounds to the number of symmetric words (*cf.* Sect. 3 of CP) of length smaller than n and to the number of semiperiodic words [2] of length n. Moreover, we prove that the fraction of the words of length n which are periodic-like [3] is exactly given by $\langle K\rangle_n - \langle K\rangle_{n-1}$.

In Section 4 we show that the points of maximum of $D_G(i,n)$, viewed as a function of i with n fixed but sufficiently large, lie between $\lfloor \log_d n \rfloor - 1$ and $\lceil 2\log_d n + \log_d \log_d n \rceil - 1$. Similarly, the points of maximum of $D_R(i,n)$ and $D_K(i,n)$ lie, respectively, between $\lfloor \log_d n - \log_d \log_d n \rfloor - 4$ and $\lceil 2\log_d n + \log_d \log_d n \rceil$ and between 0 and $\lceil \log_d n + \log_d \log_d n \rceil + 2$.

1. Preliminaries

Let A be a non-empty set, or *alphabet*, of cardinality $d > 0$. We denote by A^* the set of all finite sequences of elements of A, including the empty sequence, denoted by ϵ. The elements of A are usually called *letters* and those of A^* *words*. The word ϵ is called *empty word*. We set $A^+ = A^* \setminus \{\epsilon\}$. A word $w \in A^+$ can be written uniquely as a sequence of letters as

$$w = a_1 a_2 \cdots a_n,$$

with $a_i \in A$, $1 \le i \le n$, $n > 0$. The integer n is called the *length* of w and denoted by $|w|$. By definition, the length of ϵ is equal to 0. For any $n \ge 0$ we set $A^n = \{w \in A^* \mid |w| = n\}$.

A word w is called *primitive* if it cannot be written as $w = u^r$ with $u \ne \epsilon$ and $r > 1$. Two words $u, v \in A^*$ are *conjugate* if there exist words $r, s \in A^*$ such that $u = rs$ and $v = sr$. As is well known (see [9]), conjugacy is an equivalence relation in A^*. Moreover, any conjugate of a primitive word is primitive. A conjugacy class of a primitive word will be called *primitive*.

Let $w \in A^*$. The word $u \in A^*$ is a *factor* (or *subword*) of w if there exist words λ, μ such that $w = \lambda u \mu$. A factor u of w is called *proper* if $u \ne w$. If $w = u\mu$, for some word μ (resp. $w = \lambda u$, for some word λ), then u is called a *prefix* (resp. *suffix*) of w. For any word w, we denote respectively by $\mathrm{Fact}(w)$, $\mathrm{Pref}(w)$, and $\mathrm{Suff}(w)$ the sets of its factors, prefixes, and suffixes.

Let $u \in \mathrm{Fact}(w)$. Any pair $(\lambda, \mu) \in A^* \times A^*$ such that $w = \lambda u \mu$ is called an *occurrence* of u in w. If $\lambda = \epsilon$ (resp. $\mu = \epsilon$), then the occurrence of u is called *initial* (resp. *terminal*). An occurrence is called *internal* if it is neither initial nor terminal. A factor u of w is *repeated* if it has at least two distinct occurrences in w, otherwise it is called *unrepeated*.

A word s is called a *right* (resp. *left*) *special factor* of w if there exist two letters $x, y \in A$, $x \ne y$, such that $sx, sy \in \mathrm{Fact}(w)$ (resp. $xs, ys \in \mathrm{Fact}(w)$).

With each word w one can associate the word k_w (resp. h_w) defined as the shortest suffix (resp. prefix) of w which is an unrepeated factor of w.

In the following, for any non-empty word w, we shall denote by k'_w (resp. h'_w) the longest repeated suffix (resp. prefix) of w. One has, trivially, $k_w = xk'_w$ and $h_w = h'_w y$ with $x, y \in A$.

For any word w, we shall consider the parameters $K_w = |k_w|$ and $H_w = |h_w|$. Moreover, we shall denote by R_w the minimal natural number such that there is no right special factor of w of length R_w and by L_w the minimal natural number such that there is no left special factor of w of length L_w.

For any $w \in A^+$ we set $B_w = \{a \in A \mid k'_w a \in \mathrm{Fact}(w)\}$. Thus, B_w is the set of letters of A extending on the right k'_w in w. Moreover, we set $B_\epsilon = A$. In Section 2 of CP we proved that for all $w \in A^*$ and any $x \in B_w$ one has

$$K_{wx} = K_w + 1 \quad \text{and} \quad R_{wx} = R_w. \tag{8}$$

Let $w = a_1 \cdots a_n$, $a_r \in A$, $1 \le r \le n$. A *repetition of length* $q > 0$ in w is any pair (i, j), $1 \le i < j \le n - q + 1$, such that

$$a_i a_{i+1} \cdots a_{i+q-1} = a_j a_{j+1} \cdots a_{j+q-1}. \tag{9}$$

Moreover, a repetition of length 0 is any pair (i, j) with $1 \le i < j \le n + 1$.

The maximal length of a repetition in w, that is the maximal length of a repeated factor of w, is denoted by G_w. As proved in Section 3 of CP, for all $w \in A^+$ one has

$$G_w + 1 \le |w| \le G_w + d^{R_w}. \tag{10}$$

In particular, if $d > 1$ one derives

$$G_w \ge \lfloor \log_d |w| \rfloor - 1. \tag{11}$$

The following lemmas will be useful in the sequel:

Lemma 1.1. *Let $d > 1$. For any $w \in A^+$ one has*

$$G_w + R_w \ge 2\lfloor \log_d |w| \rfloor - 1.$$

Proof. Let $w \in A^n$. By equation (10), $G_w + R_w \ge G_w + \log_d(n - G_w)$. Since the second derivative of the function $x + \log_d(n - x)$ is negative, the minimal value of this function in the interval $[\lfloor \log_d n \rfloor - 1, n - 1]$ is equal to

$$\min\{\lfloor \log_d n \rfloor - 1 + \log_d(n - \lfloor \log_d n \rfloor + 1), n - 1\}.$$

By equations (10) and (11), $\lfloor \log_d n \rfloor - 1 \le G_w \le n - 1$ so that one derives

$$G_w + R_w \ge \min\{\lfloor \log_d n \rfloor - 1 + \log_d(n - \lfloor \log_d n \rfloor + 1), n - 1\}. \tag{12}$$

Since for $d \ge 2$ one has $n \ge d\lfloor \log_d n \rfloor$, one gets

$$n - 1 \ge 2\lfloor \log_d n \rfloor - 1 \tag{13}$$

and

$$n - \lfloor \log_d n \rfloor + 1 \ge n - \frac{n}{d} + 1 > \frac{n}{d},$$

so that

$$\lfloor \log_d n \rfloor - 1 + \log_d(n - \lfloor \log_d n \rfloor + 1) > \lfloor \log_d n \rfloor - 1 + \log_d \frac{n}{d} \geq 2\lfloor \log_d n \rfloor - 2. \tag{14}$$

By equations (12, 13), and (14) one obtains $G_w + R_w > 2\lfloor \log_d n \rfloor - 2$, from which the conclusion follows. □

We observe that the lower bound in the preceding lemma is effectively reached. In fact, if w is a de Bruijn word of order m (*cf.* Sect. 3 of CP) one has $R_w = m$, $G_w = m - 1$, and $m = \lfloor \log_d |w| \rfloor$.

Lemma 1.2. *Let $d > 1$. For any $w \in A^+$ such that $R_w < \lfloor \log_d |w| \rfloor$ one has $R_w + K_w \geq 2\lfloor \log_d |w| \rfloor$.*

Proof. By Lemma 1.1 one has $G_w \geq 2\lfloor \log_d |w| \rfloor - 1 - R_w \geq \lfloor \log_d |w| \rfloor > R_w$. This implies $K_w = G_w + 1$ so that $R_w + K_w \geq 2\lfloor \log_d |w| \rfloor$. □

Let us denote by $P_w(q)$ the number of all repetitions of length q in w. For instance, in the case of the word $w = aabaababbab$, as one easily verifies, one has $P_w(1) = 25$, $P_w(2) = 10$, $P_w(3) = 3$, $P_w(4) = 1$, and $P_w(5) = 0$.

Lemma 1.3. *For any $w \in A^+$ and $q \geq 0$, one has*

$$P_w(q) \geq G_w - q + 1.$$

Proof. The result is trivially true if $q = 0$ or $q > G_w$. Thus suppose $0 < q \leq G_w$. Let us write w as $w = a_1 \cdots a_n$ with $a_r \in A$, $1 \leq r \leq n$. Since G_w is the maximal length of a repeated factor of w there exist integers i and j such that $1 \leq i < j \leq n - G_w + 1$ and

$$a_i a_{i+1} \cdots a_{i+G_w-1} = a_j a_{j+1} \cdots a_{j+G_w-1}.$$

Thus, the $G_w - q + 1$ pairs

$$(i, j), (i + 1, j + 1), \ldots, (i + G_w - q, j + G_w - q)$$

are repetitions of length q of w. Hence, $P_w(q) \geq G_w - q + 1$. □

2. Exact computations

In the sequel we shall assume that the alphabet A contains at least two letter, *i.e.*, $d > 1$.

In this section we give explicit arithmetic expressions for $D_R(i, n)$, $D_K(i, n)$, $D_G(i, n)$, $D_K^*(i, n)$, and $D_G^*(i, n)$, involving the Möbius function, at least when $i > n/2$. In view of the diagonal behaviour of these functions, these expressions give upper bounds to the values of the preceding maps, when $i \leq n/2$.

Let $w = a_1a_2\cdots a_n$ be a word, $a_i \in A$, $i = 1, \ldots, n$. We recall (*cf.* [9]) that a positive integer $p \leq n$ is called a *period* of w if for all $i, j \in [1, n]$ such that $i \equiv j \pmod{p}$, one has $a_i = a_j$. For any word w, we denote by π_w its *minimal period*. A word w is called *periodic* if $|w| \geq 2\pi_w$.

The notion of period is also related to the notion of *border* of a word. A word u is called a *border* of w if it is both a proper prefix and a proper suffix of w. The longest border of the word w will be called the *maximal border* of w. It is well known (*cf.* [9]) that the maximal border of a word w has length $|w| - \pi_w$.

Let $\psi : \mathbb{N}_+ \to \mathbb{N}_+$ be the function counting, for any positive integer n the number of primitive words of length n on the alphabet A. As is well known [9], for any n, $\psi(n)$ is given by

$$\psi(n) = \sum_{m|n} \mu(m) d^{\frac{n}{m}},$$

where μ is the Möbius function (see, for instance [7]).

Lemma 2.1. *Let n and p be positive integers such that $n \geq 2p - 2$. The number of words of length n having minimal period p is given by $\psi(p)$.*

Proof. Let u be a primitive word of length p and prolong u on the right in a word w of length n having period p. Let us show that p is the minimal period of w. Indeed, suppose that w has a minimal period $q < p$. Since $n \geq 2p - 2 \geq p + q - 1$, by the theorem of Fine and Wilf [6], w has also the period $\gcd(p, q)$ which has to be equal to q, since q is the minimal period of w. Thus, $p = rq$ with $r > 1$. Since u has the period q, it follows that u is not primitive, which is a contradiction. Conversely, let w be a word of length n having the minimal period p. Then the prefix u of length p of w has to be primitive as, otherwise, the minimal period of w would be less than p.

In conclusion, if $n \geq 2p - 2$, the number of words of length n having minimal period p coincides with the number of primitive words of length p, *i.e.*, $\psi(p)$. □

Lemma 2.2. *Let m be a positive integer and w a word. One has $K_w \geq m$ if and only if there exists $u \in \mathrm{Suff}(w)$ such that $|u| = \pi_u + m - 1$.*

Proof. Let us suppose that $K_w \geq m$. Thus w has a repeated suffix v of length $m - 1$. Let u be the shortest suffix of w with two occurrences of v. This implies that v is a border of u with no internal occurrence in u. Moreover, v is the longest border of u, otherwise v would have an internal occurrence in u. Hence, the minimal period of u is given by $\pi_u = |u| - |v| = |u| - m + 1$.

Conversely, suppose that there exists $u \in \mathrm{Suff}(w)$ such that $|u| = \pi_u + m - 1$. Then, u has a maximal border v of length $|v| = |u| - \pi_u = m - 1$. The word v is a repeated suffix of w so that $K_w \geq |v| + 1 = m$. □

Lemma 2.3. *Let w be a word and $m > |w|/2$. Then there is at most one suffix u of w such that $|u| = \pi_u + m - 1$.*

Proof. Suppose that $u, u' \in \mathrm{Suff}(w)$ are such that $|u| = \pi_u + m - 1$, $|u'| = \pi_{u'} + m - 1$, and $|u'| \leq |u|$. Since $|u| \leq |w| \leq 2m - 1$, it follows that $\pi_u \leq m$ so that

$$|u'| = \pi_{u'} + m - 1 \geq \pi_{u'} + \pi_u - 1.$$

Since u' is a suffix of u, u' has also the period π_u. By the theorem of Fine and Wilf [6], it follows that u' has the period $\gcd(\pi_u, \pi_{u'})$. Thus, since $\pi_{u'}$ is the minimal period of u' one has $\pi_{u'} = \gcd(\pi_u, \pi_{u'})$, so that π_u is a multiple of $\pi_{u'}$. Since $\pi_u \leq |u'|$ and π_u is a multiple of $\pi_{u'}$ it follows that $\pi_{u'}$ is a period of u. Consequently, $\pi_{u'} = \pi_u$ and $|u| = |u'|$ which implies $u = u'$. □

We recall that the map D_K^* is defined for all $i, n \geq 0$ by

$$D_K^*(i, n) = \mathrm{Card}(\{w \in A^n \mid K_w \geq i\}) = \sum_{m \geq i} D_K(m, n).$$

By equation (3) one easily derives (*cf.* Sect. 5 of CP) that for all $i, n \geq 0$,

$$D_K^*(i, n) \leq D_K^*(i + 1, n + 1), \tag{15}$$

where equality holds if and only if $i > n/2$.

Proposition 2.4. *Let m and n be integers with $0 \leq m \leq n$. One has*

$$D_K^*(m, n) \leq \sum_{i=1}^{n-m+1} d^{n-m-i+1}\psi(i),$$

where equality holds if and only if $m > n/2$. In particular, for $0 \leq m \leq n$ one has

$$D_K^*(m, n) \leq (n - m + 1)d^{n-m+1}.$$

Proof. First, we suppose that $m > n/2$. Then, for any $i = 1, \ldots, n - m + 1$, one has $i + m - 1 \geq 2i - 1$ so that, by Lemma 2.1 there are exactly $\psi(i)$ words of minimal period i and length $i + m - 1$. These words can be prolonged on the left into $d^{n-m-i+1}\psi(i)$ words of length n satisfying the condition in Lemma 2.2. Since $m > n/2$, by using Lemma 2.3, one derives that, starting from distinct values of the period i, distinct words are obtained. We conclude that the total number of words w of length n such that $K_w \geq m$, *i.e.*, $D_K^*(m, n)$ is given by $\sum_{i=1}^{n-m+1} d^{n-m-i+1}\psi(i)$.

If, on the contrary, $m \leq n/2$, then by equation (15) and the first part of the proof one has

$$D_K^*(m, n) < D_K^*(m + n + 1, 2n + 1) = \sum_{i=1}^{n-m+1} d^{n-m-i+1}\psi(i).$$

To conclude the proof, we observe that for any $i \geq 1$, trivially $\psi(i) \leq d^i$, so that

$$\sum_{i=1}^{n-m+1} d^{n-m-i+1}\psi(i) \leq \sum_{i=1}^{n-m+1} d^{n-m+1} = (n-m+1)d^{n-m+1}.$$

□

In the sequel, we follow the convention that a sum $\sum_{i=t}^{s} a_i$ holds 0 if $t > s$.

Proposition 2.5. *Let m and n be integers with $0 \leq m \leq n$. One has*

$$D_K(m,n) \leq \psi(n-m+1) + d^{n-m}(d-1)\sum_{i=1}^{n-m} d^{-i}\psi(i),$$

where equality holds if and only if $m > n/2$.

Proof. One can write

$$D_K(m,n) = \sum_{i \geq m} D_K(i,n) - \sum_{i \geq m+1} D_K(i,n) = D_K^*(m,n) - D_K^*(m+1,n).$$

If $m > n/2$, by Proposition 2.4 one has

$$\begin{aligned} D_K(m,n) &= \sum_{i=1}^{n-m+1} d^{n-m-i+1}\psi(i) - \sum_{i=1}^{n-m} d^{n-m-i}\psi(i) \\ &= \psi(n-m+1) + d^{n-m}(d-1)\sum_{i=1}^{n-m} d^{-i}\psi(i). \end{aligned}$$

If, on the contrary, $m \leq n/2$, then, by an iterated application of equation (3), one derives

$$D_K(m,n) < D_K(m+n+1, 2n+1).$$

Since $m+n+1 > (2n+1)/2$, by the previous argument it follows

$$D_K(m,n) < \psi(n-m+1) + d^{n-m}(d-1)\sum_{i=1}^{n-m} d^{-i}\psi(i),$$

which concludes the proof. □

Let us introduce now the function ω defined for any $p \geq 0$ by

$$\omega(p) = \sum_{t=1}^{p} d^{-t-1}\psi(t)(d + (d-1)(p-t)). \tag{16}$$

Proposition 2.6. *Let m, n be integers with $0 \leq m \leq n$. One has*

$$D_R(m,n) \leq (d-1)d^{n-m}\omega(n-m),$$

where equality holds if and only if $m \geq n/2$.

Proof. The statement is trivial for $n = 0$. Let us then suppose $n > 0$.

First we consider the case $m \geq n/2$. By equation (6) one has

$$D_R(m,n) = (d-1)d^{n-m}\sum_{t=1}^{n-m} d^{-t}D_K(t, 2t-1).$$

For $1 \leq t \leq n-m$ one has $(2t-1)/2 < t \leq 2t-1$ so that, by Proposition 2.5

$$D_K(t, 2t-1) = \psi(t) + d^{t-1}(d-1)\sum_{i=1}^{t-1} d^{-i}\psi(i).$$

Thus,

$$D_R(m,n) = (d-1)d^{n-m}\left(\sum_{t=1}^{n-m} d^{-t}\psi(t) + \frac{d-1}{d}\sum_{t=1}^{n-m}\sum_{i=1}^{t-1} d^{-i}\psi(i)\right). \tag{17}$$

Since

$$\sum_{t=1}^{n-m}\sum_{i=1}^{t-1} d^{-i}\psi(i) = \sum_{i=1}^{n-m}(n-m-i)\psi(i)d^{-i},$$

equation (17) becomes

$$\begin{aligned} D_R(m,n) &= (d-1)d^{n-m}\sum_{t=1}^{n-m} d^{-t}\psi(t)\left(1 + \frac{d-1}{d}(n-m-t)\right) \\ &= (d-1)d^{n-m}\omega(n-m). \end{aligned}$$

In the general case, by equation (4) one derives $D_R(m,n) \leq D_R(n-m, 2(n-m))$ and, by the previous argument, $D_R(m,n) \leq (d-1)d^{n-m}\omega(n-m)$, where equality holds if and only if $m \geq n/2$. This proves our assertion. □

We recall that the map D^*_G is defined for $i \geq 0$ and $n > 0$ by

$$D^*_G(i,n) = \mathrm{Card}(\{w \in A^n \mid G_w \geq i\}) = \sum_{m \geq i} D_G(m,n).$$

As proved in Section 5 of CP, for any $i \geq 0$ and $n > 0$ one has

$$D^*_G(i,n) \leq D^*_G(i+1, n+1), \tag{18}$$

where equality holds if and only if $i \geq \lfloor n/2 \rfloor$.

Proposition 2.7. *Let m, n be integers with $0 \leq m < n$. One has*

$$D_G^*(m,n) \leq d^{n-m}\omega(n-m),$$

where equality holds if and only if $m \geq \lfloor n/2 \rfloor$.

Proof. Let us first suppose that $\lfloor n/2 \rfloor \leq m < n$. Since $m+1 \geq (n+1)/2$, by Proposition 2.6 one has $D_R(m+1, n+1) = (d-1)d^{n-m}\omega(n-m)$. Moreover, by equation (7), $D_R(m+1, n+1) = (d-1)D_G^*(m,n)$. This implies $D_G^*(m,n) = d^{n-m}\omega(n-m)$.

In the case $m < \lfloor n/2 \rfloor$, by equation (18), one derives $D_G^*(m,n) < D_G^*(m+n, 2n)$. Since $m+n \geq n$, one has $D_G^*(m+n, 2n) = d^{n-m}\omega(n-m)$, and this proves the assertion. □

The previous proposition gives an upper bound to the number of words having at least one repeated factor of length m. Observe that, since $\psi(t) \leq d^t$, $t \geq 1$, for any $p > 0$, one has

$$\omega(p) < \sum_{t=1}^{p} d^{-t}\psi(t)(1+p-t) < \sum_{t=1}^{p}(1+p-t) = \binom{p+1}{2}.$$

Thus, a less sharp but simpler upper bound to $D_G^*(m,n)$ is given by the following corollary. A similar upper bound was proved recently in [10].

Corollary 2.8. *For $0 \leq m < n$ the following holds:*

$$D_G^*(m,n) < d^{n-m}\binom{n-m+1}{2}.$$

An interpretation of this upper bound can be given in terms of the total number $P(m,n)$ of repetitions of length m in all the words of length n, *i.e.*,

$$P(m,n) = \sum_{w \in A^n} P_w(m).$$

Proposition 2.9. *Let n and m be integers such that $0 \leq m < n$. The following holds:*

$$P(m,n) = d^{n-m}\binom{n-m+1}{2}.$$

Proof. If $m = 0$, then the result is trivial. Thus, we suppose $m > 0$. For any pair of integers i and j such that $1 \leq i < j \leq n-m+1$, we count the words $w = a_1 \cdots a_n$, $a_r \in A$, $1 \leq r \leq n$, satisfying equation (9) with $q = m$. Let us prove that a word w satisfying equation (9) is uniquely determined by the word $a_1 \cdots a_{j-1} a_{j+m} \cdots a_n$. Indeed, by equation (9) one has $a_{j+p} = a_{i+p}$, $0 \leq p \leq m-1$. One derives $a_j = a_i$. If we suppose of knowing all the letters up to a_{j+p-1}, $0 < p \leq m-1$, then $a_{j+p} = a_{i+p}$ and a_{i+p} is already known since $i+p < j+p$. This proves that the

number of the words w satisfying equation (9) is given by d^{n-m}. Since there are $(n-m+1)(n-m)/2$ pairs (i,j) satisfying the condition $1 \leq i < j \leq n-m+1$ the result follows. □

From Proposition 2.9, one obtains a different proof of Corollary 2.8 since, trivially, $D^*_G(m,n) \leq P(m,n)$.

Example 2.10. Let us consider a binary alphabet and let $m = 5$ and $n = 18$. In this case by means of a computer one obtains $D^*_G(5,18) = \sum_{i=5}^{18} D_G(i,18) = 223\,250$ (*cf.* Tab. 3 of CP). By Proposition 2.7 one has $D^*_G(5,18) \leq 2^{13}\omega(13) = D^*_G(12,25) = 363\,874 < P(5,18) = 745\,472$.

Proposition 2.11. *Let m and n be integers with $0 \leq m \leq n$. One has*

$$D_G(m,n) \leq \psi(n-m) + (d-1)d^{n-m-2} \sum_{t=1}^{n-m-1} d^{-t}\psi(t)(2+(d-1)(n-m-t+1)),$$

where equality holds if and only if $m \geq \lfloor n/2 \rfloor$.

Proof. If $m \geq \lfloor n/2 \rfloor$, by Proposition 2.7 one has

$$D_G(m,n) = D^*_G(m,n) - D^*_G(m+1,n) = d^{n-m}\omega(n-m) - d^{n-m-1}\omega(n-m+1).$$

In view of equation (16), by simple algebraic manipulations, one derives

$$D_G(m,n) = \psi(n-m) + (d-1)d^{n-m-2} \sum_{t=1}^{n-m-1} d^{-t}\psi(t)(2+(d-1)(n-m-t+1)).$$

In the general case, by equation (5) one derives $D_G(m,n) \leq D_G(n-m-1, 2(n-m)-1)$, where equality holds if and only if $m \geq \lfloor n/2 \rfloor$. By the previous argument the result follows. □

In conclusion of this section, we compute the exact value of $D_K(2,n)$, $n \geq 0$, in the case of a binary alphabet. We denote by Fib_n the sequence of Fibonacci numbers, defined by

$$\mathrm{Fib}_0 = 0, \quad \mathrm{Fib}_1 = 1, \quad \mathrm{Fib}_{n+1} = \mathrm{Fib}_n + \mathrm{Fib}_{n-1}, \ n \geq 1.$$

Proposition 2.12. *Let $d = 2$. For all $n > 1$ one has*

$$\frac{1}{2} D_K(2,n) = \mathrm{Fib}_{n-1} + n - 2.$$

Proof. For $n = 2$, the result is trivial since $D_K(2,2) = 2$ and $\mathrm{Fib}_1 = 1$. Let us suppose $n > 2$.

We have to count the number of words $w \in \{a,b\}^n$ having $K_w = 2$. For symmetry reasons, we shall count only the words ending by the letter a.

First, we consider the case $k_w = ba$. Thus the letter a, but not ba, has to appear in the prefix of w of length $n-2$. Any occurrence of a in this prefix either is the prefix of w of length 1 or is preceded by another a. Hence, the only possible words of this kind are $a^j b^m a$ with $j, m > 0$ and $j + m + 1 = n$. Their number is $n - 2$.

Now, we consider the case $k_w = aa$. We observe that one has $k_w = aa$ if and only if $w = bu$ or $w = abu$ with $k_u = aa$. Indeed, since $|w| > 2$, w cannot begin by aa, so that either $w = bu$ or $w = abu$. Moreover, aa is an unrepeated suffix of w if and only if it is an unrepeated suffix of u. Let us denote by $g(n)$ the number of words $w \in \{a,b\}^n$ such that $k_w = aa$. The previous argument shows that, for $n > 2$, $g(n) = g(n-1) + g(n-2)$. Since $g(1) = 0 = \mathrm{Fib}_0$ and $g(2) = 1 = \mathrm{Fib}_1$, it follows that $g(n) = \mathrm{Fib}_{n-1}$.

Therefore, the total number of words w of length n ending by a and having $K_w = 2$ is given by $\mathrm{Fib}_{n-1} + n - 2$. □

From the previous proposition, in the case $d = 2$ one easily derives that for $n \geq 4$ one has $D_K(2,n) = D_K(2,n-1) + D_K(2,n-2) - 2(n-5)$. An interesting problem is to determine in the case $i > 2$ similar recursive relations for $D_K(i,n)$.

3. Average values

In this section we shall be mainly interested in the average values $\langle R\rangle_n$, $\langle K\rangle_n$, $\langle G\rangle_n$ of R_w, K_w, and G_w on the words w of length n on the alphabet A. First we evaluate the most frequent values of the characteristic parameters and of the maximal length of a repetition in the set of words of length n. From this, we show that $\langle G\rangle_n$ and $\langle K\rangle_n$ are upperbounded, respectively, by $\lceil 2\log_d n\rceil - 1/2$ and $\lceil \log_d n\rceil + 2$ while $\langle R\rangle_n$ is lowerbounded by $\lfloor \log_d(n-1)\rfloor$. Moreover, one has $\lim_{n\to\infty}\langle K\rangle_n/\log_d n = 1$, $\lim_{n\to\infty}(\langle R\rangle_n - \langle G\rangle_n) = 1$, and $\lim_{n\to\infty}(\langle G\rangle_n - \langle G\rangle_{n-1}) = \lim_{n\to\infty}(\langle R\rangle_n - \langle R\rangle_{n-1}) = \lim_{n\to\infty}(\langle K\rangle_n - \langle K\rangle_{n-1}) = 0$. We also obtain upper bounds to the number of symmetric words of length smaller than n and to the number of semiperiodic words [2] of length n. Moreover, we show that the number of periodic-like words of length n is equal to $d^n(\langle K\rangle_n - \langle K\rangle_{n-1})$.

We start with the following proposition showing that the words of length n having a repeated factor of length significantly larger than $2\log_d n$ are a small fraction of all words of length n. This result was first proved in [10].

Proposition 3.1. *Let $n > 1$ and $r \geq 0$. The following holds:*

$$\mathrm{Card}(\{w \in A^n \mid G_w \geq \lceil 2\log_d n\rceil + r\}) < \frac{1}{2}d^{n-r}.$$

Proof. Let us set $m = \lceil 2\log_d n\rceil + r$. If $m \geq n$, the result is trivially true, because the set $\{w \in A^n \mid G_w \geq m\}$ is empty. Let us then suppose $m < n$. Since $m > 0$,

we can write

$$\binom{n-m+1}{2} d^{n-m} < \frac{n^2}{2} d^{n-m} < \frac{1}{2} d^{n-r}.$$

From Corollary 2.8 the result follows: □

By the previous proposition one has that the number of words w of length n such that $G_w < \lceil 2\log_d n \rceil + r$ is greater than $d^n(1 - d^{-r}/2)$. In particular, if one takes $r = \lfloor \log_d \log_d n \rfloor$, then by the preceding formula and equation (11) one derives that, for a sufficiently large n, the maximal length of repetitions in the overwhelming majority of the words of A^n will lie in the interval

$$[\lfloor \log_d n - 1 \rfloor,\ 2\log_d n + \log_d \log_d n] .$$

Proposition 3.2. *Let $n > 0$ and $r \geq 0$. The following holds:*

$$\mathrm{Card}(\{w \in A^n \mid K_w \geq \lceil \log_d n \rceil + r\}) \leq d^{n-r+1}.$$

Proof. If $r = 0$, the result is trivially true. Let us then suppose $r > 0$ and set $m = \lceil \log_d n \rceil + r$. If $m > n$, then the set $\{w \in A^n \mid K_w \geq m\}$ is empty so that the statement follows. Let us then suppose $m \leq n$. By Proposition 2.4, one has

$$\mathrm{Card}(\{w \in A^n \mid K_w \geq m\}) \leq (n - m + 1)d^{n-m+1} \leq nd^{n-m+1}.$$

Since $m = \lceil \log_d n \rceil + r$ one has $nd^{n-m+1} \leq d^{n-r+1}$, which proves the statement. □

Now, we introduce the sequence ϕ_n of real numbers defined for any $n > 1$ by

$$\phi_n = \log_d n - 2\log_d \log_d n = \log_d \frac{n}{\log_d^2 n} \cdot \tag{19}$$

Proposition 3.3. *One has*

$$\lim_{n \to +\infty} \frac{1}{d^n} \mathrm{Card}(\{w \in A^n \mid K_w \leq \phi_n\}) = 0.$$

Proof. First, we verify that for $1 \leq i \leq n$ one has

$$\frac{1}{d^n} \mathrm{Card}(\{w \in A^n \mid K_w \leq i\}) \leq \left(1 - \frac{1}{d^i}\right)^{\lfloor n/i \rfloor - 1}. \tag{20}$$

Indeed, set $q = \lfloor n/i \rfloor$ and $r = n - iq$. Any word $w \in A^n$ such that $K_w \leq i$ can be factorized as

$$w = s' s_1 s_2 \cdots s_q,$$

with $s_q \in A^i$, $s_1, \ldots, s_{q-1} \in A^i \setminus \{s_q\}$, $s' \in A^r$, since s_q is unrepeated in w. The number of words of this kind is $d^r d^i (d^i - 1)^{q-1}$, so that

$$\mathrm{Card}(\{w \in A^n \mid K_w \leq i\}) \leq d^r d^i (d^i - 1)^{q-1}.$$

Dividing by d^n, one obtains equation (20). Using the inequality $1 + x \leq \mathrm{e}^x$ holding for all real number x, one derives

$$\left(1 - \frac{1}{d^i}\right)^{\lfloor n/i \rfloor - 1} \leq \mathrm{e}^{-d^{-i}\lfloor n/i - 1 \rfloor}. \tag{21}$$

Since both ϕ_n and $d^{-\phi_n} n/\phi_n$ are diverging sequences, if one replaces i by $\lfloor \phi_n \rfloor$ in the right hand side of equation (21), one obtains a sequence which vanishes when n diverges. Thus, the conclusion follows by taking $i = \lfloor \phi_n \rfloor$ in equation (20). □

By taking $r = \lfloor \log_d \log_d n \rfloor$ in Proposition 3.2 and using Proposition 3.3, one derives that

$$\lim_{n \to +\infty} \frac{1}{d^n} \mathrm{Card}(\{w \in A^n \mid \log_d n - 2 \log_d \log_d n < K_w < \log_d n + \log_d \log_d n\}) = 1.$$

Thus, for a sufficiently large n, the minimal length of an unrepeated suffix in the overwhelming majority of the words of A^n will be in the interval

$$[\log_d n - 2 \log_d \log_d n,\ \log_d n + \log_d \log_d n].$$

Now consider the map D_R^* defined for all $i, n \geq 0$ (see Sect. 5 of CP) by

$$D_R^*(i, n) = \mathrm{Card}(\{w \in A^n \mid R_w \geq i\}) = \sum_{m \geq i} D_R(m, n).$$

Since for any $w \in A^n$ one has $R_w \leq G_w + 1$, one derives that for $0 < m \leq n$

$$D_R^*(m, n) \leq D_G^*(m - 1, n). \tag{22}$$

Consequently, for a sufficiently large n the maximal length of right special factors in the overwhelming majority of the words of A^n will not exceed $\lceil 2 \log_d n + \log_d \log_d n \rceil$.

Proposition 3.4. *Let $n > 0$ and $r > 0$. The following holds:*

$$\mathrm{Card}(\{w \in A^n \mid R_w \leq \lfloor \log_d n \rfloor - r\}) \leq d^{n-r+2}.$$

Proof. By Lemma 1.2, if $R_w \leq \lfloor \log_d n \rfloor - r$, then

$$K_w \geq 2\lfloor \log_d n \rfloor - R_w \geq \lfloor \log_d n \rfloor + r \geq \lceil \log_d n \rceil + r - 1.$$

Hence, by Proposition 3.2, the result follows. □

By taking $r = \lfloor \log_d \log_d n \rfloor$ in Propositions 3.4 and 3.1 and using equation (22), one derives that

$$\lim_{n\to+\infty} \frac{1}{d^n} \operatorname{Card}(\{w \in A^n \mid \log_d n - \log_d \log_d n < R_w < 2\log_d n + \log_d \log_d n\}) = 1.$$

In other terms, for a sufficiently large n the maximal length of right special factors in the overwhelming majority of the words of A^n lies in the interval

$$[\log_d n - \log_d \log_d n,\ 2\log_d n + \log_d \log_d n].$$

Now, for any $n > 0$, let us denote by $\langle G\rangle_n$, $\langle R\rangle_n$, and $\langle K\rangle_n$, the average values of the parameters G_w, R_w, and K_w on the words of length n, *i.e.*

$$\langle G\rangle_n = \frac{1}{d^n} \sum_{w\in A^n} G_w, \quad \langle R\rangle_n = \frac{1}{d^n} \sum_{w\in A^n} R_w, \quad \langle K\rangle_n = \frac{1}{d^n} \sum_{w\in A^n} K_w.$$

Note that, for any $n > 0$, one has

$$\langle G\rangle_n = \frac{1}{d^n} \sum_{i=0}^{n} iD_G(i,n) = \frac{1}{d^n} \sum_{i=1}^{n} D_G^*(i,n). \tag{23}$$

In a similar way, one has

$$\langle R\rangle_n = \frac{1}{d^n} \sum_{i=0}^{n} iD_R(i,n) = \frac{1}{d^n} \sum_{i=1}^{n} D_R^*(i,n) \tag{24}$$

and

$$\langle K\rangle_n = \frac{1}{d^n} \sum_{i=0}^{n} iD_K(i,n) = \frac{1}{d^n} \sum_{i=1}^{n} D_K^*(i,n). \tag{25}$$

In the case $d = 2$, the values of $\langle R\rangle_n$, $\langle K\rangle_n$, and $\langle G\rangle_n$ for $1 \le n \le 26$ are given in Table 1.

We recall that, as proved in Sections 2 and 3 of CP, for all $w \in A^+$ the following relations hold:

$$K_w \le K_{xw} \le 1 + K_w, \quad R_w \le R_{xw} \le 1 + R_w, \quad G_w \le G_{xw} \le 1 + G_w. \tag{26}$$

Proposition 3.5. *For any $n > 0$ one has:*

$$\langle K\rangle_n < \langle K\rangle_{n+1} < 1 + \langle K\rangle_n,$$
$$\langle R\rangle_n < \langle R\rangle_{n+1} < 1 + \langle R\rangle_n,$$
$$\langle G\rangle_n < \langle G\rangle_{n+1} < 1 + \langle G\rangle_n.$$

Proof. By equation (26), for any $w \in A^+$ and any $x \in A$ one has $K_w \leq K_{xw} \leq 1 + K_w$. Moreover, for any $n > 0$ there exist certainly a word $w \in A^n$ and letters $x, y \in A$ such that $K_w < K_{xw}$ and $K_{yw} < 1 + K_w$. For instance, one can take $w = a^n$, $x = a$, and $y = b \neq a$. Since

$$d^{n+1}\langle K\rangle_{n+1} = \sum_{u \in A^{n+1}} K_u = \sum_{w \in A^n} \sum_{x \in A} K_{xw},$$

one derives

$$d \sum_{w \in A^n} K_w < d^{n+1}\langle K\rangle_{n+1} < d \sum_{w \in A^n} (1 + K_w).$$

Dividing by d^{n+1} one derives $\langle K\rangle_n < \langle K\rangle_{n+1} < 1 + \langle K\rangle_n$.

TABLE 1. $\langle R\rangle_n$, $\langle K\rangle_n$, and $\langle G\rangle_n$, $1 \leq n \leq 26$.

n	$\langle R\rangle_n$	$\langle K\rangle_n$	$\langle G\rangle_n$
1	0.0000	1.0000	0.0000
2	0.5000	1.5000	0.5000
3	1.0000	2.0000	1.2500
4	1.6250	2.3750	1.6250
5	2.1250	2.7500	2.1875
6	2.6563	3.0625	2.5938
7	3.1250	3.3438	2.9688
8	3.5469	3.5859	3.2969
9	3.9219	3.8125	3.6523
10	4.2871	4.0117	3.9648
11	4.6260	4.1895	4.2422
12	4.9341	4.3516	4.4980
13	5.2161	4.5005	4.7446
14	5.4803	4.6372	4.9758
15	5.7281	4.7633	5.1934
16	5.9607	4.8800	5.3961
17	6.1784	4.9886	5.5889
18	6.3836	5.0903	5.7730
19	6.5783	5.1856	5.9480
20	6.7632	5.2754	6.1146
21	6.9389	5.3602	6.2736
22	7.1062	5.4405	6.4255
23	7.2658	5.5167	6.5705
24	7.4182	5.5893	6.7094
25	7.5638	5.6586	6.8427
26	7.7033	5.7249	6.9709

By equation (26), for any $w \in A^+$ and any $x \in A$ one has $R_w \leq R_{xw} \leq 1 + R_w$ and $G_w \leq G_{xw} \leq 1 + G_w$. Moreover, for any n there exist certainly a word $w \in A^n$ and letters $x, y \in A$ such that $R_w < R_{yw}$, $R_{xw} < 1 + R_w$, $G_w < G_{xw}$ and $G_{yw} < 1 + G_w$. For instance, one can take $w = a^n$, $x = a$, and $y = b \neq a$. Thus, similarly to the case of K_w, one derives $\langle R \rangle_n < \langle R \rangle_{n+1} < 1 + \langle R \rangle_n$ and $\langle G \rangle_n < \langle G \rangle_{n+1} < 1 + \langle G \rangle_n$. □

Proposition 3.6. *For any $n > 1$,*

$$\langle G \rangle_n \leq \lceil 2 \log_d n \rceil - \frac{1}{2}.$$

Proof. By Lemma 1.3 and Proposition 2.9 one derives, for any $q > 0$,

$$\sum_{w \in A^n} G_w \leq \sum_{w \in A^n} (P_w(q) + q - 1) = \binom{n-q+1}{2} d^{n-q} + d^n(q-1).$$

Hence,

$$\langle G \rangle_n \leq \binom{n-q+1}{2} d^{-q} + q - 1. \tag{27}$$

For $q = \lceil 2 \log_d n \rceil$ one has

$$\binom{n-q+1}{2} d^{-q} + q - 1 \leq \frac{1}{2} + \lceil 2 \log_d n \rceil - 1,$$

which concludes the proof. □

Remark 3.7. In the case $d = 2$ by taking $q = \lceil 2 \log_2 n \rceil - 1$ in equation (27) one gets for any $n > 1$, $\langle G \rangle_n \leq \lceil 2 \log_2 n \rceil - 1$.

We observe that by equation (11) one has trivially for all $n > 0$

$$\langle G \rangle_n \geq \lfloor \log_d n \rfloor - 1. \tag{28}$$

Proposition 3.8. *For any $n > 0$,*

$$\langle K \rangle_n \leq \lceil \log_d n \rceil + 2.$$

Proof. By equation (25), one has

$$\langle K \rangle_n = \frac{1}{d^n} \sum_{m=1}^{n} D_K^*(m, n).$$

We set $t = \lceil \log_d n \rceil + 1$. In the previous sum, we majorize $D_K^*(m, n)$ by d^n for $1 \leq m < t$ and by d^{n-m+t} for $t \leq m \leq n$, in view of Proposition 3.2. One has

then

$$\langle K\rangle_n \leq \frac{1}{d^n}\left(\sum_{m=1}^{t-1} d^n + \sum_{m=t}^{n} d^{n-m+t}\right) \leq t-1+\frac{d}{d-1} \leq t+1.$$

Thus $\langle K\rangle_n \leq \lceil \log_d n\rceil + 2$. □

Proposition 3.9. *One has*

$$\lim_{n\to+\infty} \frac{\langle K\rangle_n}{\log_d n} = 1.$$

Proof. One has

$$d^n\langle K\rangle_n = \sum_{w\in A^n} K_w \geq \mathrm{Card}(\{w\in A^n \mid K_w > \phi_n\})\phi_n,$$

where ϕ_n is the sequence defined by equation (19). Thus in view of Proposition 3.8 one has

$$\frac{1}{d^n}\,\mathrm{Card}(\{w\in A^n \mid K_w > \phi_n\})\frac{\phi_n}{\log_d n} \leq \frac{\langle K\rangle_n}{\log_d n} \leq \frac{\lceil \log_d n\rceil + 2}{\log_d n}\,. \tag{29}$$

By Proposition 3.3 one derives

$$\lim_{n\to+\infty} \frac{1}{d^n}\,\mathrm{Card}(\{w\in A^n \mid K_w > \phi_n\}) = 1.$$

Moreover, $\lim_{n\to+\infty} \phi_n/\log_d n = \lim_{n\to+\infty}(\lceil \log_d n\rceil + 2)/\log_d n = 1$, so that the conclusion follows from equation (29). □

Proposition 3.10. *For $n > 1$, one has*

$$d\langle R\rangle_n = \langle R\rangle_{n-1} + (d-1)(\langle G\rangle_{n-1} + 1).$$

Proof. By equations (24) and (2) one has

$$\begin{aligned} d^n\langle R\rangle_n &= \sum_{i=1}^{n} iD_R(i,n) \\ &= \sum_{i=1}^{n} iD_R(i,n-1) + (d-1)\sum_{i=1}^{n} iD_G(i-1,n-1) \\ &= d^{n-1}\langle R\rangle_{n-1} + (d-1)(d^{n-1}\langle G\rangle_{n-1} + d^{n-1}). \end{aligned}$$

Dividing by d^{n-1}, the statement follows. □

Corollary 3.11. *For $n > 0$, one has*

$$\langle R\rangle_n = (d-1)\sum_{m=1}^{n-1} \frac{\langle G\rangle_m + 1}{d^{n-m}}.$$

Proof. By Proposition 3.10 one has

$$\langle R\rangle_n = \frac{1}{d}\langle R\rangle_{n-1} + \frac{d-1}{d}(\langle G\rangle_{n-1} + 1).$$

By iteration, since $\langle R\rangle_1 = 0$, the result follows. □

Corollary 3.12. *For all $n > 1$, one has*

$$\langle G\rangle_{n-1} + \frac{d-2}{d-1} < \langle R\rangle_n < \langle G\rangle_{n-1} + 1.$$

Proof. By Proposition 3.5 one has $\langle R\rangle_{n-1} < \langle R\rangle_n$. Thus, from Proposition 3.10 it follows

$$d\langle R\rangle_n < \langle R\rangle_n + (d-1)(\langle G\rangle_{n-1} + 1)$$

from which one has $\langle R\rangle_n < \langle G\rangle_{n-1} + 1$.

From Propositions 3.10 and 3.5 one gets

$$d\langle R\rangle_n = \langle R\rangle_{n-1} + (d-1)(\langle G\rangle_{n-1} + 1) > \langle R\rangle_n - 1 + (d-1)(\langle G\rangle_{n-1} + 1).$$

Thus

$$\langle R\rangle_n > \langle G\rangle_{n-1} + 1 - \frac{1}{d-1},$$

from which the assertion follows. □

Corollary 3.13. *For all $n > 1$, one has*

$$\langle R\rangle_n \geq \lfloor \log_d(n-1)\rfloor.$$

Proof. By Proposition 3.10 one has

$$d\langle R\rangle_n = \langle R\rangle_{n-1} + \langle G\rangle_{n-1} + 1 + (d-2)(\langle G\rangle_{n-1} + 1). \tag{30}$$

By Lemma 1.1 one has

$$\langle R\rangle_{n-1} + \langle G\rangle_{n-1} + 1 \geq 2\lfloor \log_d(n-1)\rfloor$$

and by equation (28) one obtains

$$\langle G\rangle_{n-1} + 1 \geq \lfloor \log_d(n-1)\rfloor,$$

so that by equation (30) one derives

$$d\langle R\rangle_n \geq d\lfloor \log_d(n-1)\rfloor,$$

which concludes the proof. □

For all $i, n \geq 0$ we set

$$D_K^{>}(i,n) = \mathrm{Card}(\{w \in A^n \mid K_w = i > R_w\}) \cdot \tag{31}$$

In Section 4 of CP we proved the following relation between D_G^* and $D_K^{>}$: for $i \geq 0$ and $n > 1$,

$$D_G^*(i,n) = dD_G^*(i,n-1) + D_K^{>}(i+1,n). \tag{32}$$

From it the following noteworthy proposition follows.

Proposition 3.14. *For $n > 1$, one has*

$$\langle G\rangle_n = \langle G\rangle_{n-1} + \frac{1}{d^n}\,\mathrm{Card}(\{w \in A^n \mid K_w > R_w\})\cdot$$

Proof. By equation (32) one has

$$\sum_{i=0}^{n} D_G^*(i,n) = d\sum_{i=0}^{n} D_G^*(i,n-1) + \sum_{i=0}^{n} D_K^{>}(i+1,n).$$

Since, by equation (23),

$$\sum_{i=0}^{n} D_G^*(i,n) = d^n(1 + \langle G\rangle_n)$$

and

$$\sum_{i=0}^{n} D_K^{>}(i+1,n) = \mathrm{Card}(\{w \in A^n \mid K_w > R_w\}),$$

one derives

$$d^n\langle G\rangle_n + d^n = d^n\langle G\rangle_{n-1} + d^n + \mathrm{Card}(\{w \in A^n \mid K_w > R_w\}),$$

which proves the assertion. □

Lemma 3.15. *Let $\mathcal{C}$ be a conjugacy class of A^+. Then one has*

$$\mathrm{Card}(\{w \in \mathcal{C} \mid K_w > R_w\}) \leq 2\left(\max_{w\in\mathcal{C}} K_w - 1\right).$$

Proof. Set $t = \max_{w\in\mathcal{C}} K_w$ and let $v \in \mathcal{C}$ be such that $K_v = t$. Let us verify that for any $u \in \mathcal{C}$, if k'_v has two non-terminal occurrences in u, then $R_u \geq K_u$. Indeed, either k'_v is a right special factor of u or it can be extended on the right in a repeated factor $k'_v x$, $x \in A$, of u of length t. In the first case,

$$R_u \geq t \geq K_u\,;$$

in the second case,

$$G_u \geq t \geq K_u\,,$$

so that, in view of equation (1), $G_u = R_u - 1$ and $K_u < R_u$.

To complete the proof, it is sufficient to observe that $\mathcal{C}$ contains at most $2(t-1)$ words which do not have two non-terminal occurrences of k'_v. □

Proposition 3.16. *For any $n > 1$ one has*

$$\frac{1}{d(n-1)} \leq \frac{1}{d^n}\,\mathrm{Card}(\{w \in A^n \mid K_w > R_w\}) < \frac{6\log_d n + 1}{n}\cdot$$

Proof. Let t be a fixed integer such that $1 \leq t \leq n/2$. Let S be the set of the conjugacy classes $\mathcal{C} \subseteq A^n$ such that $\max_{v\in\mathcal{C}} K_v \leq t$ and T be the set of the remaining conjugacy classes $\mathcal{C} \subseteq A^n$. One has that

$$\mathrm{Card}(\{w \in A^n \mid K_w > R_w\}) \leq \sum_{\mathcal{C}\in S} \mathrm{Card}(\{w \in \mathcal{C} \mid K_w > R_w\}) + \sum_{\mathcal{C}\in T} \mathrm{Card}(\{w \in \mathcal{C} \mid K_w > R_w\})\cdot$$

A conjugacy class $\mathcal{C} \in S$ is primitive since for a non-primitive word w, $K_w > n/2$. Since the number of primitive words of length n is $\psi(n)$, there are $\psi(n)/n$ primitive conjugacy classes included in A^n and, therefore,

$$\mathrm{Card}(S) \leq \frac{\psi(n)}{n} < \frac{d^n}{n}\cdot$$

By Lemma 3.15, for any $\mathcal{C} \in S$, one has

$$\mathrm{Card}(\{w \in \mathcal{C} \mid K_w > R_w\}) \leq 2(t-1),$$

so that

$$\sum_{\mathcal{C}\in S} \mathrm{Card}(\{w \in \mathcal{C} \mid K_w > R_w\}) < 2\frac{d^n}{n}(t-1)\,.$$

Any conjugacy class $\mathcal{C} \in T$ contains at least one word w such that $K_w > t$. By Proposition 2.4, the number of the words of length n such that $K_w > t$ is

upperbounded by nd^{n-t}. Thus, $\mathrm{Card}(T) \leq nd^{n-t}$. Since any conjugacy class of a word of length n contains at most n elements, one derives

$$\sum_{\mathcal{C}\in T} \mathrm{Card}(\{w \in \mathcal{C} \mid K_w > R_w\}) \leq n^2 d^{n-t}.$$

Thus

$$\mathrm{Card}(\{w \in A^n \mid K_w > R_w\}) < d^n \left(\frac{2(t-1)}{n} + \frac{n^2}{d^t}\right).$$

Let us suppose that $n/2 \geq \lceil 3\log_d n\rceil$. Then, in the previous equation, we can take $t = \lceil 3\log_d n\rceil$, obtaining

$$\mathrm{Card}(\{w \in A^n \mid K_w > R_w\}) < d^n \frac{6\log_d n + 1}{n}.$$

If, on the contrary, $n/2 < \lceil 3\log_d n\rceil$, then $(6\log_d n + 1)/n \geq 1$, so that in any case one derives

$$\frac{1}{d^n}\mathrm{Card}(\{w \in A^n \mid K_w > R_w\}) < \frac{6\log_d n + 1}{n}.$$

Now, let us verify that

$$\frac{1}{d(n-1)} < \frac{1}{d^n}\mathrm{Card}(\{w \in A^n \mid K_w > R_w\}).$$

We remark that in any conjugacy class $\mathcal{C}$ there is at least one word w such that $K_w \geq R_w$. Indeed, let $t = \max_{v\in\mathcal{C}} G_v$. Then there is at least one word $w \in \mathcal{C}$ having a repeated suffix of length t. For such a w, $K_w = 1 + t = 1 + G_w \geq R_w$. For any $n > 1$, the number of conjugacy classes included in A^{n-1} is greater than or equal to $d^{n-1}/(n-1)$. Hence,

$$\mathrm{Card}(\{w \in A^{n-1} \mid K_w \geq R_w\}) \geq \frac{d^{n-1}}{n-1}.$$

By equation (8) for any word $w \in A^{n-1}$ such that $K_w \geq R_w$ there exists at least one letter $x \in B_w$ such that $K_{wx} = K_w + 1$ and $R_{wx} = R_w$, so that $K_{wx} > R_{wx}$. Then one derives that

$$\mathrm{Card}(\{v \in A^n \mid K_v > R_v\}) \geq \mathrm{Card}(\{w \in A^{n-1} \mid K_w \geq R_w\}) \geq \frac{d^{n-1}}{n-1}.$$

From this, the result follows. □

Remark 3.17. The class of words $w \in A^*$ such that $R_w < K_w$ has been introduced in [2]. We recall that these words can be also characterized as the words $w \in A^*$ which can be prolonged on the right in ultimately periodic words without

adding new factors of length $1 + R_w$. This class properly contains the class of *semiperiodic words*, *i.e.*, the words w such that $R_w < H_w$ [2]. In fact, as proved in [2], for a semiperiodic word w, $R_w = L_w < H_w = K_w$. As a consequence of Proposition 3.16, one has that *the fraction of the words of length n which are semiperiodic is upperbounded by* $(6 \log_d n + 1)/n$.

The following corollary shows that $\lim_{n\to\infty}(\langle G\rangle_n - \langle G\rangle_{n-1}) = 0$.

Corollary 3.18. *For $n > 1$, one has*

$$\frac{1}{d(n-1)} \le \langle G\rangle_n - \langle G\rangle_{n-1} < \frac{6 \log_d n + 1}{n}\cdot$$

Proof. The result follows from the preceding proposition and Proposition 3.14. □

Lemma 3.19. *For any $n > 0$ set $V = \{v \in A^n \mid K_v > R_v\}$. One has*

$$\sum_{w\in V} K_w < \frac{12 d^n}{n}(\log_d n + 1)^2.$$

Proof. Set $t = \lceil 2 \log_d n\rceil$. One has

$$\begin{aligned}\sum_{w\in V} K_w &= \sum_{i\ge 1} i \,\mathrm{Card}(\{w \in A^n \mid K_w = i \text{ and } K_w > R_w\})\\ &= \sum_{i\ge 1} \mathrm{Card}(\{w \in A^n \mid K_w \ge i \text{ and } K_w > R_w\})\\ &\le \sum_{i=1}^{t} \mathrm{Card}(\{w \in A^n \mid K_w > R_w\}) + \sum_{i\ge t+1} \mathrm{Card}(\{w \in A^n \mid K_w \ge i\})\cdot\end{aligned}$$

By Proposition 3.16, as $t < 2 \log_d n + 1$, one has

$$\sum_{i=1}^{t} \mathrm{Card}(\{w \in A^n \mid K_w > R_w\}) < t d^n \frac{6 \log_d n + 1}{n} < \frac{d^n}{n}(12 \log_d^2 n + 8 \log_d n + 1).$$

By Proposition 2.4 one has

$$\sum_{i\ge t+1} \mathrm{Card}(\{w \in A^n \mid K_w \ge i\}) \le \sum_{i\ge t+1} n d^{n-i+1} = n d^{n-t} \sum_{m=0}^{\infty} d^{-m} \le 2\frac{d^n}{n}\cdot$$

Thus,

$$\sum_{w\in V} K_w < \frac{d^n}{n}(12 \log_d^2 n + 8 \log_d n + 3) \le \frac{12 d^n}{n}(\log_d n + 1)^2,$$

which concludes the proof. □

The following proposition shows that for a sufficiently large n, $\langle R\rangle_n$ is approximately given by $\langle G\rangle_n + 1$:

Proposition 3.20. *One has* $\lim_{n\to\infty}(\langle R\rangle_n - \langle G\rangle_n) = 1$.

Proof. For any $n > 0$ one has:

$$\langle G\rangle_n + 1 - \langle R\rangle_n = \frac{1}{d^n}\sum_{w\in A^n}(G_w + 1 - R_w).$$

By equation (1), if $K_w \le R_w$, then $G_w + 1 - R_w = 0$. If, on the contrary, $K_w > R_w$, then $G_w + 1 - R_w = K_w - R_w \le K_w$. Hence,

$$\langle G\rangle_n + 1 - \langle R\rangle_n \le \frac{1}{d^n}\sum_{w\in V} K_w,$$

where $V = \{v \in A^n \mid K_v > R_v\}$. By Lemma 3.19, and since $\langle R\rangle_n \le \langle G\rangle_n + 1$, one derives

$$0 \le \langle G\rangle_n + 1 - \langle R\rangle_n < 12\frac{(\log_d n + 1)^2}{n}.$$

From this, the statement follows. □

Proposition 3.21. *One has* $\lim_{n\to\infty}(\langle R\rangle_n - \langle R\rangle_{n-1}) = 0$.

Proof. By Proposition 3.10 one has

$$\begin{aligned}\langle R\rangle_n - \langle R\rangle_{n-1} &= (d-1)(\langle G\rangle_{n-1} + 1 - \langle R\rangle_n)\\ &= (d-1)(\langle G\rangle_{n-1} - \langle G\rangle_n + \langle G\rangle_n + 1 - \langle R\rangle_n).\end{aligned}$$

By the previous proposition, $\lim_{n\to\infty}(\langle G\rangle_n + 1 - \langle R\rangle_n) = 0$. By Corollary 3.18, $\lim_{n\to\infty}(\langle G\rangle_{n-1} - \langle G\rangle_n) = 0$, that implies $\lim_{n\to\infty}(\langle R\rangle_n - \langle R\rangle_{n-1}) = 0$. □

We recall that a *symmetric word* of *order* m is any word w such that $R_w = K_w = m$. Let $S(m,n)$ denote the class of all symmetric words of length n and order m on the alphabet A. Let $D_K^{>}$ be the map defined by equation (31). The following result was proved in Section 4 of CP: for $0 \le i \le n$ one has

$$D_K^{>}(i,n) = \sum_{m=1}^{i}\sum_{w\in S(i-m,n-m)} \mathrm{Card}(B_w). \tag{33}$$

Proposition 3.22. *For any* $n > 0$ *one has*

$$\mathrm{Card}(\{w \in A^* \mid |w| < n \text{ and } R_w = K_w\}) < d^n\frac{6\log_d n + 1}{n}.$$

Proof. For $0 \leq i \leq n$ by equation (33) one has

$$D_K^{>}(i,n) = \sum_{m=1}^{i} \sum_{w \in S(i-m,n-m)} \mathrm{Card}(B_w),$$

so that, since for any word w, $\mathrm{Card}(B_w) \geq 1$,

$$\sum_{i=1}^{n} D_K^{>}(i,n) \geq \sum_{i=1}^{n} \sum_{m=1}^{i} \mathrm{Card}(S(i-m,n-m)) \geq \mathrm{Card}\left(\bigcup_{i=1}^{n} \bigcup_{m=1}^{i} S(i-m,n-m) \right).$$

As one easily verifies, one has

$$\bigcup_{i=1}^{n} \bigcup_{m=1}^{i} S(i-m,n-m) = \bigcup_{j=0}^{n-1} \bigcup_{p=j}^{n-1} S(j,p) = \{w \in A^* \mid |w| < n \text{ and } R_w = K_w\}.$$

Hence,

$$\begin{aligned} \mathrm{Card}(\{w \in A^* \mid |w| < n \text{ and } R_w = K_w\}) &\leq \sum_{i=1}^{n} D_K^{>}(i,n) \\ &= \mathrm{Card}(\{w \in A^n \mid K_w > R_w\}). \end{aligned}$$

From Proposition 3.16 the conclusion follows. □

As a consequence of the previous proposition, one derives that the fraction of the words of length n which are symmetric is upperbounded by the quantity $6d(\log_d(n+1)+1)/(n+1)$.

We recall that a word w is called *periodic-like* [3] if k'_w (or h'_w) has no internal occurrence in w. As proved in [3], the class of periodic words is properly included in the class of semiperiodic words and this latter is properly included in the class of periodic-like words. Let P be the set of periodic-like words of A^* and D_P the map defined for all $i, n \geq 0$ by

$$D_P(i,n) = \mathrm{Card}(\{w \in P \cap A^n \mid K_w = i\}).$$

In other terms, $D_P(i,n)$ gives the number of periodic-like words of length n having the shortest unrepeated suffix of length i. In Section 4 of CP the following relation between the maps D_K and D_P was proved: for all $i, n > 0$ one has

$$D_K(i,n) = dD_K(i,n-1) + D_P(i,n) - D_P(i+1,n-1). \tag{34}$$

Proposition 3.23. *For any $n > 0$ one has*

$$\frac{\mathrm{Card}(P \cap A^n)}{d^n} = \langle K \rangle_n - \langle K \rangle_{n-1}.$$

Proof. From equation (34) one derives

$$\sum_{i=1}^{n} iD_K(i,n) = d\sum_{i=1}^{n} iD_K(i,n-1) + \sum_{i=1}^{n} iD_P(i,n) - \sum_{i=1}^{n} iD_P(i+1,n).$$

Now,

$$\sum_{i=1}^{n} iD_P(i,n) - \sum_{i=1}^{n} iD_P(i+1,n) = \sum_{i=1}^{n} D_P(i,n) = \mathrm{Card}(P \cap A^n).$$

Thus, by equation (25) one obtains

$$d^n \langle K \rangle_n = d^n \langle K \rangle_{n-1} + \mathrm{Card}(P \cap A^n),$$

which proves our assertion. □

Proposition 3.24. *One has*

$$\lim_{n \to +\infty} \frac{\mathrm{Card}(P \cap A^n)}{d^n} = 0\,.$$

Proof. Let ϕ_n be the sequence defined by equation (19). By Proposition 3.3 one has that

$$\frac{1}{d^n}\,\mathrm{Card}(\{w \in P \cap A^n \mid K_w < \phi_n\})$$

vanishes when n diverges. We recall [3] that the minimal period of a periodic-like word $w \in A^n$ is given by $\pi_w = n - K_w + 1$. Therefore, since the number of words of length n having minimal period p is not larger than d^p, one has

$$\begin{aligned}\frac{1}{d^n}\,\mathrm{Card}(\{w \in P \cap A^n \mid K_w \geq \phi_n\}) &\leq \frac{1}{d^n}\,\mathrm{Card}(\{w \in A^n \mid \pi_w \leq n - \phi_n + 1\}) \\ &\leq \frac{1}{d^n} \sum_{p=1}^{n-\phi_n+1} d^p < d^{-\phi_n+2}.\end{aligned}$$

Since ϕ_n diverges with n, one derives that $\mathrm{Card}(P \cap A^n)/d^n$ vanishes when n diverges. □

By Propositions 3.23 and 3.24 one obtains the following:

Corollary 3.25.

$$\lim_{n \to +\infty} (\langle K \rangle_n - \langle K \rangle_{n-1}) = 0\,.$$

4. Points of maximum

In this section we show that the points of maximum of $D_G(i,n)$, viewed as a function of i with n fixed but sufficiently large, lie between $\lfloor \log_d n \rfloor - 1$ and $\lceil 2\log_d n + \log_d \log_d n \rceil - 1$. Similarly, the points of maximum of $D_K(i,n)$ and $D_R(i,n)$ lie, respectively, between 0 and $\lceil \log_d n + \log_d \log_d n \rceil + 2$ and between $\lfloor \log_d n - \log_d \log_d n \rfloor - 4$ and $\lceil 2\log_d n + \log_d \log_d n \rceil$.

Proposition 4.1. *There exists an integer n_0 such that for all $n \geq n_0$ one has*

$$\max_{0\leq i\leq n} D_G(i,n) = \max\{D_G(i,n) \mid \lfloor \log_d n \rfloor - 1 \leq i < \lceil 2\log_d n + \log_d \log_d n \rceil\}.$$

Proof. Let us take $n \geq d$ and set $m = \lceil 2\log_d n + \log_d \log_d n \rceil$. Since $m > 0$, from Corollary 2.8 one has

$$\sum_{i\geq m} D_G(i,n) = \mathrm{Card}(\{w \in A^n \mid G_w \geq m\}) < \frac{d^n}{2\log_d n}\,. \tag{35}$$

Thus,

$$\max_{i\geq m} D_G(i,n) < \frac{d^n}{2\log_d n}\,. \tag{36}$$

By equation (35), one has

$$\mathrm{Card}(\{w \in A^n \mid G_w < m\}) > \frac{d^n}{2\log_d n}(2\log_d n - 1)\,. \tag{37}$$

Thus, since by equation (11), $D_G(i,n) = 0$ for $i < \lfloor \log_d n \rfloor - 1$, one has

$$\frac{d^n}{2\log_d n}(2\log_d n - 1) < \sum_{i=\lfloor \log_d n \rfloor - 1}^{m-1} D_G(i,n) \leq M(\log_d n + \log_d \log_d n + 3)$$

where $M = \max\{D_G(i,n) \mid \lfloor \log_d n \rfloor - 1 \leq i < m\}$. Hence,

$$M > \frac{d^n}{2\log_d n}\ \frac{2\log_d n - 1}{\log_d n + \log_d \log_d n + 3}.$$

There exists an integer $n_0 \geq d$ such that for all $n \geq n_0$ one has

$$\frac{2\log_d n - 1}{\log_d n + \log_d \log_d n + 3} \geq 1\,.$$

Thus, for $n \geq n_0$ one has

$$M > \frac{d^n}{2\log_d n}\,,$$

which, in view of equation (36), proves our assertion. □

Proposition 4.2. *There exists an integer n_0 such that for all $n \geq n_0$ one has*

$$\max_{0\leq i\leq n} D_K(i,n) = \max\{D_K(i,n) \mid 0 \leq i < \lceil \log_d n + \log_d \log_d n + 2\rceil\}\cdot$$

Proof. Let $n \geq d$ and set $m = \lceil \log_d n + \log_d \log_d n + 2\rceil$. From Proposition 2.4 one has

$$\sum_{i\geq m} D_K(i,n) < nd^{n-m+1} \leq \frac{d^{n-1}}{\log_d n}\,, \tag{38}$$

so that

$$\max_{i\geq m} D_K(i,n) < \frac{d^{n-1}}{\log_d n}\cdot \tag{39}$$

By equation (38), one has

$$\mathrm{Card}(\{w \in A^n \mid K_w < m\}) > d^n\left(1 - \frac{1}{d\log_d n}\right) = \frac{d^{n-1}}{\log_d n}(d\log_d n - 1)\,. \tag{40}$$

Thus,

$$\frac{d^{n-1}}{\log_d n}(d\log_d n - 1) < \sum_{i=1}^{m-1} D_K(i,n) \leq M(\log_d n + \log_d \log_d n + 2)\,,$$

where $M = \max\{D_K(i,n) \mid 1 \leq i < m\}$. Hence,

$$M > \frac{d^{n-1}}{\log_d n}\ \frac{d\log_d n - 1}{\log_d n + \log_d \log_d n + 2}.$$

There exists an integer $n_0 \geq d$ such that for all $n \geq n_0$ one has

$$\frac{d\log_d n - 1}{\log_d n + \log_d \log_d n + 2} \geq 1\,.$$

Hence, for $n \geq n_0$ one has

$$M > \frac{d^{n-1}}{\log_d n}\,,$$

which, in view of equation (39), proves our assertion. □

Proposition 4.3. *There exists an integer n_0 such that for all $n \geq n_0$ one has*

$$\max_{0 \leq i \leq n} D_R(i,n) = \max_{m' < i \leq m} D_R(i,n),$$

where $m' = \lfloor \log_d n - \log_d \log_d n - 4 \rfloor$ and $m = \lceil 2 \log_d n + \log_d \log_d n \rceil$.

Proof. Let us observe that

$$\mathrm{Card}(\{w \in A^n \mid R_w > m\}) \leq \mathrm{Card}(\{w \in A^n \mid G_w \geq m\}) = D^*_G(m,n)\,,$$

so that, by equation (35),

$$\mathrm{Card}(\{w \in A^n \mid R_w > m\}) < \frac{d^n}{2 \log_d n}\cdot$$

Since, as one easily derives, $m' \leq \lfloor \log_d n \rfloor - \lceil 2 + \log_d(2 \log_d n) \rceil$, by Proposition 3.4 one has

$$\mathrm{Card}(\{w \in A^n \mid R_w \leq m'\}) \leq \frac{d^n}{2 \log_d n}\cdot$$

Consequently,

$$\mathrm{Card}(\{w \in A^n \mid m' < R_w \leq m\}) > d^n \left(1 - \frac{1}{\log_d n}\right).$$

Thus, one derives that

$$\max_{i > m} D_R(i,n) \leq \sum_{i > m} D_R(i,n) = \mathrm{Card}(\{w \in A^n \mid R_w > m\}) < \frac{d^n}{2 \log_d n}$$

and

$$\max_{i \leq m'} D_R(i,n) \leq \sum_{i \leq m'} D_R(i,n) = \mathrm{Card}(\{w \in A^n \mid R_w \leq m'\}) \leq \frac{d^n}{2 \log_d n}\cdot$$

Moreover,

$$\begin{aligned}\max_{m' < i \leq m} D_R(i,n) &\geq \frac{\sum_{i=m'+1}^{m} D_R(i,n)}{m - m'} \\ &= \frac{\mathrm{Card}(\{w \in A^n \mid m' < R_w \leq m\})}{m - m'} > \frac{d^n}{\log_d n} \frac{\log_d n - 1}{m - m'}\cdot\end{aligned}$$

As

$$\lim_{n \to \infty} \frac{\log_d n - 1}{m - m'} = 1\,,$$

one derives that for all sufficiently large n

$$\max_{m'<i\leq m} D_R(i,n) > \frac{d^n}{2\log_d n}\,,$$

which proves our assertion. □

Acknowledgements. The authors are thankful to Istituto di Cibernetica "E. Caianiello" del C.N.R. for its valuable support to this research.

References

[1] A. Carpi and A. de Luca, Words and special factors. *Theoret. Comput. Sci.* **259** (2001) 145-182.
[2] A. Carpi and A. de Luca, Semiperiodic words and root-conjugacy. *Theoret. Comput. Sci.* (to appear).
[3] A. Carpi and A. de Luca, Periodic-like words, periodicity, and boxes. *Acta Informatica* **37** (2001) 597-618.
[4] A. Carpi and A. de Luca, On the distribution of characteristic parameters of words. *RAIRO: Theoret. Informatics Appl.* **36** (2002) 99-128.
[5] A. Carpi, A. de Luca and S. Varricchio, Words, univalent factors, and boxes. *Acta Informatica* **38** (2002) 409-436.
[6] N.J. Fine and H.S. Wilf, Uniqueness theorem for periodic functions. *Proc. Amer. Math. Soc.* **16** (1965) 109-114.
[7] G.H. Hardy and E.M. Wright, *An Introduction to the Theory of Numbers*. Oxford University Press, Oxford, UK (1979).
[8] J.D. Kececioglu and E.W. Myers, Combinatorial algorithms for DNA sequence assembly. *Algorithmica* **13** (1995) 7-51.
[9] M. Lothaire, *Combinatorics on Words*, 2nd Edition. Cambridge Mathematical Library, Cambridge University Press, Cambridge, UK (1997).
[10] F. Mignosi, A. Restivo and M. Sciortino, Forbidden factors and fragment assembly. *RAIRO: Theoret. Informatics Appl.* (to appear).

Communicated by J. Berstel.
Received November 15, 2001. Accepted May 27, 2002.