MORE KOLAKOSKI SEQUENCES

Bernd Sing
Department of Computer Science, Mathematics & Physics,
University of the West Indies, Bridgetown, Barbados, West Indies
bernd.sing@cavehill.uwi.edu

Received: 9/16/10, Revised: 2/7/11, Accepted: 7/15/11, Published: 12/2/11

Abstract

Our goal in this article is to review the known properties of the mysterious Kolakoski sequence and at the same time look at generalizations of it over arbitrary two letter alphabets. Our primary focus here will be the case where one of the letters is odd while the other is even, since in the other cases the sequences in question can be rewritten as (well-known) primitive substitution sequences. We will look at word and letter frequencies, squares, palindromes and complexity.

1. Introduction

A one-sided infinite sequence z over the alphabet $\mathcal{A} = \{1, 2\}$ is called a (classical) *Kolakoski sequence*, if it equals the sequence defined by its run-lengths, i.e.:

$$z = \underbrace{22}_{2}\ \underbrace{11}_{2}\ \underbrace{2}_{1}\ \underbrace{1}_{1}\ \underbrace{22}_{2}\ \underbrace{1}_{1}\ \underbrace{22}_{2}\ \underbrace{11}_{2}\ \underbrace{2}_{1}\ \underbrace{11}_{2}\ \ldots \\ \ldots = z.$$

Here, a *run* is a maximal subword consisting of identical letters. The sequence $z' = 1z$ is the only other sequence which has this property.

This sequence was introduced by Kolakoski in [22] who asked "What is the nth term? Is the sequence periodic?"[1] This sequence has attracted attention over the years since, although it is easy to define, it resists any attempt to reveal even some of its most basic properties like recurrence or the frequency of its letters. There is even some prize money offered for answering some of these question about its properties, see [20, 21]. The maybe most basic question is known as *Keane's question* [19]:

> Does the frequency of the symbol 1 in $z = 221121\ldots$ exist, and is it equal to $\frac{1}{2}$?

[1] The first question is still studied today, see [32] and [15]. In these articles, recursive formulae for the nth term are derived thus answering the first question.

The line of attack in trying to answer this question has often been to detect some structure by rewriting the generation rule of the Kolakoski sequence in some sort of generalized substitution rule, see for example [31, 14], [28, Section 4.4] and references therein. However, these attempts have not been successful in answering Keane's question.

Our goal in this article is more humble – we want to give an overview of (the little) what is known about the Kolakoski sequence but at the same time look at generalizations to arbitrary two-letter alphabets $\mathcal{A} = \{r, s\}$ where r and s are natural numbers (with $r \neq s$). We note that "10" or "439" in this generalization is *one* letter not two or three, and we can (well, if we really want to) examine the Kolakoski sequence(s) over the alphabet $\mathcal{A} = \{10, 439\}$.

If we do this generalization, then we find that there is a drastically different behavior depending on whether $r + s$ is odd (i.e., one of the letters is odd while the other even) or $r + s$ is even (i.e., either both are even or both are odd). In particular, in the latter case we can answer Keane's question immediately. For this we use the observation made in [10]: One can obtain the Kolakoski sequence z above by starting with 2 as a seed and iterating the two substitutions

$$\sigma_0 : \begin{array}{ccc} 1 & \mapsto & 1 \\ 2 & \mapsto & 11, \end{array} \quad \text{and} \quad \sigma_1 : \begin{array}{ccc} 1 & \mapsto & 2 \\ 2 & \mapsto & 22 \end{array}$$

alternatingly, i.e., σ_0 substitutes letters on even positions and σ_1 letters on odd positions:

$$2 \mapsto 22 \mapsto 2211 \mapsto 221121 \mapsto 221121221 \mapsto \ldots$$

Clearly, the iterates converge to the Kolakoski sequence z (in the obvious product topology), and z is the unique (one-sided) fixed point of this iteration.

Similarly, a (generalized) Kolakoski sequence over an alphabet $\mathcal{A} = \{r, s\}$, which is again also equal to the sequence of its run-lengths, can be obtained by iterating the two substitutions

$$\sigma_0 : \begin{array}{ccc} r & \mapsto & r^r \\ s & \mapsto & r^s \end{array} \quad \text{and} \quad \sigma_1 : \begin{array}{ccc} r & \mapsto & s^r \\ s & \mapsto & s^s \end{array}$$

alternatingly. Here, a^b denotes a run of b a's, i.e., $a^b = a \ldots a$ (b times).

Let us now assume that both r and s are even numbers. Building blocks of two letters $A = rr$ and $B = ss$ and applying the alternating substitution rule to them, one actually obtains a usual substitution rule for A and B:

$$\sigma : \begin{array}{ccc} A & \mapsto & A^m B^m \\ B & \mapsto & A^n B^n \end{array}$$

where $m = \frac{r}{2}$ and $n = \frac{s}{2}$. In fact, from this (primitive) substitution rule it is easy to see that the frequency of the letters r and s in the original sequence must be equal, see [29, 30].

Let us now assume that both r and s are odd numbers. Again, building blocks of two letters helps, although we need three such blocks here: $A = rr$, $B = rs$ and $C = ss$. For these three letters one again obtains a usual (primitive) substitution rule:

$$\sigma : \begin{array}{ccc} A & \mapsto & A^m B C^m \\ B & \mapsto & A^m B C^n \\ C & \mapsto & A^n B C^n \end{array} \tag{1}$$

where $m = \frac{r-1}{2}$ and $n = \frac{s-1}{2}$. From this representation it is straightforward to calculate the letter frequencies in the corresponding Kolakoski sequence. However, here the frequencies of r and s are not equal[2], see [1, 29].

We will therefore look at the generalizations of the Kolakoski sequence in this article where one of the letters in the alphabet is odd while the other is even. We will not look at generalizations to three-letter alphabets (see for example [2]) since there the situation is in general[3] certainly worse than for two-letter (where we only alternate between two letters).

2. Derivatives and Primitives

Broadly speaking, there are (currently) two approaches to study a Kolakoski sequence: Either takes the "global perspective" and tries to examine the set of all infinite sequences over $\mathcal{A} = \{r, s\}$ with the property that their run-length sequence is also a sequence over the same alphabet $\mathcal{A} = \{r, s\}$ (and the run-length sequence of the run-length sequence – and so on – is also a sequence over $\mathcal{A} = \{r, s\}$). Or, one takes the "local perspective" and tries to study the set of all possible (finite) subwords (or factors) of the Kolakoski sequence. This leads to the study of so-called C^∞-words. In some sense, these two approaches are two sides of the same coin, nevertheless we will split our exposition into these two parts. We will introduce C^∞-words in this and the next section, and will show how the former approach via sequences is used in Section 4.

We start with some basic definitions. Let $\mathcal{A}$ be an alphabet, which throughout this article will always be a two-letter alphabet $\mathcal{A} = \{r, s\}$ where $r, s \in \mathbb{N}$. Then $z \in \mathcal{A}^{\mathbb{N}}$ is a (one-sided infinite) *sequence* of letters in $\mathcal{A}$. Any $w = w_1 w_2 \dots w_n \in \mathcal{A}^n$ where $n \in \mathbb{N}$ is a *word* of length n and we use the notation $|w| = n$ to denote the length of w. We denote the *empty word* by ε. Furthermore, we use the notation

[2] One can show that the substitution in (1) is a Pisot substitution with cubic Pisot-Vijaraghavan number if $2(r+s) \geq (r-s)^2$. It is a unimodular Pisot substitution if $r = s \pm 2$. In the case $2(r+s) < (r-s)^2$, all roots of the corresponding substitution matrix are greater than 1 in modulus (and cubic algebraic numbers). A formula for the letter frequencies in the case that one of the odd numbers is 1 can be found in [4].

[3] Of course, there are also simple cases where we can rewrite everything using one substitution rule: If the three letters are equal modulo 3, building blocks of three letters is the key. At least, if we alternate the three letters periodically in the original sequence.

$|w|_r$ and $|w|_s$ for the number of r's and s's in the word w, and, moreover, $|w|_v$ for the number of occurrences of the word v in the word w.

Since we are working in a two-letter alphabet, we can define the following two properties: Let $\tilde{\cdot}$ be the operation that exchanges letters, i.e., $\tilde{r} = s$ and $\tilde{s} = r$ extended to any word $w = w_1 w_2 \dots w_n$ by $\tilde{w} = \tilde{w}_1 \tilde{w}_2 \dots \tilde{w}_n$. Then a sequence z is called *mirror invariant* if

$$w \text{ occurs in } z \qquad \Longleftrightarrow \qquad \tilde{w} \text{ occurs in } z.$$

Similarly, the operation $\overleftarrow{\cdot}$ denotes the reversed word $\overleftarrow{w} = w_n w_{n-1} \dots w_2 w_1$ of a word $w = w_1 w_2 \dots w_{n-1} w_n$, and we say that a sequence z is *reversal invariant* if

$$w \text{ occurs in } z \qquad \Longleftrightarrow \qquad \overleftarrow{w} \text{ occurs in } z.$$

Our goal is now to see if we can say something about these properties in the case where one of the letters $\{r, s\}$ is odd while the other is even[4]. Here, we will closely follow [14, Section 3]. From now on, we will use the following convention:

$$r = \min\{r, s\} \qquad \text{and} \qquad s = \max\{r, s\}.$$

Let w be a word over $\mathcal{A} = \{r, s\}$. We define the following "*differentiation*" rule for w: The derivative $D(w)$ of w is, in principle, the run-length sequence of w except for (possibly) the first and last symbol. If w is a single run of length less than s, we set $D(w) = \varepsilon$. If w consists of more than one run and the first (last) run of w is of length less than or equal to r, we discard this run from it. If w consists of more than one run and the first (last) run of w has length between $r + 1$ and s, we extend it to a run of length s. The word $D(w)$ is now the run-length sequence of this altered word and might be the empty word ε (we use the convention $D(\varepsilon) = \varepsilon$). We say that w is *differentiable* if $D(w)$ is again a word over the same alphabet $\mathcal{A} = \{r, s\}$. Let us look at some examples using the alphabet $\{2, 5\}$:

$$\begin{array}{llll} D(255555222) = 55 & D(2555552) = 5 & D(2255) = \varepsilon & D(222555) = 55 \\ D(25555552) = 6 & D(25252) = 111 & D(222522) = 51 & D(2555222) = 35 \end{array}$$

Note that the words in the second line are not differentiable!

[4] Let us have a look back at those Kolakoski sequences where either r and s are both odd or both even:

- If r and s are both even, the corresponding Kolakoski sequence is not mirror invariant, e.g., the sequence $z = 2244222244442\dots$ has the subword 4444224 but not 2222442.
  This example also shows that such a Kolakoski sequence is not reversal invariant (e.g., 2442222 does not appear in the previous z).
- If r and s are both odd, the corresponding Kolakoski sequence is also not mirror invariant, e.g., 313331 appears in the Kolakoski sequence $z = 3331113331313331\dots$ while 131113 does not. However, in this case one has reversal invariance.

The definition of differentiable is chosen such that every subword of a Kolakoski sequence is differentiable. In fact, every subword of a Kolakoski sequence is *smooth* or a C^∞*-word* with respect to this differentiation rule over the respective alphabet, i.e., it is arbitrarily often differentiable.

We say that a word v is a *primitive* of a word w if $D(v) = w$. From our differential rule (discarding and/or extending the first and last run), one can conclude that each (nonempty) word has at least $2r^2$ and at most $2s^2$ primitives (the factor 2 appears since we have $D(v) = w = D(\tilde{v})$, i.e., a word and its mirrored word have the same derivative). E.g., over the alphabet $\{2, 3\}$ the primitives of 33 are: 222333, 3222333, 33222333, 2223332, 22233322, 32223332, 332223332, 322233322, 333222, 2333222, 22333222, 3332223, 33322233, 23332223, 223332223, 233322233, 3322233322, 2233322233.

One can now use the differentiation rule to prove the following statements:

Theorem 1 (i) *Kolakoski sequences are not eventually periodic (where a sequence z is called* eventually periodic *if there exist $m, q \in \mathbb{N}$ such that $z_{i+1} \dots z_{i+q} = z_{i+q+1} \dots z_{i+2\,q}$ for all $i \geq m$).*

(ii) *For a Kolakoski sequence, mirror invariance implies recurrence (where a sequence z is called recurrent if any word that occurs in z does so infinitely often).*

(iii) *For a Kolakoski sequence, mirror invariance holds iff each C^∞-word occurs in it.*

Proof.

(i) Compare [22] and [12, Example 4]. The reason is that a (minimal) period of length q in a sequence z yields a period of length $q' < q$ in its run-length sequence. Thus such a sequence z cannot be equal to its run-length sequence.

(ii) The proof of [14, Proposition 3.1] also applies here.

(iii) The proof of [13, Proposition 2] also applies here. □

For all Kolakoski sequence over one even and one odd symbol[5], nothing seems to be known beyond the above implications. We don't know whether or not all C^∞-words occur in such a Kolakoski sequence, or whether or not it is recurrent. In fact, it is even not known whether or not a Kolakoski sequence is *repetitive* (or *uniformly recurrent*), i.e., whether every word that occurs in the sequence does so with bounded gaps. The problem with the last property is, of course, that the

[5] For the case where the letters $\{r, s\}$ are both even or odd, see Footnote 4 on p. 4. Note that since they can be constructed using a primitive substitution rule, they are recurrent and even repetitive.

gap might be quite large, thus one has to be careful with claims based on numerical studies (as in [24, Section 4.1.4]). But one can use C^∞-words to answer the following question[6]: Given a word w, what is the maximal possible length of v such that wvw is a C^∞-word and w is not a subword of v? For the classical Kolakoski sequence over $\{1,2\}$ one obtains the following table:

$\lvert w\rvert$	1	2	3	4	5	6	7	8	9	10	11
maximal $\lvert v\rvert$	2	7	7	36	36	37	173	172	171	170	1230

So, at least all words of length less than 12 must occur with bounded gaps in the classical Kolakoski sequence supporting the conjecture that it is repetitive. Note that making this observation precise would prove that the Kolakoski sequence is repetitive, because this list tells us that there is no C^∞-word of length greater than $2\cdot 11+1230 = 1252$ such that its prefix of length (less than) 11 does not occur again within this word. The jumps in this list are closely related to the "degree" that we introduce in the next section.

3. C^∞-words and the "Kolakoski Measure"

We say that a C^∞-word has *degree* j if

$$D^j(w) \neq \varepsilon, \qquad D^{j+1}(w) = \varepsilon.$$

We call C^∞-words of degree 0, i.e., the primitives of the empty word ε, *fundamental words*. Note that a fundamental word has length less than $\max\{s, 2r+1\}$.

We now define a function μ on the *cylinder sets* $[w]$ of $\mathcal{A}^{\mathbb{N}}$, i.e., $[w] = [w_1 \dots w_n] = \{z \in \mathcal{A}^{\mathbb{N}} \mid z_1 = w_1, \dots, z_n = w_n\}$, by

$$\mu([w]) = \begin{cases} \mu\left(\left[D^j(w)\right]\right) \cdot \frac{1}{(r+s)^j} & \text{if } w \text{ is a } C^\infty\text{-word of degree } j, \\ 0 & \text{if } w \text{ is not a } C^\infty\text{-word.} \end{cases}$$

Here, we have to fix the function μ for all fundamental words, and we do so by requiring that $\mu([w]) = \mu([\tilde{w}])$, $\mu([w]) = \mu([\overleftarrow{w}])$ and $\mu([wr]) + \mu([ws]) = \mu([w])$ for all fundamental words w and that $\sum_{w\in\mathcal{A}^n} \mu([w]) = 1$ for $1 \leq n < \max\{s, 2r+1\}$. For example, one has for the fundamental words

using $\mathcal{A} = \{1,2\}$:	$\mu([1]) = \frac{1}{2}$	$\mu([2]) = \frac{1}{2}$	$\mu([12]) = \frac{1}{3}$	$\mu([21]) = \frac{1}{3}$
using $\mathcal{A} = \{2,3\}$:	$\mu([2]) = \frac{1}{2}$	$\mu([3]) = \frac{1}{2}$	$\mu([23]) = \frac{1}{5}$	$\mu([32]) = \frac{1}{5}$
	$\mu([22]) = \frac{3}{10}$	$\mu([33]) = \frac{3}{10}$	$\mu([223]) = \frac{1}{5}$	$\mu([332]) = \frac{1}{5}$
	$\mu([233]) = \frac{1}{5}$	$\mu([322]) = \frac{1}{5}$	$\mu([2233]) = \frac{1}{5}$	$\mu([3322]) = \frac{1}{5}$

[6] For the question "Given $|v| \leq n$, what is the maximal possible length of w such that wvw is a C^∞-word?" see [7, Proposition 7]: Based on the computations in [9], this length is bounded $O(n^{1.002})$, and it is conjectured to be $O(n)$. Also see [11, Section 6.3] and [8] on this question and its connection to Keane's question.

We note that there are $2 \cdot (r^2 + s - 1)$ fundamental words[7].

Clearly, one has the property $\mu([D(w)]) = (r+s) \cdot \mu([w])$ for any C^∞-word of length greater than or equal to $\max\{s, 2r+1\}$, and one can use this to show:

Theorem 2 *For any $\mathcal{A} = \{r, s\}$, the function μ extends to a Borel-measure (also denoted μ) on $\mathcal{A}^{\mathbb{N}}$. This measure is mirror invariant, reversal invariant and shift invariant.*

Proof. A careful case study as in [14, Theorem 5.1] works in the general case. □

The aim of introducing this measure is to connect it somehow to the frequencies of subwords w in a Kolakoski sequence. Indeed, one can show:

Theorem 3 *Suppose that z is a Kolakoski sequence over $\mathcal{A} = \{r, s\}$, where one of the numbers $r, s \in \mathbb{N}$ is odd and the other even, and that the frequencies $f_w = \lim\limits_{n\to\infty} \frac{|z_1 \dots z_n|_w}{n}$ exist and satisfy $f_w = f_{\tilde{w}}$ for all words occurring in z. Then for all words w we have $f_w = \mu([w])$.*

Proof. The proof of [14, Proposition 5.1] carries over to the general case, see [29, Proposition 2.5]. □

This is a nice result – if only we would know that the frequencies satisfy the required properties. In fact, one can state Keane's question for all Kolakoski sequences where one of the letters is odd and other one is even:

> Does the frequency of r exist in a Kolakoski sequence over $\{r, s\}$ (where one letter is odd and the other one is even)? And if so, does the letter frequency equal $\frac{1}{2}$?

Much computing time has been dedicated to find evidences for or against the conjecture that the letter frequency is $\frac{1}{2}$. The numerical evidences against it are usually dismissed by looking at larger and larger parts of the Kolakoski sequence, see [32].

Since already the existence of the letter frequency is in question, one can try to find bounds on $\limsup_{n\to\infty} |z_1 \dots z_n|_r/n$ and $\liminf_{n\to\infty} |z_1 \dots z_n|_r/n$ using the C^∞-words. A brute force approach is, of course, to generate all C^∞-words of a certain length, say n, and check for those with the least number[8] a of r's (since for any C^∞-word w, its mirrored version $\tilde{w}$ is also a C^∞ word, the maximal number of r's is $n - a$). One then has[9]

$$\frac{a}{n} \le \liminf_{n\to\infty} \frac{|z_1 \dots z_n|_r}{n} \le \frac{1}{2} \le \limsup_{n\to\infty} \frac{|z_1 \dots z_n|_r}{n} \le \frac{n-a}{n}.$$

[7] There are $2 \cdot (s-1)$ "single run" fundamental words (e.g., 2, 22, 3, 33) and $2 \cdot r^2$ "two runs" fundamental words (each of the two runs might be of length between 1 and r). The factor 2 appears because with each fundamental word also its mirrored word is a fundamental word.

[8] I.e., we have $a = \min\{|w|_r \mid |w| = n \text{ and } w \text{ is a } C^\infty\text{-word}\}$

[9] For a proof see [23, Section 3.2].

For example, one finds the following numbers for alphabets with $r+s \le 7$:

alphabet	$\{1,2\}$	$\{2,3\}$	$\{1,4\}$	$\{3,4\}$	$\{2,5\}$	$\{1,6\}$
length n	1355	8003	6003	1000	1000	1000
$a = \min_{\lvert w\rvert=n} \lvert w\rvert_r$	669	3989	2750	493	481	451
letter freq.	0.5±0.0063	0.5±0.0016	0.5±0.0419	0.5±0.007	0.5±0.019	0.5±0.049

Alternatively, one can use a generating function approach; see [23] (based on [26]) as follows. For each word w on the alphabet $\{r,s\}$, one defines the its *weight* as polynomial $x^{|w|_r}\, y^{|w|_s}\, t^{|w|}$. By summing these weights over all C^∞-words[10], a lower bound on the frequency is obtained by looking at the minimal degree of x for a given power t^n. The bound $\frac{1}{2} \pm \frac{17}{762} \approx 0.5 \pm 0.0223097$ was obtained using this method for the alphabet $\mathcal{A} = \{1,2\}$.

4. Chvatal's Bound on the Letter Frequency

Instead of considering C^∞-words, Chvatal [9] in his unpublished technical report looked at infinite words over $\{1,2\}$ with the property that their run-length sequence is also a sequence over the same alphabet. A sequence over $\{r,s\}$ is said to be 1-special. If only runs of length r and s occur in this sequence, we say that the sequence is 2-special. And if in this run-length sequence only runs of length r and s occur, we call the original sequence 3-special. We continue in this way and note that a Kolakoski sequence is d-special for all $d \in \mathbb{N}$.

Note that the (finite) subwords of a 2-special sequence are once differentiable, the subwords of a 3-special sequence are twice differentiable etc. So, one can think of $(d+1)$-speciality as extension of d times differentiability to infinite words. However, when considering an (iterated) run-length sequence, we do not discard the first letter/run of it here. E.g., while a Kolakoski sequence $2255222225\ldots$ over $\mathcal{A} = \{2,5\}$ is d-special for all d, the sequence $52255222225\ldots$ is only 1-special although all its subwords are C^∞-words as well.

We now write (a d-special) sequence and its d iterated run-length sequences in a special way in an array: The first row is the original sequence, the first row its run-length sequence and so on, but we align them appropriately in the columns. E.g., for the classical Kolakoski sequence (here we use the Kolakoski sequence starting

[10] In fact, it is computationally more feasible to sum over all words that just avoid to be C^∞-words, i.e., words on $\{r,s\}$ that are not C^∞-words but any of its (genuine) subwords is. This is the method used in [23, 26].

with 1), we write

1	2	2	1	1	2	1	2	2	1	2	2	1	1	2	1	1	2	2	1	...
1		2		2	1	1		2	1		2		2	1		2		2	1	...
1				2		2		1	1				2	1				2		...
1						2			2				1	1						...

We now call the ith element in a the original sequence *d-special* if the sequence itself is d-special and the ith column in this array has length at least d (there are no blanks in the first d lines of this column). We call this column (of length d) the *type* of the corresponding d-special element in the sequence. E.g., the third letter in the Kolakoski sequence above is 2-special of type 22, while the 7th letter is 4-special of type 1122 (it is also 2-special of type 11).

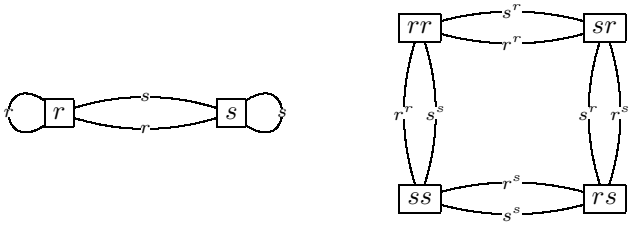
Now, the observation is that the type of a d-special element determines the first d terms of the type of the previous d-special element as well as all the letters between them, and the type of a d-special element and the last term of the type of the next d-special element determine the remaining terms of the type of this next d-special element. These properties can be used to iteratively build graphs G_d, $d \geq 1$. Since the same observations can be made about Kolakoski sequences on any alphabet $\{r, s\}$, we describe the more general case here.

The vertices of the graph G_d are the types of the d-special elements. Since all elements of $\mathcal{A}^d$ occur as types, the graph G_d has 2^d vertices. We connect two vertices u and v by a directed edge $u \xrightarrow{w} v$ labelled w, if v is the next d-special type after u in a d-special sequence z. If v is the ith element of such a d-special sequence and u the jth element, then the label w is the word $z_{i+1}z_{i+2}\dots z_{j-1}z_j$. So if we follow any (infinite) directed path in such a graph and read the edge-labels, we get a d-special sequence. Conversely, any d-special sequence arises as such an infinite path.

The trick is that one can build the graph G_{d+1} from the graph G_d.

- A path $Ar \xrightarrow{w_1} B_1 s \xrightarrow{w_2} B_2 s \xrightarrow{w_3} \dots \xrightarrow{w_r} B_r s$ in the graph G_d gives rise to the edges $Arr \xrightarrow{w_1 w_2 \dots w_r} B_r sr$ and $Ars \xrightarrow{w_1 w_2 \dots w_r} B_r sr$ in G_{d+1}.
- A path $As \xrightarrow{w_1} B_1 r \xrightarrow{w_2} B_2 r \xrightarrow{w_3} \dots \xrightarrow{w_r} B_r r$ in the graph G_d gives rise to the edges $Asr \xrightarrow{w_1 w_2 \dots w_r} B_r rr$ and $Ass \xrightarrow{w_1 w_2 \dots w_r} B_r rr$ in G_{d+1}.
- A path $Ar \xrightarrow{w_1} B_1 s \xrightarrow{w_2} B_2 s \xrightarrow{w_3} \dots \xrightarrow{w_s} B_s s$ in the graph G_d gives rise to the edges $Arr \xrightarrow{w_1 w_2 \dots w_r} B_s ss$ and $Ars \xrightarrow{w_1 w_2 \dots w_r} B_s ss$ in G_{d+1}.
- A path $As \xrightarrow{w_1} B_1 r \xrightarrow{w_2} B_2 r \xrightarrow{w_3} \dots \xrightarrow{w_s} B_s r$ in the graph G_d gives rise to the edges $Asr \xrightarrow{w_1 w_2 \dots w_s} B_s rs$ and $Ass \xrightarrow{w_1 w_2 \dots w_s} B_s rs$ in G_{d+1}.

The graphs G_1 and G_2 are :



To get bounds on the letter frequencies from a graph G_d, one associates to an edge with edge label w the cost $x \cdot |w| - |w|_r$. If one now uses for x a number $\frac{1}{2} \le x < 1$ that is smaller than the maximal possible letter frequency that can occur for a d-special sequence, then one finds a negative cycle in this graph, i.e., a cycle such that the sum of the costs over its edges is negative. Applying this method to G_6 for alphabets with $r + s \le 7$, one finds the following bounds:

alphabet	$\{1,2\}$	$\{2,3\}$	$\{1,4\}$	$\{3,4\}$	$\{2,5\}$	$\{1,6\}$
upp. bound	$^{12}/_{23}$	$^{53}/_{105}$	$^{592}/_{1085}$	$^{46}/_{91}$	$^{4834}/_{9527}$	$^{1478}/_{2821}$
letter freq.	0.5±0.0218	0.5±0.0048	0.5±0.0457	0.5±0.0055	0.5±0.0075	0.5±0.0240

By a clever use of the structure of the graphs G_d and efficient programming, Chvatal used G_{22} in [9] which yields the upper bound 616 904/1 231 743 for the classical Kolakoski sequence over $\mathcal{A} = \{1,2\}$, i.e., the letter frequencies are confined to 0.5 ± 0.000838.

5. Squares (and Cubes)

The question which (and how many) squares occur in the classical Kolakoski sequence was asked by [27]. Shortly thereafter, Carpi [6, 7] and Lepistö [25] answered the question by finding all squares that occur: in the classical Kolakoski sequence (i.e., using the alphabet $\mathcal{A} = \{1,2\}$) only squares of length 1, 2, 3, 9 and 27 ([25, Theorem 1], [6, Proposition 1], [7, Proposition 3]) occur; in particular, it is cube-free ([25, Corollary 1],[6, Proposition 2], [7, Proposition 4]). Here, a square w of length n is a C^∞-word with $|w| = n$ such that ww is appears in the respective Kolakoski sequence (in particular, ww is then also a C^∞-word).

The algorithm for finding squares is based on the following observations: If ww is a square, its derivative has the form $D(ww) = uvu$ where $|uv|$ has to be even (otherwise, not ww but $w\tilde{w}$ will be a primitive) and $|v| \le 1$ (we have $D(w) = u$ and v arises because of the rule on how to derive first and last runs). There is one speciality, though, if $r < \frac{s}{2}$: In these cases, v might also be a "negative" power r^{-1} or s^{-1} of length -1, meaning that in uv the v cancels the last symbol of u.

length n	number of squares	complexity $\gamma(n)$	max. MOOR
1	2	2	2
2	2	4	2
3	6	6	$2\,{}^{2}\!/_{3}$
9	12	42	$2\,{}^{1}\!/_{9}$
27	24	486	$2\,{}^{1}\!/_{27}$

Table 1: The classical Kolakoski case over $\mathcal{A} = \{1, 2\}$: There are 46 squares and no cubes.

length n	number of squares	thereof also cubes	complexity $\gamma(n)$	max. MOOR
1	2	2	2	3
4	4		8	$2\,{}^{1}\!/_{2}$
6	4		14	$2\,{}^{1}\!/_{6}$
10	20	2	30	3
15	30		58	$2\,{}^{7}\!/_{15}$
25	100	16	130	$3\,{}^{4}\!/_{25}$

Table 2: For $\mathcal{A} = \{2, 3\}$, there are 160 squares and 20 cubes among the C^∞-words.

If one continues differentiating, one gets a sequence of words $D(ww) = u_1v_1u_1$, $D^2(ww) = u_2v_2u_2$, $\ldots D^k(ww) = u_kv_ku_k$. But one can show that in this sequence the length of $|v_i|$ is bounded by $-1 \leq |v_k| \leq 2s + 1$ where the lower bound -1 only can appear if $r < \frac{s}{2}$, see [29, Lemma 4.4] (compare to [25, Lemma 1] for $\mathcal{A} = \{1, 2\}$). Furthermore, we must always have that $|u_iv_i|$ is even for all i. Thus, one now has an algorithm to find squares in Kolakoski sequences: We start with all C^∞-words of the form uvu where u is a fundamental word, $-1 \leq |v| \leq 2s+1$ and $|uv|$ is even (and/or $v = \varepsilon$ and thus we already have a square uu). Then construct all primitives which are again of the form $u'v'u'$ with either $|u'v'|$ even and $-1 \leq |v'| \leq 2s + 1$, and/or which happen to be a square $u'u'$. Continue in this way. If there are eventually no more words of this form left, the algorithm stops and one has calculated all squares among the C^∞-words. Lastly, one checks that each square indeed occurs in the respective Kolakoski sequence.

We note, however, that it is a priori not clear whether this algorithm will indeed stop, or if there are only finitely many squares in a Kolakoski sequence besides the classical one(s). However, we used this algorithm to check the C^∞-words over the alphabets $\{2, 3\}$ and $\{1, 4\}$: Both, similar to the classical Kolakoski sequence, have only finitely many squares – there are a total of 160 different squares of smooth words

over $\{2,3\}$ but $59,964$ squares of smooth words in $\{1,4\}$. We list the numbers in this case together with the classical Kolakoski sequence in Tables 1–3. We also listed the number of cubes and fourth powers in these cases together with the complexity at this word length. The *maximal order of repetition*, for short *MOOR*, or *repetition exponent* for a C^∞-word $w = uv$, is given by the maximum of $|ww\dots wu|/|w|$ such that $ww\dots wu$ is also a C^∞-word.

Since there are only finitely many squares, the corresponding Kolakoski sequences cannot be obtained by a (usual) substitution rule; see [25, Theorem 2] (if a substitution sequence has one square, one gets infinitely many using the substitution rule repeatedly). Also, the following conjecture was stated in [7]: For any repetition exponent $q > 1$, the length of C^∞-words having this exponent is bounded.

6. Palindromes

If a C^∞-word is a palindrome, i.e., if we have $w = \overleftarrow{w}$, then $D(w)$ is also a palindrome. Conversely, only palindromes of odd length have primitives that are also palindromes. We have a look at the following table:

palindrome	primitives
22	1122, 21122, 11221, 211221, 2211, 12211, 22112, 122112
212	**11211**, 211211, 112112, **2112112**, **22122**, 122122, 221221, **1221221**
121	**121121**, **212212**

Thus, together with a further observation, one has in fact an algorithm on how to construct palindromes as follows (compare to [24, Section 4.1.3]; see also [5, 3]): Start with all palindromic fundamental words. Palindromes of odd length where the letter in the middle is odd, will have palindromes of odd length among their primitives. Palindromes of odd length where the letter in the middle is even, will have palindromes of even length among their primitives. Palindromes of even length do not have palindromic primitives.

Since w is a palindrome if and only if $\tilde{w}$ is a palindrome, palindromes (in fact, palindromic fundamental words) of odd length where the letter in the middle is odd play a special role and can be used to construct all palindromes. For example, by repeatedly constructing primitives, one gets the following palindromic two-sided infinite sequence with the number 1 "in the middle" when starting with the fundamental word 1 over $\mathcal{A} = \{1,2\}$:

$$\dots 122121122 \,\big|\, 1 \,\big|\, 221121221 \dots$$

Applying the operation $\tilde{\cdot}$ to this word yields:

$$\ldots 211212211 \mid 2 \mid 112212112 \ldots$$

The primitives of this infinite sequence are:

$$\begin{array}{c|c|c} \ldots 12112212 & 11 & 21221121 \ldots \\ \ldots 21221121 & 22 & 12112212 \ldots \end{array}$$

Consequently, looking at the symmetric part of these sequences, one has two palindromes of each length (for details see the cited literature). In fact, one always has the single letters as fundamental words, so there are, for all alphabets, at least two palindromic C^∞-words for each length. For example, for $\mathcal{A} = \{2, 3\}$, the same construction as before works, where we now have

$$\ldots 2223322332223|3|3222332233222 \ldots$$

The situation gets a bit more complicated if there is more than one palindromic fundamental word of an odd length with an odd letter in the middle. For example, for the alphabet $\mathcal{A} = \{1, 4\}$ the two fundamental words with the stated property are 1 and[11] 111. Thus, additional palindromes appear (but for each length, one has at most two times the number of palindromic fundamental words of odd length with odd letter in its middle):

length	palindromes
1	1, 4
2	11, 44
3	111, 414, 444, 141
4	1111, 4444
5	44144, 14141, 11411, 41414
6	411114, 144441
7	4441444, 1114111
8	14111141, 44111144, 41444414, 11444411

Generalizations of palindromes, namely words of the form $\overleftarrow{w}vw$ ("palindromes with a gap in the middle"), have been studied in [16, 17].

7. Complexity

It is clear that the set of subwords of a Kolakoski sequence is a subset of the C^∞-words over the same alphabet. Since one, in fact, conjectures that the two sets are even identical, one tries to establish bounds on the number of C^∞-words for a given

[11] Note that 111 is a single run of length $3 < 4 = s$ and thus we have $D(111) = \varepsilon$.

length. We denote the complexity of C^∞-words, i.e., the number of C^∞-words of length n, by $\gamma(n)$.

Again, one can straightforwardly generalize results by Dekking.

Theorem 4 *Let $\gamma(n)$ be the number of C^∞-words of length n in the alphabet $\mathcal{A} = \{r, s\}$. Then*

(i) *there is an $N \in \mathbb{N}$ such that $\gamma(n) \le n^\alpha$ where $\alpha = \frac{\ln(2s^2)}{\ln(\frac{2rs}{r+s})}$ for all $n \ge N$.*

(ii) *there is an $N \in \mathbb{N}$ and a constant $C > 0$ such that $\gamma(n) \ge C \cdot n^\beta$ where $\beta = \frac{\ln(r+s)}{\ln(\frac{r^2+s^2}{r+s})}$ for all $n \ge N$.*

Proof. For a proof in the classical case $\mathcal{A} = \{1, 2\}$, see [13, Propositions 3 & 4]. For the generalizations, see [29, Propositions 4.1 & 4.3]. □

For the alphabet $\mathcal{A} = \{1, 2\}$ these bounds have recently been improved in [18] (based on previous work [33]). In this case, there are positive constants C_1, C_2 such that

$$C_1\, n^{2.7087} < \gamma(n) < C_2\, n^{2.7102}$$

for all $n \in \mathbb{N}$.

In fact, one can conjecture:

There are positive constants C_1, C_2 such that

$$C_1 \cdot n^\delta \le \gamma(n) \le C_2 \cdot n^\delta, \text{ where } \delta = \frac{\ln(r+s)}{\ln \frac{r+s}{2}}.$$

Noting that for $\mathcal{A} = \{1, 2\}$ we have $\delta = \ln 3 / \ln \frac{3}{2} \approx 2.7095$, we see that this conjecture is well supported by the above result. For $\mathcal{A} = \{2, 3\}$ and $\mathcal{A} = \{1, 4\}$, we refer to numerical results that we show in Fig. 1.

References

[1] M. Baake and B. Sing, Kolakoski(3, 1) is a (deformed) model set, *Canad. Math. Bull.* **47**(2):168–190 (2004).

[2] V. Berthé, S. Brlek and P. Choquette, Smooth words over arbitrary alphabets, *Theor. Comput. Sci.* **341**: 293–310 (2005).

[3] S. Brlek, S. Dulucq, A. Ladouceur and L. Vuillon, Combinatorial properties of smooth infinite words, *Theor. Comput. Sci.* **352**: 306–317 (2006).

[4] S. Brlek, D. Jamet and G. Paquin, Smooth words on 2-letter alphabets having same parity, *Theor. Comput. Sci.* **393**(1–3): 166–181 (2008).

[5] S. Brlek and A. Ladouceur, A note on differentiable palindromes, *Theor. Comput. Sci.* **302**: 167–178 (2003).

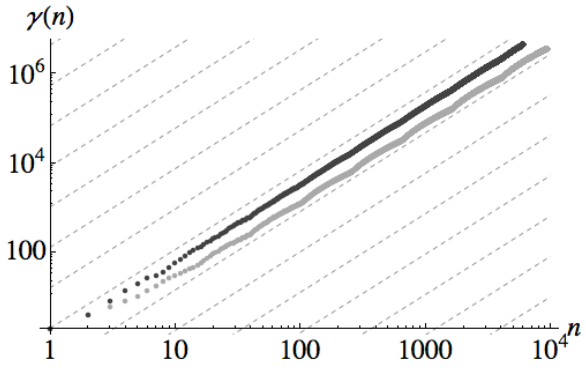


Figure 1: Complexity $\gamma(n)$ vs. length n for $\mathcal{A} = \{1, 4\}$ (dark gray) and $\mathcal{A} = \{2, 3\}$ (light gray). The dotted lines are the graphs of $f_k(n) = 2 \cdot 8^k \cdot n^{\ln 5/\ln \frac{5}{2}}$ (where $k \in \mathbb{Z}$) in this double-log plot.

[6] A. Carpi, Repetitions in the Kolakovski sequence, *Bull. EATCS* **50**: 194–196 (1993).

[7] A. Carpi, On repeated factors in C^∞-words, *Inf. Process. Lett.* **52**: 289–294 (1994).

[8] A. Carpi and V. D'Alonzo, On the repetitive index of infinite words, *Int. J. Alg. Comput.* **19**(2): 145–158 (2009)

[9] V. Chvátal, Notes on the Kolakoski sequence, *DIMACS Technical Report* 93-84 (1994).

[10] K. Culik II, J. Karhumäki and A. Lepistö, Alternating iteration of morphisms and the Kolakovski sequence, in: G. Rozenberg und A. Salomaa (eds.), Lindenmayer Systems, Springer, Berlin, 1992, pp. 93–106.

[11] V. D'Alonzo, On the repetitive index of infinite words, Dottorato thesis, Università degli Studi di Napoli "Federico II", 2009.

[12] F.M. Dekking, Regularity and irregularity of sequences generated by automata (exposé no. 9), *Sém. Th. Nombres Bordeaux* 1979–1980, 901–910.

[13] F.M. Dekking, On the structure of selfgenerating sequences (exposé no. 31), *Sém. Th. Nombres Bordeaux* 1980–1981, 3101–3106.

[14] 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, Kluwer, Dordrecht, pp. 115–125 (1997).

[15] J.-M. Fédou and G. Fici, Some remarks on differentiable sequences and recursivity, *J. Integer Sequences* **10**(3): 10.3.2 (2010), 7 pp.

[16] Y.B. Huang, About the number of C^∞-words of the form $\tilde{w}xw$, *Theor. Comput. Sci.* **393**(1–3): 280–286 (2008).

[17] Y.B. Huang, The complexity of C^∞-words of the form $\tilde{w}xw$, *Theor. Comput. Sci.* **410**(47–49): 4892–4904 (2009).

[18] Y.B. Huang and W.D. Weakley, A note on the complexity of C^∞-words, *Theor. Comput. Sci.* **411**(40–42): 3731–3735 (2010).

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

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

[21] C. Kimberling, http://faculty.evansville.edu/ck6/integer/index.html and http://faculty.evansville.edu/ck6/integer/unsolved.html

[22] W. Kolakoski, Self generating runs, Problem 5304, *Amer. Math. Monthly* **72**: 674 (1965). Solution by N. Üçoluk in: *Amer. Math. Monthly* **73**: 681–682 (1966).

[23] E.J. Kupin and E.S. Rowland, Bounds on the frequency of 1 in the Kolakoski word, *preprint*. http://arxiv.org/abs/0809.2776

[24] A. Ladouceur, Outil logiciel pour la combinatoire des mots, Mém. Maitrise Math., Université du Québec 'a Montréal, AC20U5511 M6258, 1999. http://www.mevis-research.de/ allouche/ladouceur.ps

[25] A. Lepistö, Repetitions in the Kolakoski sequence, in: G. Rozenberg and A. Salomaa (eds.), "Developements in Language Theory", World Scientific, Singapore, pp. 130–143 (1994).

[26] J. Noonan and D. Zeilberger, The Goulden-Jackson cluster method: extensions, applications, and implementations, *Journal of Difference Equations and Applications* **5**: 355–377 (1999). Algorithms available from: D. Zeilberger, DAVID_IAN, a Maple package, http://math.rutgers.edu/ zeilberg/gj.html.

[27] G. Păun, How much Thue is Kolakoski?, *Bull. EATCS* **49** (1993), 183–185.

[28] N. Pytheas Fogg, Substitutions in Dynamics, Arithmetics and Combinatorics, *Lecture Notes in Mathematics* **1784**, edited by V. Berthé, S. Ferenczi, C. Mauduit and A. Siegel, Springer, 2002.

[29] B. Sing, Spektrale Eigenschaften der Kolakoski-Sequenzen, Diploma thesis, Universität Tübingen, 2002 (available from the http://www.cavehill.uwi.edu/fpas/cmp/staff/bsing/pub.html#kolakoski).

[30] B. Sing, Kolakoski-$(2m, 2n)$ are limit-periodic model sets, *J. Math. Phys.* **44**(2):899–912 (2003).

[31] R. Steacey, Structure in the Kolakoski sequence, *Bull. EATCS* **59**: 173–182 (1996).

[32] B. Steinsky, A recursive formula for the Kolakoski sequence A000002, *J. Integer Sequences* **9**(3): 06.3.7 (2006), 5pp.
The corresponding numerical study of the letter frequencies is available at http://www.lirmm.fr/ monteil/blog/BruteForceKolakoski/

[33] W.D. Weakley, On the number of C^∞-words of each length, *J. Combin. Theory* **A51**: 55–62 (1989).

length n	number of squares	thereof also cubes	thereof also fourth powers	complexity $\gamma(n)$	max. MOOR
1	2	2	2	2	4
2	4	2		4	3
5	10	8		20	$3\,{}^{2}\!/_{5}$
8	6			36	$2\,{}^{1}\!/_{4}$
13	4			96	$2\,{}^{1}\!/_{13}$
20	30			198	$2\,{}^{7}\!/_{20}$
40	8			630	$2\,{}^{1}\!/_{40}$
50	152			964	$2\,{}^{12}\!/_{25}$
100	48			3124	$2\,{}^{3}\!/_{50}$
116	364			4160	$2\,{}^{59}\!/_{116}$
134	400			5438	$2\,{}^{65}\!/_{134}$
174	8			8658	$2\,{}^{1}\!/_{174}$
241	144			14694	$2\,{}^{20}\!/_{241}$
259	100			16588	$2\,{}^{2}\!/_{37}$
272	864			18358	$2\,{}^{145}\!/_{272}$
308	960			23288	$2\,{}^{1}\!/_{2}$
317	960			24554	$2\,{}^{160}\!/_{317}$
353	1044			29738	$2\,{}^{169}\!/_{353}$
408	4			38462	2
417	16			40046	$2\,{}^{1}\!/_{139}$
453	28			46474	$2\,{}^{2}\!/_{151}$
644	2072			82292	$2\,{}^{177}\!/_{322}$
716	2252			100990	$2\,{}^{375}\!/_{716}$
734	2312			106410	$2\,{}^{375}\!/_{734}$
806	2492			126570	$2\,{}^{399}\!/_{806}$
975	12			177330	$2\,{}^{2}\!/_{975}$
1065	12			208018	$2\,{}^{2}\!/_{1065}$
1529	4960			376874	$2\,{}^{845}\!/_{1529}$
1691	5404			451208	$2\,{}^{929}\!/_{1691}$
1709	5404			460688	$2\,{}^{896}\!/_{1709}$
1745	5500			480304	$2\,{}^{178}\!/_{349}$
1871	5860			550730	$2\,{}^{983}\!/_{1871}$
1925	12020			581470	$2\,{}^{989}\!/_{1925}$
2105	6508			684994	$2\,{}^{1049}\!/_{2105}$

Table 3: For $\mathcal{A} = \{1, 4\}$, there are 59 964 squares, 12 of which are also cubes and only 2 are also fourth powers among the C^{∞}-words.