Annales Univ. Sci. Budapest., Sect. Comp. 8 (1987) 69-90

ON THE d-COMPLEXITY OF WORDS

ANTAL IVÁNYI

Dedicated to Professor Imre Kátai
on the occasion of his fiftieth birthday

Introduction

Sequences of elements of given sets of symbols have a great importance in different branches of natural science. For example, in biology the 4-letter set $\{A, C, G, T\}$ containing the nucleotids (adenine, cytosine, guanine and thymine) and the 20-letter one $\{a, c, d, e, f, g, h, i, k, l, m, n, p, q, r, s, t, v, w, y, \}$, containing the amino-acids (alanine, cysteine, asparagine-acid, glutamine-acid, phenyl, glycine, histidine, isoleucine, lysine, leucine, methionine, asparagine, proline, glutamine, arginine, serine, threonine, valine, triptophan, tyrosine) play an important role.

Complexity is an important characteristic of symbol sequences, since it affects the cost of storage and reproduction, and the quantity of information stored in the symbol sequences. The usual complexity measures of symbol sequences are based on the time (or memory) needed for generating or recognizing them.

In this paper a new complexity measure, d-complexity is studied. This measure is also intended to express the average quantity of information included in a sequence. The background of the new complexity measure lies in biology. Some natural sequences, as amino-acid sequences in proteins or nucleotid sequences in DNS-moleculas have a winding structure [1] and some bends can be cut forming new and, of course, shorter sequences. The parameter d is the bound for the length of bends, which can be cut, or, in other word, d is the maximum permissible distance between any two remaining consecutive elements of the sequence.

This concept covers some known complexity measures studied earlier, such as subword complexity (case $d = 1$) and subsequence complexity (case $d = \infty$).

We use the basic concepts and notations of formal language theory [2] and graph theory [3].

1. Basic notations and definitions

Let n and k be positive integers, $X = \{A_1, \dots, A_n\}$ an alphabet, X^k the set of words of length k over X, X^+ the set of finite nonempty words over X. The length of a word $p \in X^+$ is denoted by $L(p)$.

DEFINITION 1 [4]. Let d, r and s be positive integers, $p = x_1 \dots x_r \in X^r$ and $q = y_1 \dots y_s \in X^s$. p is a *d-subword* of q ($p \subset_d q$) iff there exists a sequence $i_1, \dots, i_r$ with $1 \le i_1$, $i_r \le s$, $1 \le i_{j+1} - i_j \le d$ for $j = 1, \dots, r-1$ such, that $x_j = y_{i_j}$, $j = 1, \dots, s$. If for given p, q and d there exist several such sequences, then the sequence belonging to p, q and d is the lexicographically minimal one of such sequences.

DEFINITION 2 [4]. For $p \in X^+$ the *d-complexity* $K_d(p)$ of p is defined as

$$K_d(p) = \sum_{i=1}^{L(p)} f(p, i, d),$$

where $f(p, i, d) = \mid S(p, i, d) \mid$, $S(p, i, d) = S(p, d) \bigcap X^i$ for $i = 1, \dots, L(p)$ and $S(p, d) = \{q \mid q \subset_d p\}$.

EXAMPLE 1. Let X be the English alphabet, $p = ELTE$, then $S(p,1,1) = S(p,1,2) = S(p,1,3) = \{E, L, T\}$, $S(p,2,1) = \{EL, LT, TE\}$, $S(p,2,2) = \{EL, ET, LT, LE, TE\}$, $S(p,2,3) = \{EL, ET, EE, LT, LE, TE\}$, $S(p,3,1) = \{ELT, LTE\}$, $S(p,3,2) = S(p,3,3) = \{ELT, ELE, ETE, LTE\}$, $S(p,4,1) = S(p,4,2) = S(p,4,3) = \{ELTE\}$ and $K_1(p) = 3+3+2+1 = 9$, $K_2(p) = 3+5+4+1 = 13$, $K_3(p) = 3+6+4+1 = 14$.

DEFINITION 3 [4]. The *divided, modified* and *normalized d-complexities* $D_d(p)$, $M_d(p)$ and $N_d(p)$ are defined by

$$D_d(p) = \frac{K_d(p)}{L(p)}, \quad M_d(p) = \frac{L(p) \cdot K_d(p)}{\max\{K_d(q) \mid L(q) = L(p)\}},$$

$$N_d(p) = \frac{K_d(p)}{\max\{K_d(q) \mid L(q) = L(p)\}},$$

respectively.

DEFINITION 4 [4]. A complexity measure $G(p)$ is said to be *monotonically increasing (decreasing)* iff $G(px) \geq G(p)$ $(G(px) \leq G(p))$ for any $p \in X^+$ and $x \in X$. $G(p)$ is said to be *strictly monotonically increasing (decreasing)*, iff $G(px) > G(p)$ $(G(px) < G(p))$ holds for any $p \in X^+$ and $x \in X$.

DEFINITION 5 [4]. A complexity measure $G(p)$ is said to be *subadditive (supadditive)*, iff $G(pq) \leq G(p) + G(q)$ $(G(pq) \geq G(p) + G(q))$ for any pair of words $p, q \in X^+$, and is said to be additive, iff $G(pq) = G(p) + G(q)$ for any $p, q \in X^+$.

DEFINITION 6 [4]. For the complexity measure $G(p)$ and words $p, q \in X^+$ the *complexity ratio* $R(G, p, q)$ is defined by

$$R(G,p,q) = \frac{G(p,q)}{G(p) + G(q)}.$$

DEFINITION 7. Let d be a positive integer, $p \in X^+$ and $q \subset_d p$. If q occurs in p several times, then we consider – according to Definition 1 – the first occurence of q. Let

$$Q_{j,d}(p) = \{q \mid q \subset_d p,\ q = x_{i_1} \ldots x_{i_r} \text{ with } i_r = j\}$$

and

$$a_{j,d}(p) = | Q_{j,d}(p) | \text{ for } j = 1, \ldots, L(p),$$

$$a_{j,d}(p) = 0 \text{ for } j = -(d-1), \ -(d-2), \ldots, -1, \ 0.$$

DEFINITION 8. Let $d \geq 2$, $S_d(z) = z^d - z^{d-1} - \ldots - z - 1$ and $z_{i,d}$ $(i = 1, \ldots, d)$ denote the roots of the equation $S_d(z) = 0$, where $| z_{1,d} | \geq \ldots \geq | z_{d,d} |$ and $| z_{j,d} | = | z_{j+1,d} |$ implies $\arg(z_{j,d}) \leq \arg(z_{j+1,d})$ $(j = 1, \ldots, d-1)$.

2. Analysis of 1-complexity

Some basic features of $K_1(p)$ are analysed in [5], therefore here we only formulate its bounds, which are needed in the next part, and summarize the basic results without proofs.

Lemma 1 [5]. *For any $k \geq 1$ and $p \in X^k$ hold*

$$k \leq K_1(p) \leq 0,5k(k+1).$$

The lower bound is tight. If $n \geq k$, then the upper bound is also tight.

The following tables contain monotonicity and additivity features (Table 1), complexity bounds for nonempty words (Table 2) and complexity bounds for the words of length k (Table 3).

Table 1. Monotonicity and additivity of some complexity measures

Complexity measure G	Strictly monotone	Monotone	Additive	Sub-additive
L	yes	yes	yes	yes
K_1	yes	yes	no	yes
D_1	no	yes	no	no
M_1	no	yes	no	no
N_1	no	no	no	no

Table 2. Tight complexity bounds for nonempty words $p, q \in X^+$

Complexity measure G	Lower bound for $G(p)$	Upper bound for $G(p)$	Lower bound for $R(G,p,q)$	Upper bound for $R(G,p,q)$
L	1	∞	1	1
K_1	1	∞	1	∞
D_1	1	∞	0,5	∞
M_1	1	∞	0,5	∞
N_1	0	1	0	∞

Table 3. Tight complexity bounds for k-length words $p, q \in X^k$

Complexity measure G	Lower bound for $G(p)$	Upper bound for $G(p)$	Lower bound for $R(G,p,q)$	Upper bound for $R(G,p,q)$
L	k	k	1	1
K_1	k	0,5k(k+1)	1	0,5(k+2)
D_1	1	0,5(k+1)	0,5	0,25(k+2)
M_1	1	k	0,5	0,5(k+1)
N_1	1/k	1	0,25	0,25(k+1)

3. Existence of supercomplex words

Using n letters we can assemble n^i different words of length i, and $L(p)-i+1$ words of length i can appear in a word of length $L(p)$, therefore

$$L(p) \leq K_1(p) \leq \sum_{i=1}^{L(p)} \min\left(n^i,\ L(p)-i+1\right).$$

A. Benczur asked, whether there exists an infinite word $p = x_1 x_2 \ldots$ with

$$K_1(x_1 \ldots x_k) = \sum_{i=1}^{k} \min(n^i,\ k-i+1) \quad (k = 1, 2, \ldots),$$

that is a word, whose prefixes have maximum possible 1-complexity. Such words (infinite and finite ones too) are called *supercomplex.*

If we try to construct a supercomplex word over the alphabet $X = \{A, B\}$, then we get Figure 1. In this figure the symbol ∇ means that the given prefix cannot be continued preserving the supercomplexity. The longest supercomplex binary word consists of 9 letters.

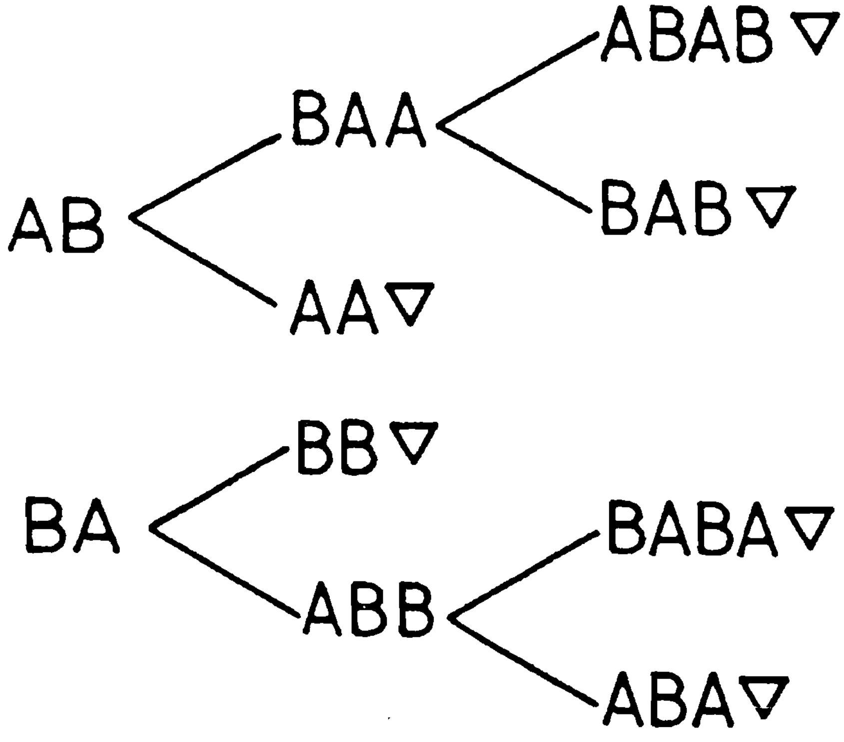

Fig. 1 Supercomplex words for $X = \{A, B\}$

If $n \geq 3$, then the answer is affirmative. To prove this fact we need some preparation.

For given n and k the graph $B(n, k)$ (the so called *de Bruijn graph*) is defined as follows. Its vertex set is X^k and its edge set is X^{k+1} in such a way that a word $p = x_1 \dots x_{k+1}$ determines an edge going from the vertex $x_1 \dots x_k$ to the vertex $x_2 \dots x_{k+1}$.

If $m \geq k$, then any word $q = y_1 \dots y_m$ determines a directed path in $B(n,k)$, which begins at the vertex $y_1 \dots y_k$, goes through the vertices $y_2 \dots y_{k+1}, \dots, y_{m-k} \dots y_{m-1}$, and ends at the vertex $y_{m-k+1} \dots y_m$.

It is known that the graphs $B(n,k)$ contain an Eulerian circuit and a Hamiltonian circuit too. If p determines a Hamiltonian circuit of $B(n,k)$, then $L(p) = k + n^k$. If k corresponds to an Eulerian circuit of $B(n,k)$, then $L(q) = k + n^{k+1}$. The following correspondence between these circuits also is known.

Lemma 2 [6]. *If $k \geq 1$, $n \geq 2$, $m = k + n^k$, then $p = x_{i_1} \dots x_{i_m}$ determines an Eulerian circuit of $B(n,k)$ iff $q = x_{i_1} \dots x_{i_m} x_{i_{k+1}}$ determines a Hamiltonian circuit in $B(n, k+1)$.*

Another useful feature of $B(n,k)$ is the following.

Lemma 3 [7]. *If $n \geq 3$, $k \geq 1$ and p determines a Hamiltonian circuit of $B(n,k)$, then p can be continued in order to get a word q, which determines an Eulerian circuit of $B(n,k)$.*

It is worth to remark that this assertion can be formulated also as follows: if $n \geq 3$ and $k \geq 1$, then after removing the edges of a Hamiltonian circuit of $B(n,k)$ the remaining partial graph is connected.

In [7] a computer program running on TPA-1140 is described. This program during 30 seconds produced the word $p =$

=0122002110001011120222121020100110102100200002201121111
1022101200121221122220202120120101000001110010021010121
1000201100221101102001020200202200012020121102111012211
0012220011200021211201111121011220102202102122022120211
1212120022222111221222102220122011011101001010102000101
0011000010211010110011110002100002001102101101201102201
0100120000002111111011201000221000121010020101121001021
2111002020102011102012002002102002200201221100121110211

2000112120102121002112001012021001122101022210022020000
2021012112102102201002220000122020112020012012102221211
0021202020221020211020222021112002122102112200022221102
2220022122001221200111211012120220120221120212121012222
1121122111120122221011112202122201112220220222212011222
2122212122012121112122222201012220

for the case $X = \{0,1,2\}$ and $L(p) = 734$. This word determines an Eulerian circuit of $B(3,5)$ and is supercomplex.

This example shows an interesting consequence of the definition of supercomplexity: for any fixed r the prefix of length $r + n^r - 1$ of a supercomplex word contains, as subword, all elements of X^r precisely once.

We remark that in [8] the maximum number of edge-disjoint Hamiltonian circuits of $B(n,k)$ is studied: for some special cases we were able to show that if p determines a Hamiltonian circuit of $B(n,k)$, then p can be continued in order to get a word q, determining $(n-1)$ edge-disjoint Hamiltonian circuits of $B(n,k)$.

But for the general case $n \geq 3$ we can prove only the following weaker assertion.

Theorem 1. *If $n \geq 3$, then there exists an infinite supercomplex word over $X = \{A_1, \ldots, A_n\}$.*

Proof. We give a constructive proof. Let us consider a Hamiltonian circuit of $B(n,1)$, e.g. the circuit given by the word $A_1 A_2 \ldots A_n A_1$. According to Lemma 3 we can continue p in order to get an Eulerian circuit of $B(n,1)$, e.g. $q = A_1 \ldots A_n A_1 A_1 A_n A_n A_{n-1} A_{n-1} \ldots A_2 A_2 A_1$ gives an Eulerian circuit in $B(n,1)$. According to Lemma 2, $q' = qA_2$ determines a Hamiltonian circuit of $B(n,2)$.

By induction we get the existence of an infinite supercomplex word. $\square$

4. Analysis of d–complexity

At first we give lower and upper bounds for $K_d(p)$.

Lemma 4 [5]. *If $n \geq 2$, $k \geq 1$, $d \geq 1$ and $p \in X^k$, then*

$$k \leq K_d(p) \leq 2^k - 1.$$

The lower bound is tight. For $d \geq k-1$ and $n \geq k$ the upper bound is also tight.

Let us consider now an infinite alphabet $X = \{A_1, A_2, \ldots\}$. The complexity $K_d(p)$ of the word $p = A_1 A_2 \ldots A_k$ (or any other k-length word consisting of different letters) is denoted by $N(k,d)$ and is called *maximal.*

According to Lemmas 1 and 4 we have $N(k,1) = 0,5k(k+1)$ and $N(k,k-1) = 2^k - 1$. In which manner does a quadratic polynomial changes into an exponential function when d increases?

In Definition 7 we have classified the d-subwords of a given word p according to the position of their last letter. Among the cardinalities of the sets $Q_{j,d}(p)$ L. Hunyadvári has found the following reccurent connection.

Lemma 5 [9]. *If $k \geq 1$, $p \in X^k$ and $K_1(p) = N(k,1)$, then*

$$a_{j,d}(p) = 1 + a_{j-1,d}(p) + a_{j-2,d}(p) + \cdots + a_{j-d,d}(p) \quad \text{for } j = 1, \ldots, k. \tag{1}$$

Proof [9]. Among the elements of $Q_{j,d}(p)$ there exists an element with unit length. The remaining elements consist of two or more letters, and their last but one letters can be located in the $(j-1)$-th,..., $(j-d)$-th positions. □

The next assertion gives the explicit form of the cardinalities $a_{j,d}$.

Lemma 6. *If $k \geq 1$, $d \geq 2$, $p \in X^k$ and $K(p) = N(k,d)$, then*

$$a_{j,d} = \frac{1}{1-d} + \sum_{i=1}^{d} k_{i,d} z_{i,d}^j \quad (j = 1, \ldots, k)$$

and

$$\mathcal{N}(k,d) = \frac{k}{1-d} + \sum_{i=1}^{d} k_{i,d} z_{i,d} \frac{z_{i,d}^k - 1}{z_{i,d} - 1},$$

where the coefficients $k_{i,d}$ $(i = 1, \ldots, d)$ are constants.

Proof. The general solution of an inhomogeneous recurrent relation equals to the sum of the general solution of the corresponding homogeneous equation and an arbitrary particular solution of the inhomogeneous one [10].

Let us suppose that $a_{j,d} = z^j$ for a suitable z. Then from (1) we get $S_d(z) = 0$, and so the general solution of the homogeneous equation has the form

$$a_{j,d} = k_{1,d} z_{1,d}^j + \cdots k_{d,d}^j \; z_{d,d}^j$$

where the constants $k_{i,d}$ $(i = 1, \ldots, d)$ are determined by the initial conditions.

Supposing $a_{j,d} = c$ for $j = -(d-1), -(d-2), \ldots, -1, 0, 1, \ldots, L(p)$ we get a particular solution of the inhomogeneous equation: if $d \geq 2$, then $c = 1/(1-d)$, which finishes the proof. □

The following lemma formulates an important property of the roots of $S_d(z)$.

Lemma 7. *If $d \geq 2$, then the equation $S_d(z) = 0$ has precisely one root $z_{1,d} > 1$. For the remaining roots $z_{i,d}$ $(i = 2, \ldots, d)$ we have $\mid z_{i,d} \mid < 1$.*

The following proof is due to Imre Kátai.

Proof [11]. a/ Due to $S_d(1) = -(d-1) < 0$ and $S_d(2) = 1 > 0$ we get $z_{1,d} > 1$.

b/ It is known [12], that if $m > 0$ is an integer number and $r_0 > r_1 > \cdots > r_m > 0$ are real numbers, then for any root y of the equation

$$r_0 + r_1 x + \cdots + r_m x^m = 0 \tag{2}$$

we have $|y|>1$.

c/ Since $S_d(0)\neq 0$, substituting $1/w$ for z and multiplying by $(-w_d)$ we change $S_d(z)$ into

$$T_d(w)=w^d+w^{d-1}+\cdots+w-1,$$

whose roots are the reciprocals of the roots of $S_d(z)$. Dividing $T_d(w)$ by $(w-w_1)$, we get

$$R_d(w)=\frac{T_d(w)}{w-w_1}=w^{d-1}+w^{d-2}(1+w_1)+$$
$$w^{d-3}(1+w_1+w_1^2)+\cdots+(1+w_1+\cdots+w_1^{d-1}).$$

If $z_{1,d}>1$, then $w_{1,d}=1/z_{1,d}\in(0,1)$, and the coefficients of $R_d(w)$ satisfy the conditions of the assertion, mentioned in part b/ of this proof. Therefore the roots of $R_d(w)$ are outside the unit circle, and so the roots of $S_d(z)$ – in except of $z_{1,d}$ – are inside the unit circle. □

Now we can formulate the main result of this paper.

Theorem 2. *If $d\geq 2$, then*

$$N(k,d)=\frac{k_{1,d}z_{1,d}}{z_{1,d}-1}z_{1,d}^k+\frac{k}{1-d}+\sum_{i=1}^{d}\frac{k_{i,d}z_{i,d}}{1-z_{i,d}}+\sum_{j=2}^{d}\frac{k_{j,d}z_{j,d}}{z_{j,d}-1}z_{j,d}^k$$

and so

$$\lim_{k\to\infty}\left(\frac{k_{1,d}z_{1,d}}{z_{1,d}-1}z_{1,d}^k+\frac{k}{1-d}+\sum_{i=1}^{d}\frac{k_{i,d}z_{i,d}}{1-z_{i,d}}-N(k,d)\right)=0.$$

Proof. We get the assertion from the expression of $N(k,d)$ in Lemma 6 using our knowledge about the roots of $S_d(z)$ formulated in Lemma 7. □

EXAMPLE 2. If $d = 2$, then $S_2(z) = z^2 - z - 1 = 0$ has the roots $z_{1,2} = \frac{1}{2}(1+\sqrt{5}) \approx 1,618034$ and $z_{2,2} = \frac{1}{2}(1-\sqrt{5}) \approx -0,618034$, therefore

$$a_{j,2} = k_{1,2}\left(\frac{1+\sqrt{5}}{2}\right)^j + k_{2,2}\left(\frac{1-\sqrt{5}}{2}\right)^j - 1 \qquad (j = 1,2).$$

Taking into account that $a_{1,2} = 1$ and $a_{2,2} = 2$, for the constants $k_{1,2}$ and $k_{2,2}$, we have the system of linear equations

$$2 = k_{1,2}\left(0,5 + \sqrt{1,25}\right) + k_{2,2}\left(0,5 - \sqrt{1,25}\right),$$
$$3 = k_{1,2}\left(1,5 + \sqrt{1,25}\right) + k_{2,2}\left(1,5 - \sqrt{1,25}\right),$$

from where $k_{1,2} = 0,5 + 0,3\sqrt{5} \approx 1,170820$ and $k_{2,2} = 0,5 - 0,3\sqrt{5} \approx -0,170820$. Substituting the constants and the roots into the formula of Lemma 6 we get

$$N(k,2) = (1,5 + 0,7\sqrt{5})(0,5 + 0,5\sqrt{5})^k +$$
$$+(1,5 - 0,7\sqrt{5})(0,5 - 0,5\sqrt{5})^k - k - 3 \approx 3,065247.1,618034^k -$$
$$-0,065247(-0,618034)^k - k - 3,$$

and so

$$\lim_{k\to\infty}\left[N(k,2) - \left((1,5 + 0,7\sqrt{5})(0,5 + 0,5\sqrt{5})^k - k - 3\right)\right] = 0.$$

If $d = 3$, then the roots are

$$z_{1,3} = \frac{1}{3}\left(1 + \sqrt[3]{19 + 3\sqrt{33}} + \sqrt[3]{19 - 3\sqrt{33}}\right) \approx 1,839287,$$

$$z_{2,3} = \frac{1}{6}\left(2 - \sqrt[3]{19+3\sqrt{33}} - \sqrt[3]{19-3\sqrt{33}}\right) +$$

$$+i\frac{\sqrt{3}}{6}\left(\sqrt[3]{19+3\sqrt{33}} - \sqrt[3]{19+3\sqrt{33}}\right) \approx -0,419643+0,606291i,$$

$$z_{3,3} = \frac{1}{6}\left(2 - \sqrt[3]{19+3\sqrt{33}} - \sqrt[3]{19-3\sqrt{33}}\right) -$$

$$-i\frac{\sqrt{3}}{6}\left(\sqrt[3]{19+3\sqrt{33}} - \sqrt[3]{19+3\sqrt{33}}\right) \approx -0,419643-0,606291i,$$

$$k_{1,3} \approx 0,736840, \quad k_{2,3} \approx -0,118420 - 0,037401i,$$

$$k_{3,3} \approx -0,118420 + 0,037401i,$$

and

$$N(k,3) \approx 1,614776 \cdot 1,839287^k - \frac{k}{2} - \frac{3}{2} +$$
$$+0,737353^k \cdot \big[0,061034 \cos(2,176234(k+1)) -$$
$$-0,052411 \sin(2,176234(k+1))\big].$$

5. Estimation of the most significant root

If $d \geq 2$, then multiplying $S_d(x)$ by $(x-1)$ we get

$$W_d(x) = x^{d+1} - 2x^d + 1.$$

By analysing of $W_d(x)$ using its derivates $W_d'(x)$ and $W_d''(x)$ we obtain Figure 2 (for even d) and Figure 3 (for odd d).

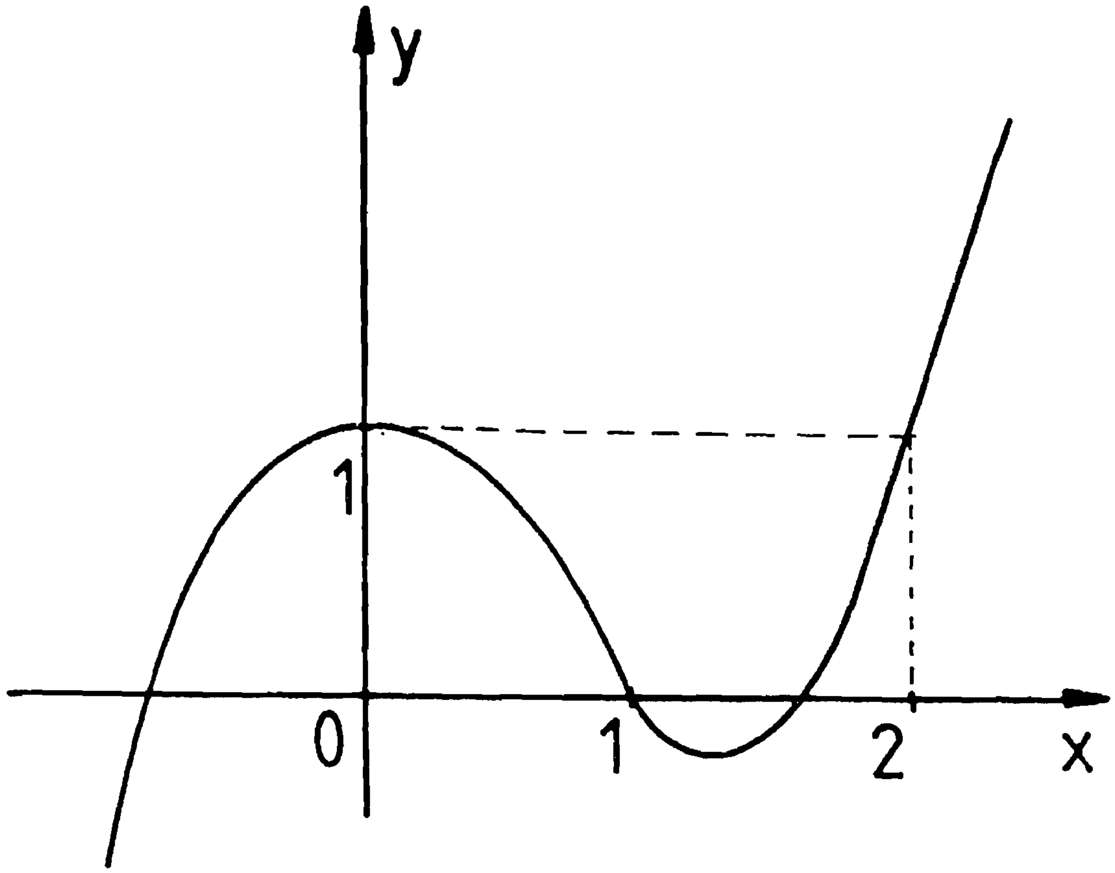

Figure 2. The plot of $y = x^{d+1} - 2x^d + 1$ for even d

According to Lemma 7 the equation $S_d(x) = 0$ has only one root $z_{1,d}$ outside the unit circle. Because of $S_d(1) = -(d-1)$ and $S_d(2) = 1$ we have $z_{1,d} \in (1,2)$.

Lemma 8. *If $d \geq 2$ then*

$$z_{chord,d} = 2 - (0,5 + \frac{1}{2d})^d < z_{1,d} < 2 - \frac{1}{2^d} = z_{tan,d}. \square$$

Proof. The function $W_d(x)$ has a local minimum at $x_0 = 2 - 2/(d+1)$. Since $W_d(x)$ is convex in the interval $(x_0, 2)$, we can give an upper bound on $z_{1,d}$ using the tangent to the curve at $x = 2$ and a lower bound using the chord belonging to the points of the curve at x_0 and $x = 2$ [13].

Since $W_d(2) = 2^d$, the equation of the tangent is $y = 2^d(x - 2) + 1$, from where we get the value $z_{tan,d} = 2 - 1/2^d$.

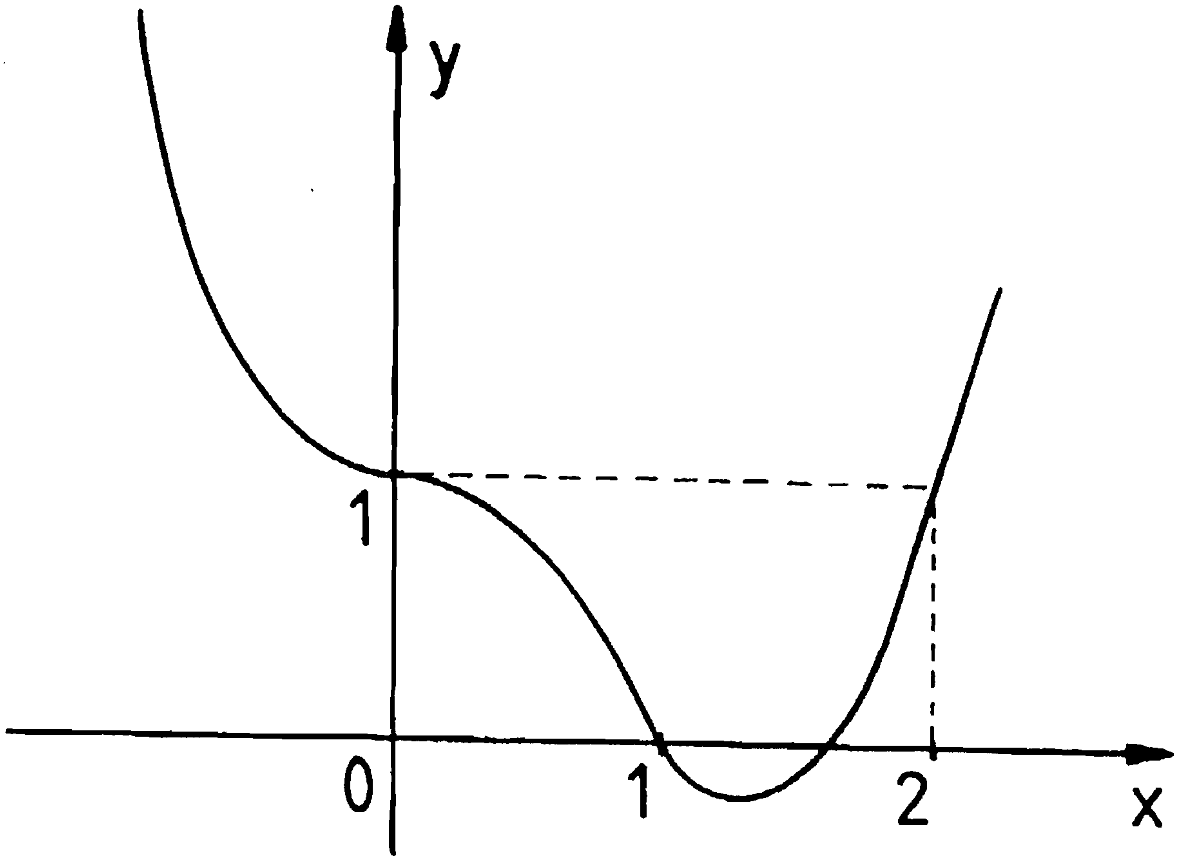

Figure 3. The plot of $y = x^{d+1} - 2x^d + 1$ for odd d

Using

$$W_d(2) = 1, \; W_d(x_0) = (2 - \frac{2}{d+1})^d + 1 - 2(2 - \frac{2}{d+1}) + 1$$

and the formula

$$y - y_1 = \frac{y_2 - y_1}{x_2 - x_1}(x - x_1)$$

we obtain the equation of the chord and the value

$$z_{chord,d} = 2 - \frac{1}{2^d(2d+1)}. \; \square$$

The following estimations are due to Keresztély Corrádi.

Lemma 9 [14]. *If $d \geq 2$ then*

$$L_{CK,d} = 2 - \frac{1}{2^{d-1}} < z_{1,d} < 2 - \frac{1}{2^d} = U_{CK,d} = z_{tan,d}.$$

Proof [14]. a) At first we show that $W_d(L_{CK,d}) < 0$. Using the well-known inequality

$$\sqrt[m]{\prod_{i=1}^{m} a_i} \leq \frac{1}{m}\sum_{i=1}^{m} a_i$$

between the geometric and arithmetic means of nonnegative numbers for $a_j = 1 - 1/2^d$ $(j = 1, \dots, 2^d)$ and $a_j = 1$ $(j = 2^d + 1, \dots, 2^{d+1})$ we get

$$(1 - \frac{1}{2^d})^{2^d} \leq (1 - \frac{1}{2^{d+1}})^{2^{d+1}}, \tag{3}$$

i.e. $(1 - 1/2^d)^{2^d}$ is an increasing function of d.

If $d \geq 2$ then $2^d \geq 2d$ therefore

$$(1 - \frac{1}{2^d})^{2d} \geq (1 - \frac{1}{2^d})^{2^d}. \tag{4}$$

From (4), taking into account (3)

$$(1 - \frac{1}{2^d})^{2d} \geq (1 - \frac{1}{2^d})^{2^d} \geq (1 - \frac{1}{2^2})^4,$$

and so extracting quadratic root we have

$$(1 - \frac{1}{2^d})^{d} \geq \frac{9}{16} > \frac{1}{2}. \tag{5}$$

Since $W_d(x) = x^d(x-2)+1$ and

$$W_d(L_{CK,d}) = (2 - \frac{1}{2^{d-1}})^d(-\frac{1}{2^{d-1}}) + 1,$$

$W_d(L_{CK,d}) < 0$ is equivalent to (5).

b) For $U_{CK,d}$ we have

$$W_d(U_{CK,d}) = (2 - \frac{1}{2^d})^d(-\frac{1}{2^d}) + 1 = 1 - (1 - \frac{1}{2^{d+1}})^d > 0. \ \square$$

We remark, that using a similar argumentation we can show

$$z_{1,d} > 2 - \frac{4}{5}\frac{1}{2^{d-1}}$$

for $d \geq 2$ and

$$z_{1,d} > 2 - \frac{2}{3}\frac{1}{2^{d-1}}$$

for $d \geq 3$.

Combining the ideas of the last two lemmas we get the following estimations.

Lemma 10. *If $d \geq 2$, then*

$$2 - \frac{1}{2^d} - \frac{1 - (1 - \frac{1}{2^{d+1}})^d}{2^d\, 2(1 - \frac{1}{2^d})^d - (1 - \frac{1}{2^{d+1}})^d} = L_d < z_{1,d} <$$

$$< 2 - \frac{1}{2^d} - \frac{1 - (1 - \frac{1}{2^{d+1}})^d}{(2 - \frac{1}{2^d})^{d-1}(2 - \frac{1+d}{2^d})} = U_d.$$

Proof. Using the values $W_d(U_{CK,d})$ and $W_d'(U_{CK,d})$ we get the equation of the tangent to the curve at $x = U_{CK,d}$

$$y - 1 + (2 - \frac{1}{2^d})^d \frac{d}{2^d} = (2 - \frac{1}{2^d})^{d-1}(2 - \frac{1+d}{2^d})(x - 2 + \frac{1}{2^d}),$$

from where the expression for U_d follows.

Using the values $U_{CK,d}$, $L_{CK,d}$, $W_d(U_{CK,d})$, $W_d(L_{CK,d})$ we get the equation of the chord belonging to the points at $x = L_{CK,d}$ and $x = U_{CK,d}$:

$$y - 1 + (1 - \frac{1}{2^{d+1}})^d = -2^d(1 - \frac{1}{2^{d+1}})^d - 2(1 - \frac{1}{2^d})^d x - 2 + \frac{1}{2^d},$$

from where the expression for L_d follows. □

The following table shows some numerical values. The roots $z_{1,d}$ are computed using Newton's method [13]. As initial value we used $U_{CK,d}$. The accuracy was $\epsilon = 10^{-8}$ in all cases.The number of necessary iteration steps is denoted by M_d. Table 4 contains the values $z_{chord,d}$, $L_{CK,d}$, $L_d, z_{1,d}$, U_d, $U_{CK,d} = z_{tan,d}$ and M_d for $d = 2, \ldots 10$.

Table 4. Approximate values of $z_{1,d}$

d	$z_{chord,d}$	$L_{CK,d}$	L_d	$z_{1,d}$	U_d	$U_{CK,d}$	M_d
2	1,43750	1,50000	1,5869565	1,6180340	1,6428571	1,75000	17
3	1,70370	1,75000	1,8323474	1,8392868	1,8416204	1,87500	9
4	1,84741	1,87500	1,9262779	1,9275620	1,9277830	1,93750	4
5	1,92224	1,93750	1,9657246	1,9659482	1,9659691	1,96875	3
6	1,96060	1,96875	1,9835452	1,9835828	1,9835848	1,98438	2
7	1,98011	1,98438	1,9919581	1,9919642	1,9919644	1,99219	1
8	1,98998	1,99219	1,9960302	1,9960312	1,9960312	1,99609	1
9	1,99496	1,99609	1,9980293	1,9980295	1,9980295	1,99805	1
10	1,99747	1,99805	1,9990186	1,9990186	1,9990186	1,99902	1

6. Computing d-complexity

Using Lemma 5 $N(k,d)$ is computable in $O(k)$ time. Using Theorem 2 we can get different approximations of $N(k,d)$.

Let

$$f_1(k,d) = \frac{k_{1,d}}{z_{1,d}-1} z_{1,d}^{k+1}, \quad f_2(k,d) = f_1(k,d) + \frac{k}{1-d},$$

$$f_3(k,d) = f_2(k,d) + \sum_{i=1}^{d} \frac{k_{i,d} z_{i,d}}{1-z_{i,d}},$$

$$f_4(k,d) = f_3(k,d) + \sum_{j=2}^{d} \frac{k_{j,d}}{z_{j,d}-1} z_{j,d}^{k+1}.$$

Table 5. 2-complexity and its approximations

k	$f_1(k,2)$	$f_3(k,2)$	$N(k,2)$
1	4,9597	0,9597	1
2	8,0249	3,0249	3
3	12,9846	6,9846	7
4	21,0095	14,0095	14
5	33,9941	25,9941	26
6	55,0036	46,0036	46
7	88,9977	78,9977	79
8	144,0014	133,0014	133
9	232,9991	220,9991	221
10	377,0005	364,0005	364
11	609,9997	595,9997	596
12	987,0002	972,0002	972
13	1596,9999	1580,9999	1581
14	2584,0001	2567,0001	2567
15	4180,99999	4162,99999	4163

Then

$$N(k,d) - f_1(k,d) = O(k), \ N(k,d) - f_2(k,d) = O(1),$$
$$N(k,d) - f_3(k,d) = o(1),$$

and

$$N(k,d) = f_4(k,d),$$

so we can estimate $N(k,d)$ with accuracy $O(k)$ or $O(1)$ in $O(1)$ time, with accuracy $o(1)$ in $O(d)$ time and can get the precise value of $N(k,d)$ also in $O(d)$ time units (of course, only if we know the values of the roots and coefficients).

Table 6. 3-complexity and its approximations

k	$f_1(k,3)$	$f_3(k,3)$	$N(k,3)$
1	2,9700	0,9700	1
2	5,4627	2,9627	3
3	10,0476	7,0476	7
4	18,4803	14,9803	15
5	33,9906	29,9906	30
6	62,5185	58,0185	58
7	114,9895	109,9895	110
8	211,4987	205,9987	201
9	389,0068	383,0068	383
10	715,4950	708,9950	709
11	1316,0005	1309,0005	1309
12	2420,5023	2413,0023	2413
13	4451,9978	4443,9978	4444
14	8188,5006	8180,0006	8180
15	15061,0007	15052,0007	15052

The results of the computations for $d = 2$ and $d = 3$, $k = 1, \ldots, 15$ are summarized in Table 5 and Table 6, where

$$f_1(k,2) = 3,065247 \cdot 1,618034^k, \quad f_3(k,2) = f_1(k,2) - k - 3,$$
$$N(k,2) = f_3(k,2) - 0,065247 \cdot (-0,618034)^k,$$
$$f_1(k,3) = 1,614776 \cdot 1,839287^k, \quad f_3(k,3) = f_1(k,3) - \frac{k}{2} - \frac{3}{2},$$

$$N(k,3) = f(k,3)+2\cdot 0,737353^{k+1}[0,061034\cos(2,176234(k+1))- \\ -0,052411\sin(2,176234(k+1))].$$

We are indebted to the colleagues mentioned in the text for proving Lemmas 5, 7 and 9 and also to András Benczur and Péter Simon for their useful critical remarks.

References

[1] EBELING, W. und FEISTEL, R., *Physik der Selbstorganisation und Evolution.* Akademie-Verlag, Berlin, 1982.

[2] SALOMAA, A., *Formal Languages.* Academic Press, New York–London, 1973.

[3] BERGE, C., *Graphs and Hypergraphs.* North-Holland Publ. Co., Amsterdam–London, 1973.

[4] HUNYADVÁRI, L., and IVÁNYI, A., On some complexity measures of words. In: *Automata, Languages and Mathematical Systems* (Ed. I. Peák). Karl Marx Univ. of Economics, Budapest, 1984, 67-82.

[5] HUNYADVÁRI, L., and IVÁNYI, A., On the subsequence complexity of words. In: *Conference of Young Programmers and Mathematicians.* Eötvös Loránd Univ., Budapest, 1984, 7-16.

[6] LOVÁSZ, L., *Combinatorial Problems and Exercises.* Akadémiai Kiadó, Budapest, 1979.

[7] VÖRÖS, N., On the complexity of symbol-sequences. In: *Conference of Young Programmers and Mathematicians.* Eötvös Loránd Univ., Budapest, 1984, 43-50.

[8] BOND, J., and IVÁNYI, A., Modelling of interconnection networks using de Bruijn graphs. In: *Third Conference of Program Designers.* Eötvös Loránd Univ., Budapest, 1987, 75-88.

[9] HUNYADVÁRI, L., *Private communication.* Budapest, 1984.

[10] GREENE, D. H., and KNUTH, D., *Mathematics for the Analysis of Algorithms.* Birkhäuser, Boston–Basel–Stuttgart, 1981.

[11] KÁTAI, I., *Private communication.* Budapest, 1984.

[12] PÓLYA, G. und SZEGŐ, G., *Aufgaben und Lehrsätze aus der Analysis.* Band I. Springer–Verlag, Berlin, 1925.

[13] KÁTAI, I., *Introduction into the Numerical Analysis* /in Hungarian/. Tankönyvkiadó, Budapest, 1975.

[14] CORRÁDI, K., *Private communication.* Budapest, 1985.

(Received December 12, 1987)

ANTAL IVÁNYI
Dept. of General Computer Science
H-1088 Budapest, Múzeum krt. 6-8.

HUNGARY