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

1.1 Problem Setup

Consider the sequence

$$1, 2, 2, 1, 1, 2, 1, 2, 2, 1, 2, 2, 1, 1, 2, 1, 1, 2, 2, \ldots$$

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

$$[1], [2, 2], [1, 1], [2], [1], [2, 2], [1], [2, 2], [1, 1], [2], [1, 1], [2, 2], \ldots$$

The block lengths, read from left to right, reproduce the original sequence. Thus, this sequence is *self-referential*. For the remainder of this paper we study properties of such sequences.

We say that this sequence is *generated* by the set of numbers used in the sequence, so the above sequence is generated by $\{1, 2\}$. We will refer to this set as the *generating set*. Observe that it is unambiguous which number to use next. After a block of 1's, the next block must be a block of 2's. There does exist ambiguity in a sequence with generating set $\{1, 2, 3\}$, since after a block of 1's, ther can be a block of 2's or a block of 3's. We assign an order to the generating set and resolve the ambiguity. The order we will use will be the order of the elements presented, for example, $\{1, 2, 3\}$ will have a order $1 \to 2 \to 3 \to 1$. Note that because order matters, the *generating set* is technically not a set, but we do not call it a "sequence" to avoid possible confusion with its corresponding self-referencing sequence.

With this rule, a self-referencing sequence generated by $\{1, 2, 3\}$ has blocks of 1's followed by blocks of 2's followed by blocks of 3's, followed by blocks of 1's, and so on. Thus, $\{1, 2, 3\}$ can generate

$$1, 2, 2, 3, 3, 1, 1, 1, 2, 2, 2, 3, 2, 1, 2, 2, 3, 3, 1, 1, \ldots$$

but not

$$1, 3, 3, 3, 2, 2, 2, 1, 1, 1, 3, 3, 2, 2, 1, 1, 3, 2, 1, 3, \ldots$$

as the second sequence has the numbers in the wrong order.

1.2 Overview

In this paper, we address a variety of questions regarding these sequences. In Section 2 we show that if we are given a generating set and a starting number for the sequence, the sequence is uniquely determined. After that, we consider how the starting number for the sequence affects what the sequence can be. In particular, in Section 3, we show that there exist limits on how much two sequences with the same generating set can resemble each other if they start with different numbers. In the following Section 4, we consider the limiting behavior of such sequences, and we conjecture that the limiting behavior of a sequence is determined only by the generating set, and not the starting number.

In Section 5, we take a different approach to analyzing these sequences by showing that there is an alternate way of generating them with an iterative process. In Section 6, we show how a large number of these sequences can actually be generated with a simple set of substitution rules. We then introduce the density problem in Section 7, and use ideas developed in Section 5 and Section 6 to answer the question for certain types of generating sets. Unfortunately, we find that the open problem of determining the density for the sequence of 1's and 2's 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 set and the starting number. To see why this is true, we can simply give an algorithm to build the sequence. Assume the generating set is $\{a_1, a_2, \ldots, a_m\}$ and that we start with some a_i. If we ever write down some a_k where $k > m$, we set this equal to $a_{k(\mathrm{mod}\ m)}$.

Algorithm:
Suppose the first number of the sequence is a_i. Then the first block of the sequence must be a block of a_i instances of a_i, so we extend the sequence to this block. If a_i is not 1, then there is a second element of that sequence. We read that second element, and then we add a block to the end of that sequence with length dictated by the value of the second element. If a_i is equal to 1, we write down a_{i+1} in the second spot, and use the value of a_{i+1} to dictate the length of the second block.

This procedure can continue deterministically so long as there is always an element to read. However, every time we write down a block, we read 1 element and write down at least 1 element, so the end of the sequence will always be after the spot we are reading. The only exception is when $a_i = 1$, but then since $a_{i+1} \neq 1$, this issue will never occur again.

Thus, a self-referencing sequence is uniquely determined by the generating set and the starting number. From this point on, we will denote by $s(A, a)$ the unique sequence generated by set A with starting number $a \in A$.

3 Prefixes

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

Definition 1. *We say that two sequences $\{a_i\}$ and $\{b_i\}$ differ by a prefix of length n if*

$$a_{n+i} = b_i,$$

for i any positive integer. if no such n exists then we say that the two sequences are independent.

Note that a pair of sequences can have prefixes of different length. For example if we have

$$\{a_i\} = \{1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5, 6, \ldots\}$$
$$\{b_i\} = \{7, 8, 9, 1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5, 6, \ldots\},$$

then a_i and b_i differ by prefixes of length $3 + 6k$ for k a nonnegative integer. Because of this we make the following definition.

Definition 2. *We say that two sequences $\{a_i\}$ and $\{b_i\}$ differ by a minimal prefix of length n if they differ by a prefix of length n, but do not differ by a prefix of length m for all $0 < m < n$.*

We note that if a pair of sequences differs by a prefix they differ by a minimal prefix. In our above example a_i and b_i differ by a minimal prefix of length 3. The following lemma is a trivial consequence of our algorithm for generating self-referential sequences:

Lemma 1. *Let $C = \{1, c_1, c_2, \ldots, c_n\}$, $s(C, 1)$ and $s(C, c_1)$ differ by a minimal prefix of length at most 1.*

Proof. Let a_i refer to the ith term of $s(C, 1)$ and b_i refer to the ith term of $s(C, c_1)$. If we use our algorithm for generating the sequence, we note that $a_1 = 1$, and so $a_2 = c_1$, and that this is the beginning of a new block of length c_1. Furthermore, $b_1 = c_1$ and it indicated the beginning of a block of length c_1. Note that our algorithm is "memory-less" in the sense that it does not maintain any state about previous elements. Because the number and the begining of the block is the same. The algorithm will generate the same sequence, shifted by 1. Therefore, $s(C, 1)$ and $s(C, c_1)$ differ by a minimal prefix of length at most 1. □

Theorem 2. *If $1 < a < b$ then the self-referencing sequence beginning with a generated over $\{1, a, b\}$ and the self-referencing sequence beginning with b generated over $\{1, a, b\}$ are independent.*

Proof. We proceed by contradiction. If the two sequences are not independent then they must differ by some minimal prefix of length n. If we let a_i be the sequence beginning with a and b_i be the sequence beginning with b the

$$a_{n+k} = b_k.$$

We know that the first b terms of the sequence which begins with a b are b's. Because $b > a > 1$ no block can be longer then b. Thus the terms immediately preceding and following this block in the sequence a_i must not be b's. Because of this the first block of b_i is the $m+1$th block of a_i, where m is some positive integer. Since the value of the sequence at position k is also the length of the kth block this means that $a_{m+1} = b_1$. In fact the kth block of b_i is the $(m+k)$th block of a_i and so

$$a_{m+k} = b_k$$

We now claim that the prefix has length strictly larger than m. Assume otherwise so $n \leq m$. This means the entry in the a_i which gives the length of the $m+1$th block occurs at or after the beginning of the $m+1$th block. By our proof that self-referential sequences are well defined we know that the second case cannot occur. Therefore the only other option would be for the $m+1$th block to start at position $m+1$, but this would require every block to have length exactly 1, but this is a contradiction because $a_1 = a > 1$, that is the first block has length greater than one. Therefore the prefix must have length strictly larger than m.

We note that based on a previous step the two sequences differ by a prefix starting at m. Since $m < n$ we have produced a prefix which starts before n which is a contradiction. Therefore the two sequences are independent. □

4 Equivalence

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

In this section, we consider a slightly weaker definition of equivalence. Consider, for example, sequences generated by $\{1, 2, 3\}$. Theorem 2 shows that the sequence starting with 2

$$2, 2, 3, 3, 1, 1, 1, 2, 2, 2, 3, 1, 2, 3, 3, 1, 1, 2, 2, 3, 3, 3, \ldots$$

and the sequence starting with 3

$$3, 3, 3, 1, 1, 1, 2, 2, 2, 3, 1, 2, 3, 3, 1, 1, 2, 2, 3, 3, 3, \ldots$$

are independent in the sense that neither sequence is exactly contained in the other. However, if we delete the first two numbers of the sequence starting with 2, and the first number of the sequence starting with 3, we get

$$3, 3, 1, 1, 1, 2, 2, 2, 3, 1, 2, 3, 3, 1, 1, 2, 2, 3, 3, 3, \ldots$$

in both cases. (Need to further justify why these two are the same, probably with a reading distance argument)

Thus, since both sequences are essentially the same if certain prefixes are removed, we call these sequences *equivalent*. To formalize, this, we say that two sequences $\{a_i\}$ and $\{b_i\}$ are equivalent if there exists some integers $n, k \geq 0$ such that

$$a_{n+i} = b_{k+i}$$

We now give the following conjecture

Conjecture 1. Any sequence generated by a set S is equivalent to any other sequence generated by the same set S.

This conjecture is partly motivated by the fact that the property holds for $\{1, 2, 3\}$. For generating sets with larger numbers, we have been unable to verify or refute the conjecture. However, in Section **??**, we show that the starting number of a sequence does not affect the densities of each number, which provides some evidence for this conjecture.

We now introduce a notion which will help us categorize sequences as equivalent.

Definition 3. *Consider a sequence x. We say define the read distance for an element x_i as the number of terms in between it and the beginning of the ith block of numbers. We denote this as $Rd(x_i)$.*

The read distance can be any nonnegative integer. For example if we consider the sequence a_i generated by $\{1, 2, 3\}$ beginning with a 2

$$2, 2, 3, 3, 1, 1, 1, 2, 2, 2, 3, 1, 2, 3, 3, 1, 1, 2, 2, 3, 3, 3, \ldots$$

the read distance for the a_1 (the first two) is 0 as the first block of numbers starts at the first number, the read distance for a_2 (the second two) 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 $0, 1, 2, 4, 6, 6, 6, 6, 7, 8, 9, \ldots$. From this we might suspect that the read distance for any self referencing sequence will also be monotonically increasing. We can in fact show a much stronger result.

Lemma 3. *Let a_i be any self referencing sequence generated by some set of positive integers. Then the read distance for the element a_i for $i > 1$ is*

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

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. For the base case $i = 2$ we must show that the read distance is $a_1 - 1$. Well we know that 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 lth block and the number at 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. Also note that 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 reading 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 the read distance is monotonically increasing. Although this shows us that read distance never decreases it does not tell us how quickly it increases, our next result gives a bound on the number of consecutive times the read distance can stay constant.

Lemma 4. *Consider a self referencing sequence x generated by $\{a_1, a_2, \ldots, a_n\}$ with $a = \max a_1, a_2, \ldots, a_n$. Suppose the read distance for x_i is d, then the read distance for x_{i+a} is at least $d + 1$.*

Proof. Assume otherwise. By our Lemma 3 we know that the read distance for x_{i+a} is at least d. By our assumption it must be exactly d. Lemma 3 also tells us that $x_{i+k} = 1$ for $k = 0, 1, 2, \ldots, a_n$. But this means that we have a block of length $a+1$, which is a contradiction of the self referencing property of the sequence. Therefore the read distance for x_i is at least $d + 1$. □

Theorem 5. *If two sequences are equivalent then the first element of the shared sequence must have the same read distance in both sequences.*

Proof. Assume otherwise. So if the two sequences are a_j and b_j and we have

$$a_{i+l} = b_{i+k} = c_i$$

such that if m or l were made smaller this relation would not hold. Suppose the read distance for b_{1+k} is x and the read distance for a_{1+l} is y. Without loss of generality let us assume that $x > y$. Therefore we know there are two separate blocks of length c_1 in the sequence, one which corresponds to a_{i+l} and another which corresponds to b_{i+k}. Because the two sequences are identical there exists some index n such that $c_{n+1} = a_m$ is the element of sequence a which corresponds to the second block of length c_1. Because the sequences are self referential we know that $c_1, c_2, \ldots, c_n$ must appears in b prior to element $i+k$. In fact the self referential nature of the sequence guarantees that $b_{i+k-1} = c_n$, that is this subsequence must end just before the two sequences become the same. Thus our sequence b is

$$\cdots c_1 c_2 c_3 \cdots c_n c_1 c_2 \cdots c_n \cdots$$

where the prefix begins at the second c_1. We now claim that the common sequence cycles through

$$c_1 c_2 \cdots c_n.$$

Well by Lemma 3 since the read distance for b_{1+k} is greater than the read distance for a_{1+l}, and the sequences are identical for subsequent indices, we know that the read distance for b_{i+k} is longer than the read distance for a_{i+l} by exactly $x - y$. Thus b_{k+i} corresponds to the same block as $a_{l+(x-y)+i}$. But $b_{k+i} = c_i$ and $a_{l+(x-y)+i} = c_{x-y+i}$ and so $c_i = c_{x-y+i}$ and so the common sequence cycles. Because the common sequence cycles and b has the same cycle, but one which begins before the common sequence l could be made smaller and have the equivalence relation hold. This is a contradiction and so the first element of the shared sequence must have the same read distance in both sequences. □

We also have a sufficient condition for equivalence

Lemma 6. *If two sequences a_i, b_i are such that for some integers m, l the read distance for a_m and b_l are the same and*

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

where k ranges from 0 to the read distance for a_m then a_i and b_i are equivalent.

Proof. We will prove this using a simple induction. Suppose the read distance for a_m is d then know that $a_{m+d} = b_{l+d}$ because $a_{m+d-1} = b_{l+d-1}$ and a_{m+d} is the beginning of a new block. Furthermore we have that the read distance for a_{m+d} equals the read distance for b_{l+d}. Now suppose that we have for $k \leq j$ that $a_{m+k} = b_{l+k}$ and the read distance for a_{m+k} equals the read distance for b_{l+k} and we would like to show that this holds for $k = j + 1$. By Lemma 3 we actually already know that the read distance for the two is the same, as it is the same for the preceding terms, which are equal.

If a_{m+k} is the last term in a block then b_{m+k} must also be the last term in its block, because the length of blocks is determined by earlier terms which we know to be equal and to have the same read distance. Since $a_{m+k} = b_{l+k}$ in this case we have $a_{m+k+1} = b_{l+k+1}$.

Otherwise a_{m+k} and a_{m+k+1} are in the same block, then we only need to show that b_{l+k} and b_{l+k+1} are in the same block as well. But as we argued above the length of blocks is determined by terms earlier in the sequence which we know to be equal and have the same read distance. Therefore we have $a_{m+k+1} = b_{l+k+1}$ for this case as well. This completes our induction and so a_i and b_i are equivalent. □

These conditions for equivalence lead us to make the following conjecture on a sufficient condition for nonequivalence.

Conjecture 2. If two sequences a_i and b_i have elements $a_k \neq b_l$, for $k, l > 1$ with the same read distance then they are not equivalent.

5 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 $\{1, 2\}$.

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:

$$\begin{aligned}
\{s^{(0)}\} &= 2 \\
\{s^{(1)}\} &= 2, 2 \\
\{s^{(2)}\} &= 2, 2, 1, 1 \\
\{s^{(3)}\} &= 2, 2, 1, 1, 2, 1 \\
\{s^{(4)}\} &= 2, 2, 1, 1, 2, 1, 2, 2, 1 \\
&\cdots
\end{aligned}$$

Let s^∞ be the limit of this process. We claim that s^∞ is the unique self-referential sequence $s(\{1, 2\}, 2)$ generated by $\{1, 2\}$ with starting number 2.

Proof. We prove this with induction. We show that if $s^{(i)}$ is a prefix of $s(\{1,2\},2)$, then $s^{(i+1)}$ is a longer prefix of $s(\{1,2\},2)$. The base case of the induction is satisfied, as $s^{(0)}$ is a prefix of $s(\{1,2\},2)$.

For the inductive step, observe that a prefix of the sequence describes the block lengths of a *longer* prefix of the sequence. We know that it describes a longer prefix because the prefix it describes has length equal to the sum of the original prefix. In other words, the length of $s^{(i+1)}$ must be equal to 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). This holds true because $s^{(i)}$ merely contains a collection of lengths, which are summed up to give the next prefix. □

It turns out that this same argument holds in general.

Theorem 7. *Let $\{s^0\} = a$ for some $a \in A$ where $a \neq 1$, and let $\{s^{(i)}\}$ for $i \geq 1$ be the unique sequence that starts with a and is such that its block lengths, read from left to right, reproduce the sequence $\{s^{(i-1)}\}$. Then the limit of this process, s^∞, is the sequence $s(A,a)$.*

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 $s^{(i)}$. The proof of this theorem is omitted, as it is essentially the same proof as for the $A = \{1,2\}, a = 2$ case.

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

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

Here, we highlight in bold the numbers that form a prefix of $s(\{1,2\},2)$. We note that $s^\infty = s(A,a)$ simply because $s^{(0)}$ contains a 2 at the beginning, whch then causes a 2,2 to appear at the start of $s^{(1)}$, which causes $2,2,1,1$ to appear at the start of $s^{(2)}$, and so on. The rest of each $s^{(i)}$ that is not a prefix of $s(1,2,2)$ we call the *tail*.

We now consider two different forms of convergence. We know that $s^{(i)}$ in some sense approaches $s(A,a)$, as arbitrarily many elements at the beginning of the sequence will match up with the beginning of $s(A,a)$ 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 $s(A,a)$. Thus, we can formalize two types of convergence: *weak convergence* and *strong convergence*. Weak convergence is satisfied if $s^\infty = s(A,a)$, and strong convergence is satisfied if $s^{(i)}$ is a prefix of $s(A,a)$ for all $i \geq k$ for some k.

6 Substitution Rules

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

The iterative process given in Section 5 is one way to generate self-referencing sequences. However, for certain generating sets A, there is a simpler ways to perform this iteration using a fixed set of substitution rules. Take $A = \{1,3\}$ as an example, and consider the sequence that starts with a 3:

$$33311133131333111333\dots$$

We can use the iterative process from Section 5, and suppose that $s^{(0)} = 33$. We get:

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

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 $s^{(i)}$.

7 Density