Self Referencing Sequences

Perry Kleinhenz, Fermi Ma, and Erik Waingarten

{pkleinhe,fermima,eaw}@mit.edu

1 Introduction

Written by Fermi Ma, edited by Perry Kleinhenz and Erik Waingarten

We will study a sequences which are self-referencing. That is, each number in the sequence will correspond to a previous number in the same sequence. In particular, we will analyze self-referencing sequences with respect to block lengths. Consider the sequence

$$1, 2, 2, 1, 1, 2, 1, 2, 2, 1, 2, 2, 1, 1, 2, 1, 1, 2, 2, \ldots \tag{1}$$

We can break up the sequence into contiguous blocks, where each block is a stretch of repeated numbers:

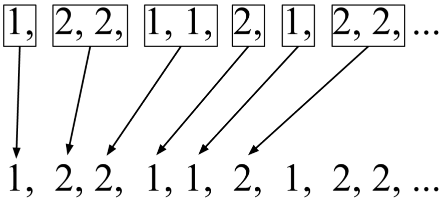

Fig. 1. Sequence broken up into contiguous blocks with lengths generating same sequence.

The block lengths, read from left to right, reproduce the original sequence (see Figure 1). We call such a sequence *self-referential*.

Above was an example where the blocks always have elements be either 1 or 2. Since we are self-referincing with respect to block lengths, the particular value that the block takes is not as important as the block lengths. Whenever the sequence can only takes two possible numbers, the values of the blocks is unambiguous; the block values simply alternate. As in the sequence above, after a block of 1's, the next block must be a block of 2's and vice versa. However if our sequence was composed of 1's, $2'$s and $3'$s there would be ambiguity over which number a new block should begin with.

For example if the first number is 1, then we could make the second number either a 2, or a 3. In particular, both:

$$1, 3, 3, 3, 2, 2, 2, \ldots \tag{2}$$

$$1, 2, 2, 1, 1, 3, 1, \ldots \tag{3}$$

are in some sense self-referencing sequences. Each time we need to specify a new block, we have a choice of two elements. We would like to avoid having a large number of choices in the generation of the sequence, and so we will use the convention that the order in which the elements are presented gives the order they are used in the sequence. For example, (123) will have a order $1 \to 2 \to 3 \to 1$.

With this rule, the self-referencing sequence generated by (123) that begins with a 1 is shown in Figure 1.

We call the set the sequence is generated over, along with its order the *generating cycle*.

Our main results are:

- The self-referencing sequence for a given generating cycle is uniquely determined by its first number.

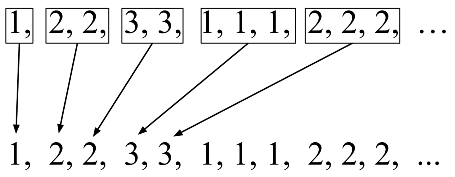

Fig. 2. Sequence generated by (123) broken up into contiguous blocks.

- For some generating cycles there exists an equivalent formulation of self-referencing sequences using substitution rules
- If a generating cycle has substitution rules then one can compute the limiting density for the numbers in the sequence if a limiting density exists.
- The limiting densities of the self-referencing sequences of (13) if such densities exists.

In Section 2 we show that if we are given a generating cycle and a starting number for the sequence, the sequence is uniquely determined. In Section 3, we consider the limiting behavior of such sequences, and we conjecture that the limiting behavior of a sequence is determined only by the generating cycle, and not the starting number.

In Section 4, we show that there is an alternate way of generating them with an iterative process. In Section 5, we show how a large number of these sequences can actually be generated with a simple set of substitution rules. We introduce the density problem in Section **??**, and use ideas developed in Section 4 and Section 5 to answer the question for certain types of generating cycles. Unfortunately, we find that the open problem of determining the density of the sequences generated by (12) is hard to solve using our methods.

2 Determinism

Written by Fermi Ma, edited by Perry Kleinhenz and Erik Waingarten

We claim that a self-referencing sequence is uniquely determined by the generating cycle C and the starting number $c \in C$. We will prove this by showing that if we are given the first number of the sequence, there is only one way to extend the sequence. In order to do this, we formalize the idea of breaking up a sequence into blocks.

Definition 1. *A block B is any maximal length subsequence of repeated numbers in a sequence. We denote by B_i the ith block of the sequence.*

Here, *maximal length* is a local property that means that the block cannot be extended to include the numbers preceding or succeeding the block, as they will not be the repeated number. Thus, if a sequence $\{a_i\}$ is

$$1, 3, 3, 3, 1, 1, 1, 3 \dots$$

the first block is $B_1 = \{1\}$, the second is $B_2 = \{3, 3, 3\}$, the third is $B_3 = \{1, 1, 1\}$, and so on.

Clearly, a sequence can be specified by giving the numbers of the sequence individually, or by specifying all the blocks B_i. This is enough to make the following claim:

Proposition 1. *A self-referencing sequence $\{a_i\}$ is uniquely specified by the generating cycle C and the starting number $c \in C$. In other words, such an $\{a_i\}$ is guaranteed to exist and is unique.*

Proof. The proof is constructive. The self-referential property of the sequence implies that the contents of block B_i are entirely determined by the values of c, i, and a_i. The value of i determines which number is repeated throughout the block. For example, when $i = 1$, the number is the starting number c, and when $i = 2$, we use the number following the starting number in C, and so on for larger values of i. The value of a_i simply dictates the length of block B_i.

Thus, the sequence is uniquely given by first "writing down" B_1, followed by B_2, followed by B_3, etc. One detail is that it remains to check that a_i is known when B_i is written down.

However, writing down a block always involves writing down at least one element of the sequence, so the value of a_i is always known when B_i is written down. Note that there is an edge case where a_i is not yet written down if the starting number is 1 and $i = 2$. However, in this case, the value of a_i is still known (as it is given by the rule) and thus B_i can still be written down.

From this point on, we will denote by $s(C, a)$ the unique sequence with generating cycle C and with starting number $a \in C$.

3 Equivalence

Written by Fermi Ma and Perry Kleinhenz, edited by Erik Waingarten

In this section, we consider long term equivalence of sequences. We introduce a formal definition of this and establish results that allow us to determine when two sequences are equivalent.

We would first like to make a formal definition of equivalence.

Definition 2. *We say that two sequence $\{a_i\}$ and $\{b_i\}$ are equivalent if there exists some integers $n, k \geq 0$ such that*

$$a_{n+i} = b_{k+i}, \quad i = 1, 2, 3, \dots$$

Note that two sequences can only be equivalent if they are generated by the same cycle.

As a basic example of equivalence let us consider sequences generated by $C = (123)$. The sequence which starts with a 2 is

$$2, 2, 3, 3, 1, 1, 1, 2, 2, 2, 3, 1, 2, 3, 3, 1, 1, 2, 2, 3, 3, 3, \dots\ ,$$

and the sequence that starts with a 3 is

$$3, 3, 3, 1, 1, 1, 2, 2, 2, 3, 1, 2, 3, 3, 1, 1, 2, 2, 3, 3, 3, \dots\ .$$

If we delete the first two numbers of the sequence starting with 2, and the first number of the sequence starting with 3, we see

$$\begin{aligned} 2, 2, &| 3, 3, 1, 1, 1, 2, 2, 2, 3, 1, 2, 3, 3, 1, 1, 2, 2, 3, 3, 3, \dots \\ 3, &| 3, 3, 1, 1, 1, 2, 2, 2, 3, 1, 2, 3, 3, 1, 1, 2, 2, 3, 3, 3, \dots\ , \end{aligned}$$

that the resultant sequences appear to be identical and so the sequences should be equivalent. At an intuitive level it seems obvious that these two truncated sequences are identical as they begin with the same two numbers and the numbers which were removed do not refer to blocks to the right of these numbers, but we cannot actually show that these two sequences are equivalent using just our definition as we would need to check equality for an infinite number of terms.

In order to develop a method to address this problem we introduce the notion of a read distance.

Definition 3. *Consider a sequence $\{a_i\}$. We define the read distance for an index i as the number of terms in between it a_i and the beginning of the ith block of numbers.*

The read distance can be any nonnegative integer. For example if we again consider generated by $C = (123)$ that starts with a 2

$$2, 2, 3, 3, 1, 1, 1, 2, 2, 2, 3, 1, 2, 3, 3, 1, 1, 2, 2, 3, 3, 3, \dots\ ,$$

the read distance for the first number, 2, is 0 since in a self-referencing sequence, the first number must refer to the first block. The read distance for the second 2 is 1 as the second block of numbers starts at the third number. We can continue to calculate the read distances for this sequence and see that they form a monotonically increasing sequence $0, 1, 2, 4, 6, 6, 6, 6, 7, 8, 9, \dots$.

This monotonicity is true in general, and we can in fact make a stronger statement on the value of the read distance at a given index.

Lemma 1. *Let $\{a_i\}$ be any self referencing sequence generated by some cycle of positive integers. Then the read distance for the number at index $i > 1$, is*

$$\sum_{k=1}^{i-1}(a_k - 1).$$

The read distance for the number at index 1 is always 0.

Proof. We will prove this using induction. We know that for $i = 1$ the read distance is 0, as the first block always starts at the first number. When $i = 2$ the read distance is $a_1 - 1$. The first block will be of length a_1 since the sequence is self referencing, therefore the second block will begin at position a_1. The number of elements between position a_1 and 2 is exactly $a_1 - 1$.

Now if we assume that our result holds for $i = l$ we would like to show that it holds for $i = l + 1$. We know that the distance between the start of the lth block and position l is

$$\sum_{k=1}^{l-1}(a_k - 1).$$

This lth block has length a_l and so the $(l + 1)$th block will begin a_l positions to the right of the beginning of the lth block. Additionally a_{l+1} is one position to the right of a_l. Therefore the beginning of the $(l + 1)$th block is

$$(a_l - 1) + \sum_{k=1}^{l-1}(a_k - 1) = \sum_{k=1}^{l}(a_k - 1)$$

positions to the right of a_{l+1}. But this is exactly the desired read distance for a_{l+1} and so we have our result.

Note that because $\{a_i\}$ is generated by positive integers we always have $a_k - 1 \geq 0$ and so read distance is increases monotonically.

Now that we have a nice expression for the value of the read distance at a given index we can show our first major result. It will allow us to check only finitely many indices in order to establish that two sequences are equivalent, thus solving the problem we encountered in our example.

Theorem 1. *Let $\{a_i\}, \{b_i\}$ be two self-referencing sequences generated by the same cycle. If there are integers m, l such that the read distance for a_m and b_l is d and*

$$a_{m+k} = b_{l+k}, \quad 0 \leq k \leq d - 1$$

then $\{a_i\}$ and $\{b_i\}$ are equivalent.

Proof. We will proceed by contradiction so assume otherwise. That is there exists some $j \geq d$ such that

$$a_{m+j} \neq b_{l+j}$$

but

$$a_{m+k} = b_{l+k}$$

for all $0 \leq k < j$.

Since the two sequences are both generated by the same cycle this implies that either a_{m+j} or b_{l+j} is part of the same block as the element proceeding it, while the other is not. Without loss of generality we will assume that $a_{m+j} = a_{m+j-1}$ and $b_{l+j} \neq b_{l+j-1}$.

The length of the block that a_{m+j-1} is a part of is a_{m+n}, for some $0 \leq n < j - 1$. Because the two sequences are identical for $0 \leq k < j$ and by our assumption that the read distance is the same for a_m and b_l, the read distance for b_{l+n} equals the read distance for a_{m+n}. This combined with the fact that $a_{m+n} = b_{l+n}$ shows that the block including a_{m+j-1} must be of the same length as the block including b_{l+j-1}.

Since the two sequences are the same for $0 \leq k < j$ the number of elements in the block containing a_{m+j-1}, with $k < j$, and the number of elements in the block containing b_{l+j-1}, with $k < j$, must be the same. Because a_{m+j} is part of the same block as the element preceding it, this means that b_{l+j} must be a part of the same block as the element proceeding it as well. This contradicts the fact that exactly one of the two elements is not in the same block as the element preceding it and so the two sequences must be equivalent.

This result also allows us to make a general statement for equivalence of sequences that begin with a 1.

Corollary 1. *If $C = (1c_1c_2 \cdots c_n)$, then $s(C, 1)$ and $s(C, c_1)$ are equivalent.*

Proof. Let $\{a_i\} = s(C, 1)$ and $\{b_i\} = s(C, c_1)$. The read distance of sequence $\{a_i\}$ for $i = 2$ is 0, and the read distance of sequence $\{b_i\}$ for $i = 1$ is 0. In addition, $a_2 = b_1 = c_1$, so by Theorem 1 the two sequences are equivalent.

We can also demonstrate a necessary condition for equivalence.

Theorem 2. *If two sequences are equivalent then the first element of the shared sequence must have the same read distance in both sequences.*

Proof. Suppose the two sequences are $\{a_j\}$ and $\{b_j\}$ and we have

$$a_{m+i} = b_{l+i} = c_i \quad 1 \leq i$$

such that if m or l were made smaller this relation would not hold. We will proceed by contradiction so assume that the read distance for a_{m+1} is x and the read distance for b_{l+1} is y such that $x < y$.

Since the two read distances are different we know there are two separate blocks of length c_1 in the common sequence, one which corresponds to a_{m+1} and another which corresponds to b_{l+1}. Note that the block corresponding to b_{l+1} is to the right of the block corresponding to a_{m+1} because the read distance for b_{l+1} is larger. Now because $\{a_i\}$ is self referential there exists some index n such that a_{m+n+1} is the element of $\{a_i\}$ which corresponds to the second block of length c_1.

We claim that

$$c_i = b_{l-n+i} \quad 1 \leq i \leq n$$

That is the subsequence $c_1, c_2, \ldots, c_n$ appears in $\{b_i\}$ such that the last term is next to the beginning of the common sequence. This is true because the length of the blocks of $\{c_i\}$ which occur before the second block of length c_1 must appear as terms of $\{b_i\}$ and must end next to the number which refers to the second block of length c_1.

Thus our sequence $\{b_i\}$ is

$$\cdots c_1c_2 \cdots c_nc_1c_2 \cdots c_n \cdots$$

where the beginning of the common sequence is the second c_1. We now claim that the common sequence is just the subsequence

$$c_1c_2 \cdots c_n,$$

repeated an infinite number of times.

Since the read distance for b_{l+1} is greater than the read distance for a_{m+1}, and the sequences are identical for subsequent indices, by Lemma 1 we know that the read distance for b_{l+i} is larger than the read distance for a_{m+i} by exactly $y - x$.

Thus b_{l+i} and $a_{m+(y-x)+i}$ correspond to the same block. But $b_{l+i} = c_i$ and $a_{m+(x-y)+i} = c_{x-y+i}$. Therefore $c_i = c_{y-x+i}$ and so the common sequence repeats.

Because the common sequence repeats and $\{b_i\}$ has the same cycle, but one which begins before the common sequence, l could be made smaller and have the equivalence relation hold. This contradicts our statement that l is as small as possible. Therefore the first element of the shared sequence must have the same read distance in both sequences.

Corollary 2. *If two self-referencing sequences are generated over the same cycle and then the read distance of their shared subsequence is the same everywhere.*

Proof. This follows from Theorem 2 and Lemma 1.

It is unclear to us whether all sequences generated by the same cycle are equivalent. This is certainly the case for $C = (123)$ and any two element cycles but we have been unable to generalize this to other three element or larger cycles. A key ingredient in establishing a counterexample would be more stringent necessary conditions for equivalence but it is unclear what these conditions should be or how to prove they are necessary. In order to establish that all sequences generated by the same cycle are in fact equivalent we would like to establish weaker or more generalizable sufficient conditions for equivalence, but based on our proof of Theorem **??** it does not seem like weaker conditions will actually be sufficient.

4 An Iterative Process

Written by Fermi Ma, edited by Perry Kleinhenz and Erik Waingarten

We change focus in this section and consider how to generate self-referential sequences with an iterative process. We first look at the sequences generated by the cycle $C = (12)$.

One possible iterative process is as follows. Let the sequence s^0 be 2, and let the sequence $s^{(i)}$ for $i \geq 1$ be the unique sequence that starts with 2 and is such that its block lengths, read from left to right, reproduce the sequence $s^{(i-1)}$. This gives:

$$\begin{aligned} s^{(0)} &= 2 \\ s^{(1)} &= 22 \\ s^{(2)} &= 2211 \\ s^{(3)} &= 221121 \\ s^{(4)} &= 221121221 \\ &\vdots \end{aligned}$$

Let s^∞ be the limit of this process. By limit, we mean the limit with respect to each term seperately, so we can define

$$s_j^{(i)} \to s_j^\infty$$

as $i \to \infty$. We claim that s^∞ is the unique self-referential sequence $s(C, 2)$. In fact we can prove a more general result about this iterative process.

Theorem 3. *Let C be some generating cycle. Let $s^0 = a$ for some $a \in C$ where $a \neq 1$, and let $s^{(i)}$ be the sequences generated by the iterative process. Then the limit of this process s^∞ exists and is the sequence $s(C, a)$.*

Proof. We prove this by induction on i. We show that if $s^{(i)}$ is a prefix of $s(C, a)$, then $s^{(i+1)}$ is a longer prefix of $s(C, a)$. The base case of the induction is satisfied, as $s^{(0)}$ is a prefix of $s(C, a)$.

For the inductive step, we observe that a prefix of the sequence describes the block lengths of a *longer* prefix of the sequence. We know that the sequence it describes is longer because the sequence it describes has length equal to the sum of the original prefix. In other words, the length of $s^{(i+1)}$ equals the sum of of the numbers in $s^{(i)}$, which is strictly greater than the length of $s^{(i)}$ (since there are numbers that are greater than 1).

The resultant sequence is a prefix of the self referencing sequence because if we apply our process to any prefix of the self referencing sequence it will produce a prefix of the self referencing sequence.

Note that we must specify that $a \neq 1$, because that is the only case in which the length of $s^{(i+1)}$ is not strictly greater than the length of $s^{(i)}$.

It turns out that we can be less restrictive about what our starting sequence, $s^{(0)}$, is. Going back to the case of $C = (12)$, we can let $s^{(0)}$ be *any* sequence of 1's and 2's that starts with a 2. And still, the limiting sequence will be the self-referential sequence. For example, suppose $s^{(0)} = 2, 1, 1, 1, 2$. The rules for generating $s^{(i)}$ give:

$$\begin{aligned} s^{(0)} &= \mathbf{2}, 1, 1, 1, 2 \\ s^{(1)} &= \mathbf{2}, \mathbf{2}, \mathbf{1}, 2, 1, 2, 2 \\ s^{(2)} &= \mathbf{2}, \mathbf{2}, \mathbf{1}, \mathbf{1}, \mathbf{2}, \mathbf{1}, 1, 2, 1, 1, 2, 2 \\ s^{(3)} &= \mathbf{2}, \mathbf{2}, \mathbf{1}, \mathbf{1}, \mathbf{2}, \mathbf{1}, \mathbf{2}, \mathbf{2}, \mathbf{1}, \mathbf{2}, 1, 1, 2, 1, 2, 2, 1, 1 \\ &\vdots \end{aligned}$$

Here, we bold the numbers that form a prefix of $s(C, 2)$. We note that $s^{\infty} = s(C, a)$ because because $s^{(0)}$ contains a 2 at the beginning, which by the above theorem will produce $s(C, a)$ in the limit. We call the rest of each $s^{(i)}$ that is not a prefix of $s(C, 2)$ the *tail.*

We now consider two different forms of convergence. We know that $s^{(i)}$ in some sense approaches $s(C, a)$, as arbitrarily many elements at the beginning of the sequence will match up with the beginning of $s(C, a)$ as i increases. However, if the tail of the sequence never entirely disappears, then there does not exist an i where $s^{(i)}$ is exactly a prefix of $s(C, a)$. Thus, we formalize two types of convergence

Definition 4. *We say that a sequence s^i is weakly convergent to $s(C, a)$ if $s^{\infty} = s(C, a)$. We say that the sequence s^i is strongly convergent if $s^{(i)}$ is a prefix of $s(C, a)$ for all $i \geq k$ for some k.*

It is unclear what conditions must be placed on the starting sequence in order to ensure strong convergence, although based on numerical simulation we make the following conjecture.

Conjecture 4 *Let $C = (1, 2)$. A sequence $s^{(i)}$ is strongly convergent to $s(C, 2)$ if the starting sequence begins with a 2 and contains no more than 3 elements. A sequence $s^{(i)}$ is strongly convergent to $s(C, 1)$ if the starting sequence begins with a 1 and contains between 2 and 3 elements. A sequence $s^{(i)}$ is weakly convergent but not strongly convergent to $s(C, a)$ if the starting sequence begins with an $a \in \{1, 2\}$ and contains 4 or more elements.*

We make this conjecture on the basis on computer simulations which executes the iterative process to at least 20 iterations. More iterations were not completed because after this point the simulations began to take impractically long amounts of time to complete.

5 Substitution Rules

Written by Fermi Ma and Erik Waingarten, edited by Perry Kleinhenz

The iterative process given in Section 4 is one way to generate self-referencing sequences. However, for certain generating cycles C, there is a simpler ways to perform this iteration using a fixed set of substitution rules. Take $C = (13)$ as an example, and consider the sequence that starts with a 3:

$$333111333131333111333\ldots$$

We can use the iterative process from Section 4, and suppose that $s^{(0)} = 33$. We get:

$$\begin{aligned} \{s^{(0)}\} &= 3, 3 \\ \{s^{(1)}\} &= 3, 3, 3, 1, 1, 1 \\ \{s^{(2)}\} &= 3, 3, 3, 1, 1, 1, 3, 3, 3, 1, 3, 1 \\ \{s^{(3)}\} &= 3, 3, 3, 1, 1, 1, 3, 3, 3, 1, 3, 1, 3, 3, 3, 1, 1, 1, 3, 3, 3, 1, 3, 3, 3, 1 \\ &\ldots \end{aligned}$$

It turns out that we can model this evolution with the following substitution rules:

$$\begin{aligned}&\text{Rule } A : 3,3 \to 3,3,3,1,1,1\\&\text{Rule } B : 3,1 \to 3,3,3,1\\&\text{Rule } C : 1,1 \to 3,1\end{aligned}$$

This is to be interpreted as follows. Starting from 3,3, the only possible rule to apply is Rule A, which gives $3,3,3,1,1,1$ at the following iteration. We then apply Rule A to $3,3$, Rule B to $3,1$ and Rule C to $1,1$. Note that at every step, we break up the sequence into chunks of 2 and then simultaneously apply *all* the rules. This will always be possible, since the rules preserve the fact that these sequences have even length. So at the following step, we have 3,3,3,1,1,1,3,3,3,1,3,1, which is $s^{(2)}$. Notice that in general, after the ith application of the substitution rules, the sequence is equal to $s^{(i)}$.

Why does this work? Well, first of all, notice that the rules are the natural applications of the self-referential sequence. Also, note that we have decided to make the rules of length 2; if we had made the self-referential rules $1 \to 1$ and $3 \to 333$, this would not have worked. In addition, we know that each block length preserves parity.

We would like to know if we can do this for any generating cycle C and starting number a. In particular we would like to be able to simulate our iterative process with substitution rules.

When we apply our substitution rules we should not need to know what happened to blocks other than the one we are currently substituting for. Therefore, each rule must somehow encode the length of blocks, the number of elements already written in the current block, and what number the current block is composed of. The first condition is simple to implement, as the rules will specify the length of blocks. In order to keep track of the numbers already written, each rule should be self-contained. That is each rule should write complete blocks. In order to comply with the third condition it is enough for the size of the generating cycle to divide the length of the rules. This is so that at the start of each rule, we always know which numbers to write.

Therefore, we can formulate a set of sufficient conditions to make rules to generate the self-referential sequence with generating cycle C.

We want to have a set of rules R where each rule $r \in R$ has length $|r| = l$, such that all $|C|^l$ strings are in the rule set and for each $r \in R$

- each rule writes blocks in the same order of C and each rule has the same starting number
- $\sum_{i \in r} i$ is divisible by l
- l is divisible by $|C|$

Given these conditions on every rule, we can guarantee that the self-referential sequence is generated. Note that in this case, the above example does not satisfy our rules until we say that $1,3 \to 3,1,1,1$.

Another example, we can have $C = (24)$, and we can have that the rules be

$$\begin{aligned}&\text{Rule } A : 2,2 \to 2,2,4,4\\&\text{Rule } B : 2,4 \to 2,2,4,4,4,4\\&\text{Rule } C : 4,2 \to 2,2,2,2,4,4\\&\text{Rule } D : 4,4 \to 2,2,2,2,4,4,4,4\end{aligned}$$

One of the reasons why generating a rule set for (12) is complicated is that since $1, ..., 1$ must be in the rule set, the length of the rules must be even (since $|C| = 2$). On the other hand, $2, 1, ..., 1$ must also be in the set and so the rule length must be odd.

6 A Density Computation

Written by Erik Waingarten and Fermi Ma, edited by Perry Kleinhenz

An open question regarding the self-referential sequence of 1's and 2's, $s((12), 1)$, is what the limiting *density* of 1's and 2's is. In this section, we consider the related problem of finding the density for sequences with generating cycle $C = (13)$ (note that the sequence that starts with a 1 differs by only one number from the sequence that starts with a 3, so we do not need to consider them separately). While we cannot prove that a limiting density exists, we give a numerical calculation of what the density is *if it exists*. We use a general technique that takes advantage of the fact that substitution rules when $C = (13)$, as shown in Section 5.

In Section 5, we found that repeated applications of the rules

$$\begin{aligned}\text{Rule 1} &: 3,3 \to 3,3,3,1,1,1\\ \text{Rule 2} &: 3,1 \to 3,3,3,1\\ \text{Rule 3} &: 1,1 \to 3,1\end{aligned}$$

to the initial sequence $\{3,1\}$ gives a growing sequence that converges to the self-referential sequence $s((13),3)$.

If we let A denote the subsequence $\{3,3\}$, B denote $\{3,1\}$ and let C denote $\{1,1\}$, then the rules can be rewritten as

$$\begin{aligned}\text{Rule 1} &: A \to ABC\\ \text{Rule 2} &: B \to AB\\ \text{Rule 3} &: C \to B,\end{aligned}$$

where AB and ABC are to be interpreted as concatenations of the subsequences.

Our technique for computing densities will work as follows. Let $N_i(A), N_i(B)$, and $N_i(C)$ denote the number of A's, B's, and C's, respectively, in the sequence given by the ith application of the substitution rules. We keep track of these counts of A's, B's, and C's at each round of rule applications, and we use these to determine limiting proportions of A's, B's and C's. From these limiting proportions, we can deduce the densities of 1's and 3's.

To start, consider the substitution process from the very first step. We start applying substitution rules to the initial sequence $\{3,1\}$. Since $\{3,1\} = B$, the initial counts are:

$$\text{Sequence at step 1} : \{3,1\},\ \text{Counts} : N_1(A) = 0, N_1(B) = 1, N_1(C) = 0.$$

After one application of the substitution rules, we have

$$\text{Sequence at step 2} : \{3,3,3,1\},\ \text{Counts} : N_2(A) = 1, N_2(B) = 1, N_2(C) = 0.$$

After one more application, we have

$$\text{Sequence at step 3} : \{3,3,3,1,1,1,3,3,3,1\},\ \text{Counts} : N_3(A) = 2, N_3(B) = 2, N_3(C) = 1.$$

For the purposes of computing densities, it turns out that only the counts are relevant and we do not need to keep track of the sequence at each step. We can compute the counts N_i from the values of N_{i-1} as follows

$$\begin{aligned}N_i(A) &= N_{i-1}(A) + N_{i-1}(B)\\ N_i(B) &= N_{i-1}(A) + N_{i-1}(B) + N_{i-1}(C)\\ N_i(C) &= N_{i-1}(A).\end{aligned}$$

The above equations are arrived at by looking at the set of rules involving A, B, and C. For example, an A at the ith iteration is written down every time there is an A or B in the previous iteration, which leads to the first equation above.

For computing densities, however, the *proportions* of A's, B's, and C's are needed. Fortunately, we can derive equations for the proportions by simply normalizing the equations for the counts. Let $P_i(X)$ denote the proportion of the sequence after iteration i that is made up of the

X subsequence. Normalizing gives the following equations:

$$\begin{aligned}
P_i(A) &= \frac{P_{i-1}(A) + P_{i-1}(B)}{3P_{i-1}(A) + 2P_{i-1}(B) + P_{i-1}(C)} \\
P_i(B) &= \frac{P_{i-1}(A) + P_{i-1}(B) + P_{i-1}(C)}{3P_{i-1}(A) + 2P_{i-1}(B) + P_{i-1}(C)} \\
P_i(C) &= \frac{P_{i-1}(A)}{3P_{i-1}(A) + 2P_{i-1}(B) + P_{i-1}(C)}.
\end{aligned}$$

The only difference is that each equation is divided on the right-hand side by $3P_{i-1}(A) + 2P_{i-1}(B) + P_{i-1}(C)$ to ensure that the proportions of A's, B's, and C's at each step add up to 1. Now, assume that there exist limiting proportions (and hence, limiting densities) as $i \to \infty$. Then at this limit, $P_i(X) = P_{i-1}(X) = P_\infty(X)$. Making these substitutions and solving for $P_\infty(A), P_\infty(B)$, and $P_\infty(C)$ in `Mathematica` gives the following values:

$$\begin{aligned}
P_\infty(A) &\approx \texttt{0.3760858894420931} \\
P_\infty(B) &\approx \texttt{0.4533976515164037} \\
P_\infty(C) &\approx \texttt{0.1705164590415032}.
\end{aligned}$$

Since $A = \{3, 3\}$, $B = \{3, 1\}$ and $C = \{1, 1\}$, the fraction of 1's in the limiting sequence is $0.5P_\infty(B) + P_\infty(C)$, and the fraction of 3's in the limiting sequence is $P_\infty(A) + 0.5P_\infty(B)$. So the density values, if they exist, are

Fraction of 1's: `0.397215284799705`
Fraction of 3's: `0.602784715200295`.

We were able to support these calcuations with numerical simulations. We generated the $s((13), 3)$ sequence up to 100,000 terms, and we found that the fraction of 1's was approximately 0.398, and that the fraction of 3's was approximately 0.602.

It is difficult to fully generalize this procedure for all substitution rule sets, as we cannot prescribe a general form for all substitution rules. However, we believe that this technique should work whenever all the inputs to the substitution rules are the same length (in this case, the inputs were all length 2), and the outputs are all lengths that are multiples of the inputs. We do not know if the resulting equations will always give exactly 1 solution, as they did in this case.

7 Density Bounds for $C = (12)$

As discussed in the Section 5, it is unclear how to establish a set of rules for $C = (12)$; however, there is some structure that we can exploit. In particular, we know that the sequences $1, 1, 1$ and $2, 2, 2$ will never appear in the self-referential sequence. This is because such sequences would be part of a block of size at least 3.

What does this mean? Well, we know that if there is a density of 1's and 2's, then that density d_2 is bounded between

$$\frac{1}{3} \leq d_2 \leq \frac{2}{3}$$

In general, this gives a procedure for computing the limiting density. For example, now we know that there must be at least one 2 for every three numbers. So now we can look at how many 2's are in any given string of length 9. We know there must be at least three 2's. There are only three cases where there are three 2's in a segment of 9 numbers in the sequence, namely

$$2, 1, 1, 2, 1, 1, 2, 1, 1$$

$$1, 2, 1, 1, 2, 1, 1, 2, 1$$

$$1, 1, 2, 1, 1, 2, 1, 1, 2$$

In each of these cases, we can read the subsequence that must have generated them part of them. In order, they are

$$2, 1, 2, 1, 2$$

$$1, 2, 1, 2, 1$$

$$2, 1, 2, 1, 2$$

but each of these are generated by $1, 1, 1$. This means that none of the sequences of length 9 are possible, so in addition to having one 2 for every group of 3, there is another 2 for every group of 9. Therefore, we have shown that the density must be bounded by

$$\frac{1}{3} + \frac{1}{9} \leq d_2 \leq 1 - \frac{1}{3} - \frac{1}{9}$$

In general, we can continue this process, and we did, to show that

$$\frac{1}{3} + \frac{1}{9} + \frac{1}{27} \leq d_2 \leq 1 - \frac{1}{3} - \frac{1}{9} - \frac{1}{27}$$

Of course, the number of cases to check increases very quickly: there are 3 choices to put the 2 in the groups of 3, 3 choices to put the 3 groups of 2's and 2 positions within these groups for the groups of 9, and so on. Its not hard to see that this number grows exponentially fast, making it computationally difficult to analyze more cases.

We believe that if there is a way to show this, we will be able to show that the density is $\frac{1}{2}$, since the density will be bounded below by $\sum_{n=1}^{\infty} \frac{1}{3^n} = \frac{1}{2}$ and bounded above by the same number. However, generalizing the procedure to improve the bound seems difficult.

8 Unresolved Problems

Written by Fermi Ma, edited by Perry Kleinhenz and Erik Waingarten

Our work leaves a number of unresolved problems.

- In Section 3, we propose the idea that in some sense, the number that we start a sequence with does not matter. In other words, any sequence that uses a certain generating cycle is equivalent to any other sequence created by the same generating cycle.
- In the same section, we propose a condition for determining when two sequences are not equivalent. We conjecture that if two sequences have unequal numbers with the same read distances, then they are not equivalent. It seems from trying cases that this condition is sufficient, but we have been unable to prove it.
- In Section 4 we make a conjecture on when sequences are strongly and weakly convergent. We have not yet been able to investigate strong and weak convergence in great detail, as we have been unable to make precise mathematical statements about these types of convergence. Most of what we know about them comes from analyzing numerical data.
- In Section 5, we look at substitution rules for certain "nice" generating cycles of numbers, such as (13) and (24). These sets allow for substitution rules because they satisfy the trio of properties we outline in that section. However, we do not know that these are the only properties. We leave unresolved the question of determining precisely which sets allow for substitution rules and which ones ones do not. The major difficulty here seems to come in proving that a certain generating cycle does *not* allow for substitution rules.
- In Section **??** we were able to compute the density of any sequence with generating cycle (13), and we were able to give bounds on the density of any sequence with generating cycle (12). However, we leave unresolved the question of what the density for (12) actually is, or even if we can obtain tighter bounds.

9 Haikus

Written by Fermi Ma

One, Two, Two, One, One,
Two, One, Two, Two, One, Two, Two,
One, One, Two, One, One...

For the final draft,
we will improve the haikus
(and the whole paper).

9.1 Poem

A sequence refering to its past
warrants many questions to be asked.
These sequences are deterministic
and their read distances an imporant statistic.

We want to compute the limiting density
so we've defined a process.
But one would be a fool to not follow our rule.
And compute the density for 1, 3.

The big question remains unanswered.
The density of 1,2 which seems to be one half.
Our rules are for numbers of a certain class.
Unfortunetly, dealing with 1,2 is a pain in the a**.