Maximal right smooth extension chains

Yun Bao Huang
Department of Mathematics
Hangzhou Normal University
Xiasha Economic Development Area
Hangzhou, Zhejiang 310036, China
huangyunbao@sina.com
huangyunbao@gmail.com

2010.12.26

Abstract. If $w = u\alpha$ for $\alpha \in \Sigma = \{1, 2\}$ and $u \in \Sigma^*$, then w is said to be a *simple right extension* of u and denoted by $u \prec w$. Let k be a positive integer and $P^k(\varepsilon)$ denote the set of all C^∞-words of height k. Set $u_1, u_2, \cdots, u_m \in P^k(\varepsilon)$, if $u_1 \prec u_2 \prec \cdots \prec u_m$ and there is no element v of $P^k(\varepsilon)$ such that $v \prec u_1$ or $u_m \prec v$, then $u_1 \prec u_2 \prec \cdots \prec u_m$ is said to be a *maximal right smooth extension (MRSE) chains* of height k. In this paper, we show that *MRSE* chains of height k constitutes a partition of smooth words of height k and give the formula of the number of *MRSE* chains of height k for each positive integer k. Moreover, since there exist the minimal height h_1 and maximal height h_2 of smooth words of length n for each positive integer n, we find that *MRSE* chains of heights h_1-1 and h_2+1 are good candidates to be used to establish the lower and upper bounds of the number of smooth words of length n respectively, which is simpler and more intuitionistic than the previous methods.

Keywords: smooth word; primitive; height; *MRSE* chain.

1. Introduction

Let $\Sigma = \{1, 2\}$, Σ^* denotes the free monoid over Σ with ε as the empty word. If $w = w_1 w_2 \cdots w_n$, $w_i \in \Sigma$ for $i = 1, 2, \cdots, n$, then n is called the length of the word w and denoted by $|w|$. For $i = 1$, 2, let $|w|_i$ be the number of i which occurs in w, then $|w| = |w|_1 + |w|_2$.

Given a word $w \in \Sigma^*$, a *factor* or *subword* u of w is a word $u \in \Sigma^*$ such that $w = xuy$ for $x, y \in \Sigma^*$, if $x = \varepsilon$, then u is said to be a *prefix* of w. A *run* or *block* is a maximal factor of the form $u = \alpha^k, \alpha \in \Sigma$. The *complement* of $u = u_1 u_2 \cdots u_n \in \Sigma^*$ is the word $\bar{u} = \bar{u}_1 \bar{u}_2 \cdots \bar{u}_n$, where $\bar{1} = 2, \bar{2} = 1$.

The Kolakoski sequence K which Kolakoski introduced in [13], is the infinite sequence over the alphabet Σ

$$K = \underbrace{1}_{1}\underbrace{22}_{2}\underbrace{11}_{2}\underbrace{2}_{1}\underbrace{1}_{1}\underbrace{22}_{2}\underbrace{1}_{1}\underbrace{22}_{2}\underbrace{11}_{2}\underbrace{2}_{1}\underbrace{11}_{2}\underbrace{22}_{2}\underbrace{\cdots}_{\cdots = K}$$

which starts with 1 and equals the sequence defined by its run lengths.

I would like to thank Prof. Jeffrey O. Shallit for introducing me the Kolakoski sequence K and eight questions on it in personal communications (Feb. 15, 1990), the fourth and eighth problems of them are respectively as follows:

(1) Prove or disprove: $|K_i|_1 \sim |K_i|_2$, which is almost equivalent to Keane's question.

(2) Prove or disprove: $|K_i| \sim \alpha(3/2)^i$ (This would imply (1)), where α seems to be about 0.873. Does $\alpha = (3 + \sqrt{5})/6$?

where $K_0 = 2$ and define K_{n+1} as the string of $1'$ and $2'$ obtained by using the elements of K_n as replication factors for the appropriate prefix of the infinite sequence $1212\cdots$.

The intriguing Kolakoski sequence K has received a remarkable attention [1, 3, 5, 11, 15, 16]. For exploring two unsolved problems, both wether K is recurrent and whether K is invariant under complement, raised by Kimberling in [12], Dekking proposed the notion of C^∞-*word* in [6]. Chvátal in [4] obtained that the letter frequencies of C^∞-words are between 0.499162 and 0.500838.

We say that a finite word $w \in \Sigma^*$ in which neither 111 or 222 occurs is *differentiable*, and its *derivative*, denoted by $D(w)$, is the word whose jth symbol equals the length of the jth run of w, discarding the first and/or the last run if it has length one.

If a word w is arbitrarily often differentiable, then w is said to be a C^∞-*word* (or *smooth word*) and the set of all C^∞-word is denoted by $\mathcal{C}^\infty$.

A word v such that $D(v) = w$ is said to be a *primitive* of w. Thus 11, 22, 211, 112, 221, 122, 2112, 1221 are the primitives of 2. It is easy to see that for any word

$w \in \mathcal{C}^\infty$, there are at most 8 primitives and the difference of lengths of two primitives of w is at most 2.

The *height* of a nonempty smooth word w is the smallest integer k such that $D^k(w) = \varepsilon$ and the height of the empty word ε is zero. We write $ht(w)$ for the height of w. For example, for the smooth word $w = 12212212$, $D^4(w) = \varepsilon$, so $ht(w)$=4.

2. Maximal right smooth extension chains

Let $\mathcal{N}$ be the set of all positive integers and $P^k(\varepsilon)$ denote the set of all smooth words of height k for $k \in \mathcal{N}$, then

$$\begin{aligned} P(\varepsilon) &= \{1, 2, 12, 21\}, & (1) \\ P^2(\varepsilon) &= \{121, 212, 11, 22, 211, 122, 112, 221, 2112, 1221, 1211, 12112, \\ &\quad 2122, 21221, 1121, 21121, 2212, 12212\}. & (2) \end{aligned}$$

Definition 1. *Let w, u, $v \in \Sigma^*$ if $w = uv$, then w is said to be a right extension of u. Especially, if $v = \alpha \in \Sigma$, then w is said to be a simple right extension of u, and is denoted by $u \prec w$.*

Definition 2. *Let u_1, $u_2, \cdots, u_m \in P^k(\varepsilon)$, where $k \in \mathcal{N}$.*

$$u_1 \prec u_2 \prec \cdots \prec u_m, \tag{3}$$

and there is no element v of $P^k(\varepsilon)$ such that

$$v \prec u_1 \text{ or } u_m \prec v, \tag{4}$$

then (3) is said to be a maximal right smooth extension (MRSE) chain of the height k. Moreover, u_1 and u_m are respectively called the first and last members of the MRSE chain (3).

Let H^k denote the set of all *MRSE* chains of the height k. For $\xi \in H^k$, $\xi = u_1 \prec u_2 \prec \cdots \prec u_m$, the complement of ξ is $\bar{u}_1 \prec \bar{u}_2 \prec \cdots \prec \bar{u}_m$, and is denoted by $\bar{\xi}$. It is clear that $\bar{\xi}$ is also a *MRSE* chain of the height k. In addition, for $A \subseteq H^k$, $\bar{A} = \{\bar{\xi} : \ \xi \in A\}$.

Definition 3. *For $\xi \in H^{k+1}$, $\xi = u_1 \prec u_2 \prec \cdots \prec u_m$, where $k \in \mathcal{N}$. If there is an element $\eta = v_1 \prec v_2 \prec \cdots \prec v_n \in H^k$ such that u_1, $u_2, \cdots, u_m$ are all the primitives of v_1, $v_2, \cdots, v_n$, then ξ is said to be a primitive of η.*

For example, $\xi = 121 \prec 1211 \prec 12112 \in H^2$ is a primitive of $\eta = 1 \prec 12 \in H^1$, $\bar{\xi} = 212 \prec 2122 \prec 21221 \in H^2$.

For a set A, let $|A|$ denote the cardinal number of A. Next we establish the formula of the number of the members of H^k. For this reason, let

$$H_1^k = \{\xi \in H^k : \xi = u_1 \prec u_2 \prec \cdots \prec u_m \textit{ and first}(u_1) = 1\}; \tag{5}$$
$$H_2^k = \{\xi \in H^k : \xi = u_1 \prec u_2 \prec \cdots \prec u_m \textit{ and first}(u_1) = 2\}. \tag{6}$$

It immediately follows that

$$H_1^k = \bar{H}_2^k; \tag{7}$$
$$H^k = H_1^k \cup H_2^k; \tag{8}$$
$$|H_1^k| = |H_2^k|. \tag{9}$$

From (1) and (2) we have

$$\begin{aligned}
H^1 &= \{1 \prec 12, 2 \prec 21\}; && (10)\\
H_1^1 &= \{1 \prec 12\};\\
H_2^1 &= \{2 \prec 21\};\\
H^2 &= \{121 \prec 1211 \prec 12112,\ 212 \prec 2122 \prec 21221,\ 11 \prec 112 \prec 1121,\\
&\quad 22 \prec 221 \prec 2212,\ 211 \prec 2112 \prec 21121,\ 122 \prec 1221 \prec 12212\}; && (11)\\
H_1^2 &= \{121 \prec 1211 \prec 12112,\ 11 \prec 112 \prec 1121,\ 122 \prec 1221 \prec 12212\};\\
H_2^2 &= \{212 \prec 2122 \prec 21221,\ 22 \prec 221 \prec 2212,\ 211 \prec 2112 \prec 21121\}.
\end{aligned}$$

Thus from (10) and (11), ones see that every *MRSE* chain of height k is uniquely determined by its first member u_1 and each member of $P^k(\varepsilon)$ exactly belongs to one *MRSE* chain of height k for $k = 1, 2$ and

$$|H^2| = 3|H^1|. \tag{12}$$

Actually, the above result holds for every $k \in \mathcal{N}$.

Theorem 4. *H^k is stated as above. Then each member of $P^k(\varepsilon)$ exactly belongs to one MRSE chain of height k, that is, H^k gives a partition of the smooth words of height k and*

$$|H^k| = 2 \cdot 3^{k-1} \textit{ for all } k \in \mathcal{N}. \tag{13}$$

Proof. We proceed by induction on k. From (12) it follows that (13) holds for $k = 1, 2$. Assume that (13) holds for $k = n - 1 \geq 1$.

Now we consider the case for $k = n$. Since for each $\eta = u_1 \prec u_2 \prec \cdots \prec u_m \in H_1^{n-1}$, from the definition 2 and (5), we see that $first(u_1) = first(u_2) = \cdots = first(u_m) = 1$, and $u_{i+1} = u_i\alpha$ where $i = 1, 2, \cdots, m-1$, $\alpha = 1, 2$. Thus if $\alpha = 1$ then the two primitives $p(u_{i+1})$ of u_{i+1} are

$$\begin{aligned} p(u_{i+1}) &= \bar{\beta}\Delta_{\beta}^{-1}(u_{i+1})\gamma \\ &= \bar{\beta}\Delta_{\beta}^{-1}(u_i)\bar{\gamma}\gamma \\ &= p(u_i)\gamma,\ \textit{where}\ \beta,\ \gamma \in \Sigma, \end{aligned}$$

so $p(u_i) \prec p(u_{i+1})$.

If $\alpha = 2$ then the four primitives $p_t(u_{i+1})$ of u_{i+1} are

$$\begin{aligned} p_t(u_{i+1}) &= \bar{\beta}\Delta_{\beta}^{-1}(u_{i+1})\gamma^t \\ &= \bar{\beta}\Delta_{\beta}^{-1}(u_i)\bar{\gamma}^2\gamma^t \\ &= p(u_i)\bar{\gamma}\gamma^t,\ \text{where}\ \beta = 1, 2,\ t = 0, 1, \end{aligned}$$

hence $p(u_i) \prec p_0(u_{i+1}) \prec p_1(u_{i+1})$. Therefore, η has exactly two primitives and the primitives of u_1, $u_2, \cdots$, and u_m all occur in the two primitives of η.

For example, $\eta = 121 \prec 1211 \prec 12112 \in H_1^2$ has exactly two primitives:

$\mu = 121121 \prec 1211212 \prec 12112122 \prec 121121221$ and $\bar{\mu}$.

Analogously, we can see that each member η of H_2^{n-1} has exactly four primitives and the primitives of u_1, $u_2, \cdots$, and u_m all occur in the four primitives of η.

For example, $\eta = 212 \prec 2122 \prec 21221 \in H_2^2$ has exactly four primitives:

$\xi_1 = 22122 \prec 221221 \prec 2212211 \prec 22122112 \prec 221221121$;

$\xi_2 = 122122 \prec 1221221 \prec 12212211 \prec 122122112 \prec 221221121$ and $\bar{\xi}_1$, $\bar{\xi}_2$.

Thus, by the induction hypothesis, it follows from (8) and (9) that

$$\begin{aligned} |H^n| &= |H_1^n| + |H_2^n| \\ &= 2 \cdot |H_1^{n-1}| + 4 \cdot |H_2^{n-1}| \\ &= 3 \cdot (|H_1^{n-1}| + |H_2^{n-1}|) \\ &= 3 \cdot |H^{n-1}| \\ &= 2 \cdot 3^{n-1}. \quad \square \end{aligned}$$

3. The number of smooth words of length n

Let $\gamma(n)$ denote the number of smooth words of length n and $p_K(n)$ the number of subwords of length n which occur in K.

Dekking in [6] proved that there is a suitable positive constant c such that $c \cdot n^{2.15} \leq \gamma(n) \leq n^{7.2}$ and brought forward the conjecture that there is a suitable positive constant c satisfying $p_K(n) \sim c \cdot n^q (n \to \infty)$, where $q = (\log 3)/\log(3/2)$.

Recall from [18] that a C^∞-word w is *left doubly extendable* (LDE) if both $1w$ and $2w$ are C^∞, and a C^∞-word w is *fully extendable* (FE) if $1w1, 1w2, 2w1$, and $2w2$ all are C^∞-words. For each nonnegative integer k, let $A(k)$ be the minimum length and $B(k)$ the maximum length of an FE word of height k.

Weakley in [18] proved that there are positive constants c_1 and c_2 such that for each n satisfying $B(k-1)+1 \leq n \leq A(k)+1$ for some k, $c_1 \cdot n^q \leq \gamma(n) \leq c_2 \cdot n^q$.

It is a pity that we don't know how many positive integers n fulfil the conditions required. Set $\gamma'(n) = \gamma(n+1) - \gamma(n)$, Weakley in [18] gave

$$\gamma(n) = \gamma(0) + \sum_{i=0}^{n-1} \gamma'(i) \ \textit{for}\ n \geq 2. \tag{14}$$

Let $F(n)$ denote the number of LDE-words of height n, Shen and Huang in [14, Proposition 3.2] established

$$F(n) = 4 \cdot 3^{n-1} \ \textit{for each positive integer}\ n. \tag{15}$$

Huang and Weakley in [9] combined (14) with (15) to show that

Theorem 5 ([9] Theorem 4)**.** *Let ξ be a positive real number and N a positive integer such that for all LDE words u with $|u| > N$ we have $|u|_2/|u| > (1/2) - \xi$. Then there are positive constants c_1, c_2 such that for all positive integers n,*

$$c_1 \cdot n^{\frac{\log 3}{\log((3/2)+\xi+(2/N))}} < \gamma(n) < c_2 \cdot n^{\frac{\log 3}{\log((3/2)-\xi)}}.$$

Let $\gamma_{a,b}(n)$ denote the number of smooth words of length n over 2-letter alphabet $\{a, b\}$ for positive integers $a < b$, Huang in [10] obtained

Theorem 6. *For any positive real number ξ and positive integer n_0 satisfying $|u|_b/|u| > \xi$ for every LFE word u with $|u| > n_0$, there exist two suitable constants c_1 and c_2 such that*

$$c_1 \cdot n^{\frac{\log(2b-1)}{\log(1+(a+b-2)(1-\xi))}} \leq \gamma_{a,b}(n) \leq c_2 \cdot n^{\frac{\log(2b-1)}{\log(1+(a+b-2)\xi)}}$$

for every positive integer n.

Now we are in a position to use the number of *MRSE* chains of the suitable height k to bound the number of smooth words of length n, which is simpler than the ones used in [9, 10].

Theorem 7. *For any positive number θ and n_0 satisfying $|u|_2/|u| > \theta$ for $|u| > n_0$, there are suitable positive constant c_1, c_2 such that*

$$c_1 \cdot n^{\frac{\log 3}{\log(2-\theta)}} \leq \gamma(n) \leq c_2 \cdot n^{\frac{\log 3}{\log(1+\theta)}} \text{ for any positive integer } n.$$

Proof. It is obvious that

$$|w| = |D(w)| + |D(w)|_2 + c \text{ for each smooth word } w, \text{ where } c = 0, 1, 2. \tag{16}$$

First, since $|u|_2/|u| > \theta$ for $|u| \geq n_0$, from (16) one has

$$|w| \geq (1+\theta)|D(w)| \text{ for } |D(w)|_2/|D(w)| > \theta,$$

which implies

$$|D(w)| < \alpha|w| \text{ for } |w| \geq N_0, \tag{17}$$

where N_0 is a suitable fixed positive integer such that $|D(w)| \geq n_0$ as soon as $|w| \geq N_0$, $\alpha = 1/(1+\theta)$.

Let s_0 be the greatest height of all smooth words with length$< N_0$ and n_1 is the least positive integer such that if $|w| = n_1$, then the height of the smooth word w is larger than s_0. Thus for any smooth word w, if $|w| \geq n_1$, then from (17) one can get

$$|D^s(w)| < \alpha^{s-s_0}|w|.$$

Hence

$$\begin{aligned} |D^s(w)| < 1 \quad &\text{only if} \quad \alpha^{s-s_0}|w| \leq 1 \\ &\iff \quad s \geq \log(|w|)/\log(1+\theta) + s_0, \end{aligned}$$

which means that the height s of w is smaller than $\log(|w|)/\log(1+\theta) + s_0 + 1$. Therefore, the maximal height $h_2(n)$ of all smooth words of length n satisfies

$$h_2(n) \quad \leq \quad \frac{\log n}{\log(1+\theta)} + t_2, \text{ where } t_2 = s_0 + 1. \tag{18}$$

Put $k = h_2(n) + 1$, then the length of every smooth word of height k is greater than n, so each smooth word of length n can be right extended to get a *MRSE* chain of height k, which suggests $\gamma(n) \leq |H^k|$. Consequently, from (13) and (18) it follows the desired upper bound of $\gamma(n)$.

Second, since the complement of any smooth word is a smooth word of the same length, the theorem's hypothesis implies that $|D(w)|_1/|D(w)| \geq \theta$, so $|D(w)|_2/|D(w)| \leq 1-\theta$. From (16) it follows that

$$|w| \leq \beta|D(w)| + q \text{ for each } C^\infty\text{-word } w, \tag{19}$$

where $\beta = 2 - \theta$, q is a suitable positive constant. Thus

$$|w| \leq \beta^{k-1}|D^{k-1}(w)| + q\frac{\beta^{k-1}-1}{\beta-1} < 2\beta^{k-1} + \frac{q\beta^{k-1}}{\beta-1} = (2 + \frac{q}{\beta-1})\beta^{k-1} = t\beta^{k-1},$$

where $t = 2 + q/(\beta - 1)$, k is the height of $|w|$. Wherefore, the length $|w|$ of a smooth word w with height k is less than $t\beta^{k-1}$ and $k - 1 > (\log|w| - \log t)/\log\beta$. Hence, the smallest height $h_1(n)$ of smooth words of length n meets

$$h_1(n) > \frac{\log n}{\log(2-\theta)} + t_1, \ \textit{where}\ t_1 = 1 - \frac{\log t}{\log\beta}. \tag{20}$$

Then the length of all smooth words with height $m = h_1(n) - 1$ is less than n, which means that the length of the last member *last* (ξ) is less than n for each $\xi \in H^m$. Since each smooth words of length no more than $n - 1$ can be extended right to a smooth word of length n, we see $\gamma(n) \geq |H^m|$. Herewith, from (13) and (20) we get the desired lower bound of $\gamma(n)$. □

4. Concluding remarks

Let a and b be positive integers of different parities and $a < b$. Lately, Sing in [15] conjectured:

There are positive constants c_1, c_2 such that

$$c_1 \cdot n^{\delta} \leq \gamma_{a,b}(n) \leq c_2 \cdot n^{\delta}, \ \text{where}\ \delta = \frac{\log(a+b)}{\log((a+b)/2)}.$$

Theorem 6 means Sing's conjecture should be revised to be of the following form

$$c_1 \cdot n^{\theta} \leq \gamma_{a,b}(n) \leq c_2 \cdot n^{\theta}, \ \text{where}\ \theta = \frac{\log(2b-1)}{\log((a+b)/2)}.$$

For 2-letter alphabet $\Sigma = \{a, b\}$ with $a < b$, let $P^j(\varepsilon)$ denote the set of smooth words of height k for $j \in \mathcal{N}$. For $\alpha \in \Sigma$, set

$$\xi_i = \alpha^i \prec \alpha^i\bar{\alpha} \prec \alpha^i\bar{\alpha}^2 \prec \cdots \prec \alpha^i\bar{\alpha}^{b-1} \text{ for } 1 \leq i \leq b-1.$$

and

$$H^1 = \{\eta | \eta = \xi_i \text{ or } \bar{\xi}_i,\ i = 1, 2, \cdots, b-1\}.$$

Let H^2 denote the set of primitives of the members in H^1, then it is easy to see H^2 constitutes a partition of $P^2(\varepsilon)$. So continue, we can define the set H^k for each $k \in \mathcal{N}$ and H^k constitutes a partition of $P^k(\varepsilon)$. Using the method similar to Theorem 7, we could establish the corresponding result to Theorem 6.

References

[1] **S. Brlek, A. Ladouceur,** *A note on differentiable Palindromes,* Theoret. Comput. Sci., 302 (2003), 167-178.

[2] **A. Carpi,** *Repetitions in the Kolakoski sequence,* Bull. of the EATCS, **50** (1993), 194-196.

[3] **A. Carpi,** *On repeated factors in C^{∞}-words,* Inform. Process. Lett., **52** (6) (1994), 289-294.

[4] **V. Chvátal,** *Notes on the Kolakoski sequence,* DIMACS Tech. Rep., 93-84 (1993).

[5] **F. M. Dekking,** *Regularity and irregularity of sequences generated by automata,* Séminaire de Théorie des Nombres de Bordeaux, 1979-80, exposé n° 9, 901-910.

[6] **F. M. Dekking,** *On the structure of selfgenerating sequences,* Séminaire de Théorie des Nombres de Bordeaux, 1980-81, exposé n° 31, 3101-3106.

[7] **F. M. Dekking,** *What is the long range order in the Kolakoski sequence?,* in R. V. Moody (ed.), The mathematics of long-range aperiodic order, Springer, New York, 1997, pp. 115-125.

[8] **J.-M. Fédou and G. Fici,** *Some remarks on differentiable sequences and recursivity,* J. Integer Sequences 10(3)(2010): 10.3.2.

[9] **Y. B. Huang, W. D. Weakley,** *A note on the complexity of C^{∞}-words,* Theoret. Comput. Sci., **411** (2010), 3731-3735.

[10] **Y. B. Huang,** *The subword complexity of smooth words on 2-letter alphabets,* submitted (2009.09), arXiv:1012.3229v1.

[11] **M. S. Keane,** *Ergodic theory and subshifts of finite type,* in: Ergodic Theory, Symbolic Dynamics and Hyperbolic Spaces, T. Bedford, M. Keane, C. Series (Eds.), Oxford University Press, Oxford, 1991, pp. 35-70.

[12] **C. Kimberling,** *Problem 6281,* Amer. Math. Monthly, **86** (1979), 793.

[13] **W. Kolakoski,** *Self Generating Runs, Problem 5304,* Amer. Math. Monthly, **72** (1965), 674. Solution: Amer. Math. Monthly, **73** (1966), 681-682.

[14] **C. L. Shen, Y. B. Huang,** *Some properties of C^∞-words with applications,* SEA Bull. Math., **20** (4) (1996), 19-30.

[15] **B. Sing,** *More Kolakoski Sequences,* arXiv:1009.4061v1 (2010).

[16] **R. Steacy,** *Structure in the Kolakoski sequence,* Bull. of the EATCS, 59 (1996), 173-182.

[17] **B. Steinsky,** *A Recursive Formula for the Kolakoski Sequence,* Journal of Integer Sequences, 9(3) (2006), 06.3.7.

[18] **W. D. Weakley,** *On the Number of C^∞-words of Each Length,* J. Comb. Theory, Ser. A, **51** (1989), 55-62.