SELF GENERATING SETS AND NUMERATION SYSTEMS

David Garth
Department of Mathematics and Computer Science, Truman State University,
Kirksville, MO 63501, USA
dgarth@truman.edu

Joseph Palmer
Department of Mathematics and Computer Science, Truman State University,
Kirksville, MO 63501, USA
jap875@truman.edu

Ha Ta
Department of Mathematics and Computer Science, Truman State University,
Kirksville, MO 63501, USA
tdh186@truman.edu

Received: 1/29/08, Revised: 9/10/08, Accepted: 11/26/08, Published: 10/13/09

Abstract
Kimberling has studied a variety of sets generated in the following way. Let \( F \) be a countable set of functions, and let \( S = S_F \) be the smallest set containing 0 that is closed under any function in \( F \). When \( F = \{2x, 4x + 1\} \), the resulting set \( S \) is precisely the set of nonnegative integers whose binary expansion does not contain the block 11. The set of all such binary expansions corresponds to the set of greedy representations of the natural numbers with respect to the Fibonacci sequence. Other examples of a similar nature can be found in the literature. In this paper we explore the following question; which self generating sets consist of integers whose digit expansions in base two correspond to the digit expansions of the natural numbers with respect to a linearly recurrent base sequence? We study this problem in the framework of abstract numeration systems. That is, we consider an abstract numeration system as an infinite language over a finite alphabet, ordered under a genealogical ordering. We then define a self generating numeration system as one that can be realized as the set of base two expansions of the integers in some self generating set. Our first result is to prove a necessary and sufficient condition for an abstract numeration system to have a base. This result is then used to prove that certain families of generating functions give rise to self generating numeration systems that have a base sequence. Finally, we prove that the base sequence in any based self generating numeration system satisfies a linear recurrence. Many of our results make use of a natural tree structure that can be put on an abstract numeration system.
1. Introduction

The purpose of this paper is to explore some of the connections between self-generating sets and abstract numeration systems. Kimberling ([9], [10], [11]) has studied a variety of sets generated in the following way. Let $F$ be a countable set of functions, and let $S = S_F$ be the smallest set containing 0 that is closed under all the functions in $F$. In other words, $0 \in S$, and if $x \in S$ and $f \in F$ then $f(x) \in S$. Moreover, no other elements are in $S$.

As an example, the set $S$ generated by $F = \{2x, 4x + 1\}$ has been considered in the literature ([1], [10], [8]). Since $2x$ simply adds a zero to the binary expansion of $x$, and since $4x + 1$ adds a 01, the ordered sequence

$$\{0, 1, 2, 4, 5, 8, 9, 10, 16, \ldots\}$$

of elements of $S$ consists precisely of those integers whose binary expansions have no adjacent ones. Allouche, Shallit, and Skordev [4] studied the self generating set $S$ arising from $F = \{2x + 1, 4x + 2\}$ and showed that in this case

$$S = \{0, 1, 2, 3, 5, 6, 7, 10, 11, 13, 14, 15, \ldots\},$$

the set of natural numbers whose binary expansions do not contain the block 00.

The connection with numeration systems becomes apparent when considering the so-called greedy representations of the natural numbers with respect to some sequence. More precisely, let $\{b_i\}_{i \geq 0}$ be a strictly increasing sequence of natural numbers, with $b_0 = 1$. The sequence $\{b_i\}$ serves as a base for a positional numeration system for the natural numbers as follows. For $n \geq 1$, let $k \geq 0$ be such that $b_k \leq n < b_{k+1}$. Using the division algorithm we write

$$n = q_kb_k + r_k, \text{ where } 0 \leq r_k < b_k.$$ 

Thus, $q_k = \lfloor \frac{n}{b_k} \rfloor$ where $\lfloor \cdot \rfloor$ is the greatest integer function. For $i = k - 1, \ldots, 0$ let $q_i = \lfloor \frac{r_i}{b_i} \rfloor$ and $r_i = r_{i+1} - b_i q_i$. It follows that $n = q_0 b_0 + \cdots + q_k b_k$, and we say that the string $q_k q_{k-1} \cdots q_1 q_0$ is the greedy representation of $n$ with respect to the sequence $\{b_i\}$. We define the greedy representation of 0 to be 0, although most authors define it to be the empty word.

Suppose $\{b_i\}$ is the Fibonacci sequence, indexed as $b_0 = 1$, $b_1 = 2$, and $b_i = b_{i-1} + b_{i-2}$ for $i \geq 2$. Using the above algorithm it is not hard to show that the set of greedy representations of the natural numbers in this case consists of 0 and all words over $\{0, 1\}$ that begin with 1 and do not contain the block 11. This is precisely the set of binary expansions of the elements of the set $S$ generated by $F = \{2x, 4x + 1\}$.

The lazy representation of a natural number $n$ with respect to the Fibonacci sequence is obtained by successively replacing all occurrences of the string 100 in the greedy Fibonacci representation of $n$ with the string 011 until the resulting
string contains no occurrences of the block 00. Any leading zeros that arise in this process are disregarded. Thus, for \( F = \{2x + 1, 4x + 2\} \), the set of base 2 expansions of elements of \( S \) is the set of lazy Fibonacci representations of the natural numbers.

These examples are easy to generalize, and some examples are considered in \([8]\). For example, if \( F = \{2x, 4x + 1, 8x + 3\} \), then

\[
S_F = \{0, 1, 2, 4, 8, 9, 10, 16, 18, 32, 33, 34, 36, \ldots\}
\]

consists of the natural numbers whose binary expansions do not contain the block 111. These binary expansions correspond to the greedy expansions of the natural numbers with respect to the base \( \{b_i\} \) of Tribonacci numbers, where \( b_0 = 1, b_1 = 2, b_2 = 4 \), and \( b_i = b_{i-1} + b_{i-2} + b_{i-3} \) for \( i \geq 3 \). If \( F = \{2x + 1, 4x + 2, 8x + 4\} \), then the elements of the resulting self generating set \( S \) are precisely those natural numbers whose binary expansions do not contain the block 000. The set of all such expansions corresponds to the set of lazy Tribonacci representations of \( \mathbb{N} \) that are defined in a manner analogous to the lazy Fibonacci representations of \( \mathbb{N} \).

In this paper we will place these results in a more general context. In Section 2 we define the notion of an abstract numeration system. Under this definition, the set of base two expansions of the elements of any self generating set can be considered as the set of representations of the natural numbers in such a system. We will then prove a result that will be useful for determining whether such an abstract numeration system has a base sequence. In Section 3 we apply this result to give conditions on the set of generating functions \( F \) which guarantee that the binary expansions of the elements of \( S_F \) correspond to the digit expansions of the natural numbers with respect to some base sequence \( \{b_i\} \). Of natural interest is the question of whether the base sequence satisfies a linear recurrence. In Section 4 we show that for any set of generating functions \( F \), if the binary expansions of the elements of \( S_F \) correspond to the expansions of the natural numbers with respect to some base sequence, then the base sequence must satisfy a linear recurrence relation.

2. Numeration Systems

Non-standard numeration systems and more generally the so-called abstract numeration systems have been considered in a variety of settings in the literature (\([2], [3], [5], [7], [12], [13], [14], [15]\)). Generally such a numeration system is regarded as a set \( \mathcal{S} \) of words over some alphabet along with a bijection between that set of words and the natural numbers. The alphabet is the set of allowable digits, the words in \( \mathcal{S} \) are the valid representations of the natural numbers, and the bijection gives a means for obtaining the numerical value of a given representation. In this paper we take our digit set to be \( \Sigma_2 = \{0, 1\} \). To define our bijection we recall a few definitions. Let \( \Sigma_2^* \) denote the set of all finite words over \( \Sigma_2 \), and let \( \Sigma_2^+ \) be the
set of nonempty words over $\Sigma_2$. For $w \in \Sigma_2^*$, let $|w|$ denote the length of $w$. The \emph{radix order} [13] on $\Sigma_2^*$ is the ordering where, for $w, v \in \Sigma_2^*$, $w < v$ if $|w| < |v|$ or if $|w| = |v|$ and $w = uaw'$ and $v = ubv'$ with $a, b \in \Sigma_2$ and $a < b$ in the natural order on $\Sigma_2$. Enumerating the elements of $S$ under the radix order induces a natural order preserving bijection $N: S \to \N$. (We will use the convention that $0 \in \N$). We are now ready for our definition of a numeration system.

\textbf{Definition 1.} A \textit{numeration system} is an ordered pair $(S, N)$, where $S$ is an infinite subset of $(\Sigma_2^+ \setminus \Omega \Sigma_2^*) \cup \{0\}$ and $N$ is the natural order preserving bijection from $S$ to $\N$ that maps the $(n + 1)^{st}$ word of $S$ to $n$. The map $N$ is referred to as the \textit{evaluation map}, and $N^{-1}$ is the \textit{representation map}. If $N(v) = n$, then $v$ is the $S$-representation of $n$.

A few remarks about this definition are in order. First, while the definition generalizes naturally to allow for more general digit sets, we will consider only numeration systems with digit set $\Sigma_2$. Also, since the evaluation map will always be obtained from the radix order, we therefore write $S$ instead of $(S, N)$. Second, notice that our definition requires that $0 \in S$. While this restriction is not necessary in general, it is natural for our purposes. Finally, our condition that $S \subseteq (\Sigma_2^+ \setminus \Omega \Sigma_2^*) \cup \{0\}$ ensures that for $n \geq 1$ the $S$-representation of $n$ begins with 1. More general numeration systems that allow for leading zeros have been considered elsewhere in the literature (e.g. [12], [14]), but will not be considered here.

Suppose $S$ is a numeration system, and let $B = \{b_i\}_{i=0}^{\infty}$ be an arbitrary sequence of natural numbers. Let $\pi_B : S \to \N$ be the function that assigns $w = w_k w_{k-1} \cdots w_0 \in S$ to

$$\pi_B(w) = \sum_{i=0}^{k} b_i w_i$$

(1)

A numeration system $S$ is \textit{based} if there exists a strictly increasing sequence $B = \{b_i\}_{i=0}^{\infty}$ of natural numbers, with $b_0 = 1$, such that $N(w) = \pi_B(w)$ for all $w \in S$. In this case, the sequence $\{b_i\}$ is called the \textit{base sequence} of $S$. The following lemma will be important in Sections 3 and 4.

\textbf{Lemma 1.} If $S$ is a based numeration system, then $1 \in S$.

\textit{Proof.} Let $B = \{b_i\}$ be the base sequence. By definition $b_0 = 1$. Since $S$ is a numeration system, there exists a $w \in S$ for which $N(w) = 1$. Since $S$ is based, $\pi_B(w) = N(w) = 1$. Since $\{b_i\}$ is an increasing sequence, if $|w| > 1$ then $\pi_B(w) > 1$. It follows then that $|w| = 1$, and therefore $w = 1$. Thus, $1 \in S$. \qed

We say a based numeration system $S$ is greedy if whenever $w \in S \setminus \{0\}$ and $v \in \Sigma_2^* \setminus \{0\}$ with

$$N(w) = \pi_B(w) = \pi_B(v),$$

it follows that $w \geq v$ under the radix order. In other words, for every $n \geq 1$ the $S$-representation of $n$ is the largest possible representation under the radix order. The
numeration system \( S \) is said to be lazy if for every \( n \geq 1 \) the \( S \)-representation of \( n \) is the smallest possible representation. The greedy and lazy Fibonacci representations mentioned in Section 1 provide examples of these definitions. In general, if \( \{ b_i \} \) is the base sequence in a greedy numeration system \( S \) and if \( \sup_{i \geq 0} \frac{b_{i+1}}{b_i} < 2 \) then the digits in the \( S \)-representation for \( n \in \mathbb{N} \) can be obtained by the algorithm given in Section 1.

It is natural to consider numeration systems with the property that \( w0 \in S \setminus \{0\} \) whenever \( w \in S \setminus \{0\} \). Such a numeration system is said to be right extendable (see [13], Section 7.3.2). Similarly, \( S \) is a Bertrand numeration system if \( w \in S \setminus \{0\} \Leftrightarrow w0 \in S \setminus \{0\} \) \[2\]. The set of greedy Fibonacci representations of the natural numbers mentioned in Section 1 is a Bertrand numeration system, while the set of lazy Fibonacci representations is not even right extendable. The following lemma gives a necessary condition for a numeration system that is also right extendable to be based.

**Lemma 2.** Let \( S \) be a right extendable based numeration system. If \( w, w1 \in S \) then \( w01 \in S \).

**Proof.** Let \( B = \{ b_i \} \) be the base sequence for \( S \). Since \( S \) is right extendable, it follows that \( w00 \in S \) and \( w10 \in S \). It must be true then that

\[
b_1 = \pi_B(10) - \pi_B(0) = \pi_B(w10) - \pi_B(w00) = N(w10) - N(w00).
\]

If \( w01 \notin S \), then \( N(w10) - N(w00) = 1 \), which implies that \( b_1 = 1 \). However, since \( b_1 > 1 \) we have a contradiction, and so \( w01 \in S \).

To any numeration system \( S \) we associate a graph \( T(S) \) with vertex set \( S \) and edge set

\[
\{(v, v\eta) : \eta \in \{0, 1\} \text{ and } v, v\eta \in S\}.
\]

In other words, we draw an edge between \( v \) and \( v\eta \) whenever both are members of \( S \). This graph structure clearly partitions \( S \) into a collection of trees. To every tree \( \tau \) in \( T(S) \) we define the root of \( \tau \) to be the vertex in \( \tau \) of minimal length under the radix order. Notice that by definition every vertex in \( T(S) \) has at most two children. We say that for \( w \in S \), if \( w0 \in S \), then \( w0 \) is the left child of \( w \), and if \( w1 \in S \), then \( w1 \) is the right child of \( w \). Whenever \( S \) is right extendable, every vertex in \( T(S) \) has at least one child. Numeration systems in which \( T(S) \) is a single rooted tree are of particular interest.

**Definition 2.** Let \( S \) be a Bertrand numeration system, with \( 1 \in S \). If \( T(S) \) is such that \( w \in S \) whenever \( w1 \in S \), then \( S \) is treelike.

The conditions of the definition guarantee that \( T(S) \) is a single rooted tree. These conditions also imply that if \( w \in S \) then every prefix of \( w \) is in \( S \). Treelike numeration systems were introduced in [5]. As an example, Figure 1 shows the first few levels of the tree for the greedy Fibonacci numeration system mentioned in the
introduction. The vertices in the tree of Figure 1 are the greedy Fibonacci representations of the natural numbers. The numbers in parentheses are the standard base 10 values of these representations.

\[
\begin{align*}
0, & \ (0) \\
1, & \ (1) \\
10, & \ (2) \\
100, & \ (3) \\
1000, & \ (5) \\
10000, & \ (8) \\
\end{align*}
\]

Figure 1: The tree \( T(S) \) for the greedy Fibonacci numeration system.

In a given numeration system \( S \), for \( k \geq 0 \) we define \( M_k \) to be the word in \( S \) having \( k \) digits that is maximal under the radix order. We take \( M_0 \) to be the the empty word. Similarly, for \( k \geq 1 \) let \( m_k \) be the minimal word having \( k \) digits in \( S \). If \( 1 \in S \) it will be convenient to define \( m_1 \) to be 1. If \( S \) has the property that \( M_k \) is a prefix of \( M_{k+1} \) for every \( k \), we let

\[
M = \lim_{k \to \infty} M_k.
\]  

This \( M \) is then referred to as the maximal word associated with \( S \). We define the minimal word \( m \) of \( S \) similarly. It is clear that in a treelike numeration system, \( m_k = 10^{k-1} \).

Our first theorem establishes a necessary and sufficient condition for a numeration system to be based. We point out that the theorem actually resembles Theorem 5.2 of [5]. That theorem is restricted to treelike numeration systems, and so the following theorem is more general. We also comment that the hypotheses of the theorem guarantee that any two words of length at least 2 share a common nonempty prefix.

**Theorem 1.** Let \( S \) be a right extendable numeration system, and assume \( 1 \in S \). Then \( S \) is based if and only if for any two consecutive words \( v, w \in S \) of length \( l \geq 2 \) having maximal common prefix \( p \) it follows that

\[
\begin{align*}
v & = p0M_{l-|p|-1} \\
w & = pm_{l-|p|}.
\end{align*}
\]

**Proof.** Suppose first that \( S \) is based, with base \( B = \{b_i\} \). Consider two arbitrary adjacent words \( v \) and \( w \) in \( S \) of the same length \( l \geq 2 \), with \( v < w \). The property
of the theorem clearly holds if \( l = 2 \). Suppose that \( l \geq 3 \), and let \( p \) be the maximal common prefix of \( v \) and \( w \). Then we can write

\[
v = p0x_{k-2} \cdots x_0 \quad \text{and} \quad w = p1y_{k-2} \cdots y_0,
\]

where \( k = l - |p| \). Note that \( k \geq 1 \). If \( k = 1 \), then \( v = p0 = p0M_0 \) and \( w = p1 = pm_1 \), and the property of the theorem is satisfied. Assume then that \( k \geq 2 \). Since \( v \) and \( w \) are adjacent in \( S \), it is clear that \( N(w) - N(v) = 1 \). Since the numeration system is based, it follows that \( N(v) = \pi_B(v) \) and \( N(w) = \pi_B(w) \), where \( \pi_B \) is as defined in (1). Thus

\[
1 = \pi_B(w) - \pi_B(v) = \pi_B(p1y_{k-2} \cdots y_0) - \pi_B(p0x_{k-2} \cdots x_0) \\
= \pi_B(1y_{k-2} \cdots y_0) - \pi_B(x_{k-2} \cdots x_0).
\]

Since the word \( x_{k-2} \cdots x_0 \) has \( k - 1 \) digits and the word \( 1y_{k-2} \cdots y_0 \) has \( k \) digits, it must be true that

\[
x_{k-2} \cdots x_0 = M_{k-1} = M_{l-|p|-1} \quad \text{and} \quad 1y_{k-2} \cdots y_0 = m_k = m_{l-|p|}.
\]

Now suppose that \( S \) has the property mentioned in the statement of the theorem. For \( i \geq 0 \) let \( b_i = N(10^i) \). We will show that \( \{b_i\} \) is a base for the numeration system. We need to show that \( N(w) = \pi_B(w) \) for every \( w \in S \). We use induction on the length of \( w \). Clearly \( N(0) = 0 = \pi_B(0) \), and \( N(1) = 1 = \pi_B(1) \). Let \( l \geq 2 \), and assume that \( N(w) = \pi_B(w) \) whenever \( |w| < l \). Let \( w_1, \ldots, w_m \) be the words in \( S \) of length \( l \) in increasing lexicographic order under \( < \). We need to show that \( N(w_j) = \pi_B(w_j) \) for \( 1 \leq j \leq m \). Since \( S \) is right extendable it follows immediately that \( w_1 = 10 \cdots 0 = 10^{l-1} \). Therefore \( N(w_1) = b_{l-1} = \pi_B(w_1) \). Assume \( 1 \leq j < m \), and that

\[
N(w_1) = \pi_B(w_1), N(w_2) = \pi_B(w_2), \ldots, N(w_j) = \pi_B(w_j).
\]

We need to show that \( N(w_{j+1}) = \pi_B(w_{j+1}) \). Let \( p \) be the maximal common prefix for \( w_j \) and \( w_{j+1} \). If \( w_j = p0 \) and \( w_{j+1} = p1 \), then \( \pi_B(w_{j+1}) = \pi_B(w_j) + 1 = N(w_j) + 1 = N(w_{j+1}) \). Assume then that

\[
w_j = p0x \quad \text{and} \quad w_{j+1} = p1y,
\]

where \( x, y \in \{0, 1\}^* \) and \( 1 \leq |x|, |y| \leq l - |p| - 1 \). Since \( w_j \) and \( w_{j+1} \) are adjacent and since \( S \) is right extendable it follows from the property in the statement of the theorem that

\[
x = M_{l-|p|-1} \quad \text{and} \quad y = 0^{l-|p|-1}
\]

Thus, by the induction hypothesis,

\[
\pi_B(w_{j+1}) - \pi_B(w_j) = \pi_B(p1y) - \pi_B(p0x) = \pi_B(1y) - \pi_B(x) = N(M_{l-|p|-1}) - N(m_{l-|p|}) = 1.
\]

Since \( N(w_{j+1}) - N(w_j) = 1 \), and \( N(w_j) = \pi_B(w_j) \) it follows that,

\[
N(w_{j+1}) = \pi_B(w_{j+1})
\]

So for all \( w \) of length \( l \), \( N(w) = \pi_B(w) \). And by induction the result follows. \( \square \)
3. Self Generating Numeration Systems

We now return to self generating sets. For a family of functions \( F \), a self generating set \( S = S_F \) is the smallest set containing 0 that is closed under the functions in \( F \). A numeration system \( S \) is *self generating* if there exists a collection of functions \( F \) for which

\[
S = \{ [m]_2 : m \in S_F \},
\]

where \([m]_2\) is the base 2 expansion of the integer \( m \in S_F \). As mentioned in Section 1, the sets of greedy and lazy representations of the natural numbers with respect to both the Fibonacci and Tribonacci sequences give rise to numeration systems that are self generating and based.

In light of these examples it is natural to consider the question of which sets of affine functions produce self generating numeration systems that are based. In this section we will show how to construct a large class of such functions. These generating functions are perhaps the most natural generalization of the sets of functions generating the greedy Fibonacci and Tribonacci numeration systems. First, we require that \( 2x \in F \). Since multiplication of \( x \) by 2 adds a 0 to the binary expansion of \( x \), the resulting numeration system will be right extendable. Furthermore, since multiplication of \( x \) by \( 2^n \) adds \( n \) zeros to the binary expansion of \( x \), it is natural to restrict our attention to those families of functions in which the coefficient on \( x \) is a power of 2. We also require that if \( f(x) = 2^n x + c \in F \), then \( 0 \leq c < 2^n \). Thus, the binary expansion of \( f(x) \) ends in the binary expansion of \( c \). Our next lemma gives us a further restriction on \( F \).

**Lemma 3.** Let \( F = \{ f_0, f_1, \ldots, f_n \} \) be a family of functions defined as follows. Let \( f_0 = 2x \) and for \( 1 \leq i \leq n \) let \( f_i(x) = 2^{k_i} x + c_i \), where \( k_i \geq 1 \) and \( 0 \leq c_i < 2^{k_i} \). Let \( S_F \) be the set generated by \( F \), and let \( S \) be the numeration system \( \{ [s]_2 : s \in S_F \} \). If \( S \) is based then \( c_i = 1 \) for some \( i \).

**Proof.** If \( S \) is based, then by Lemma 1 \( S \) must contain 1. By definition of \( S \) it follows that \( 1 \in S_F \). Thus, \( 1 = f_i(0) \) for some \( f_i \in F \) with \( 1 \leq i \leq n \). \( \square \)

In light of Lemma 3 we add the assumption that if \( 2^k \) is the smallest coefficient on \( x \) for all functions in \( F \) that have an odd constant term, then \( 2^k x + 1 \in F \). We will prove that a finite number of iterations of this “minimal” function on the function \( 2x \) produces a family of functions \( F \) that generates a based numeration system \( S \). Our method will be to show that the self generating numeration system \( S \) satisfies the conditions of Theorem 1. Our first step in this process is to show that \( S \) is treelike.

**Lemma 4.** Let \( n \geq 1 \), \( k \geq 2 \), and let \( F \) be the set of functions \( \{ f_0, \ldots, f_n \} \) where \( f_0(x) = 2x \), and

\[
f_i(x) = 2^{k-1}(f_{i-1}(x)) + 1 \quad \text{for} \quad 1 \leq i \leq n.
\]  

(4)
Let $S_F$ be the set generated by $F$, and let $S$ be the numeration system $\{[s]_2 : s \in S_F\}$. Then $S$ is tree-like.

Proof. Clearly $1 \in S$. Since $2x \in F$, $S$ is right extendable. Suppose $v_0 \in S$. Then there is a function $f \in F$ and an $s \in S_F$ such that $[f(s)]_2 = v_0$. Since $f(s)$ is even, it follows that $f = f_0$. Then $v = [s]_2 \in S$, and therefore $S$ is a Bertrand numeration system. Now let $v \in \{0, 1\}^*$ be such that $v_1 \in S$. We need to show that $v \in S$. By definition, there exists an $s \in S_F$ and an $f_i \in F$, where $1 \leq i \leq n$, such that $[f_i(s)]_2 = v_1$. Thus, $[2^{k-1}f_{i-1}(s) + 1]_2 = v_1$, and so $[2^{k-1}f_{i-1}(s)]_2 = v_0$, and $[2^{k-2}f_{i-1}(s)]_2 = v$. Notice that $v = [f_{i}^{k-2}(f_{i-1}(s))]_2$. It follows from the definition of $S$ that $v \in S$. This completes the proof that $S$ is tree-like.

The next lemma enables us to establish that the property of Theorem 1 is satisfied. If $v \in S$ and if there exists an $f \in F$ and an $s \in S_F$ such that $v = [f(s)]_2$ then we say that $v$ is produced by $f$. Recall also that for $j \geq 0$, $M_j$ is the maximal word of length $j$ in $S$.

**Lemma 5.** Let $n \geq 1$, $k \geq 2$, and let $F$ be the set of functions $\{f_0, \ldots, f_n\}$ where $f_0(x) = 2x$, and

$$f_i(x) = 2^{k-1}(f_{i-1}(x)) + 1 \quad \text{for} \quad 1 \leq i \leq n.$$

Let $S$ be the self generating numeration system generated by $F$. Let $v \in S$, and assume $v_1 \in S$. If $v_1$ is produced by $f_1$, then for $j \geq 1$, $vM_j$ is the maximal sequence in $S$ of length $|v| + j$ having $v$ as a prefix.

Proof. Since $S$ is tree-like, a maximal sequence in $S$ having $v$ as a prefix corresponds to a path in $T(S)$ beginning at $v$ and following the rightmost branches at each level. Let $s \in S$ be such that $v_1 = [f_1(s)]_2 = [2^k s + 1]_2 = [s]_2 0^{k-1}$. Let $v_0 = [s]_2$, and define

$$v_i = [2^{k-1}f_{i-1}(s) + 1]_2 = [f_i(s)]_2 \quad \text{for} \quad 1 \leq i \leq n.$$

Also, let $v_{n+1} = [2^{k-1}f_n(s)]_2$. Notice that

$$v_i = [f_i(s)]_2 = [2^{k-1}f_{i-1}(s) + 1]_2 = v_{i-1} 0^{k-1} \quad \text{for} \quad 1 \leq i \leq n,$$

and $v_{n+1} = v_0 0^{k-1}$. It follows from the definition of $T(S)$ then that for $1 \leq i \leq n$, $v_i$ is a right child of $[2^{k-2}f_{i-1}(s)]_2$ in $S$. Also, for any integer $l \geq 1$, $[2^lf_{i-1}(s)]_2$ is a left child of $[2^{l-1}f_{i-1}(s)]_2$ in $S$. Thus, for $1 \leq i \leq n$ it follows that the $v_i$ are connected by a path from $v$ to $v_{n+1}$ of length $(k-1)(n-1) + k$. For $1 \leq i \leq n$, the $v_i$ are right children in this path, and all the other vertices in this path are left children. Let $v x_1 x_2 \cdots x_j$ be the $j$th vertex in this path. We show that this is the maximal word in $S$ of length $|v| + j$ having $v$ as a prefix.

Assume to the contrary that $v y_1 y_2 \cdots y_j \in S$ is lexicographically larger than $v x_1 x_2 \cdots x_j$. We will consider two cases. First, suppose that $v x_1 \cdots x_j$ lies on the path from $v$ to $v_n$. Let $l$ be the smallest integer such that $y_l \neq x_l$. This gives that

$$x_1 = y_1, \quad x_2 = y_2, \quad \ldots, \quad x_{l-1} = y_{l-1}, \quad x_l = 0, \quad \text{and} \quad y_l = 1.$$
By Lemma 4, $\mathcal{S}$ is treelike, so every prefix of a word in $\mathcal{S}$ is also in $\mathcal{S}$. Thus there exists a $y \in \mathcal{S}$ and an integer $m$, with $1 \leq m \leq n$, such that $vy_1 \cdots y_l = [f_m(y)]_2$. Now, $f_m(y) = 2^{k-1}f_{m-1}(y) + 1$, and so by (5) we have that $[2^{k-1}f_{m-1}(y)]_2 = vy_1y_2 \cdots y_{l-1}0 = vx_1x_2 \cdots x_l$. However, since $vx_1 \cdots x_l$ is a left vertex in the aforementioned path from $v$ to $v_{n}$, there exist $s \in S_F$ and integers $i$ and $t$ with $1 \leq i < n$ and $1 \leq t \leq k - 2$ such that $vx_1 \cdots x_l = [2^t f_i(s)]_2$. Thus, $2^{k-1}f_{m-1}(y) = 2^t f_i(s)$. Therefore $2^{k-t-1}f_{m-1}(y) = f_i(s)$. Since $k - t - 1 \geq 1$, this is a contradiction, since $f_i(s)$ is odd for $i \geq 1$. Thus, $vx_1 \cdots x_l$ is the maximal word in the path from $v$ to $v_n$.

Now, assume that $vx_1 \cdots x_l$ lies on the path from $v$ to $v_{n+1}$. Then the argument of the previous paragraph holds, except that $i = n$, and $1 \leq t \leq k - 1$. We have shown that for $j \leq (k - 1)(n - 1) + k$, $vx_1 \cdots x_j$ is the maximal word in $\mathcal{S}$ having $v$ as a prefix. Notice that our argument reveals that the $x_j$'s depend on the functions in $F$, and not on $v$. Thus, if we take $v = 0$ we see that $x_1 \cdots x_j = M_j$ for $1 \leq j \leq (k - 1)(n - 1) + k$.

So far we have restricted our attention to the case that $j$ is less than the length of $[f_1(f_n(0))]_2$. Since $v_{n+1}$ is produced by $f_1$, the above argument can be extended to arbitrary $j$.

**Theorem 2.** Let $n \geq 1$, $k \geq 2$, and let $F$ be the set of functions $\{f_0, \ldots, f_n\}$ where $f_0(x) = 2x$, and

$$f_i(x) = 2^{k-1}(f_{i-1}(x)) + 1 \text{ for } 1 \leq i \leq n.$$ 

Let $S_F$ be the set generated by $F$, and let $\mathcal{S}$ be the numeration system $\{[s]_2 : s \in S_F\}$. Then $\mathcal{S}$ is based.

**Proof.** Let $v$ and $w$ be adjacent words in $\mathcal{S}$ of the same length. Assume that $v < w$. Let $p$ be the maximal common prefix of $v$ and $w$ in $\mathcal{S}$. Then $v = p\alpha x$ and $w = p\gamma y$ for some $x, y \in \{0, 1\}^*$. Since $\mathcal{S}$ is treelike, every prefix of a word in $\mathcal{S}$ is also in $\mathcal{S}$. Thus, $p1 \in \mathcal{S}$ and since $\mathcal{S}$ is self generating, there exists an $s \in \mathcal{S}$ and an $f_i \in F$ such that $[f_i(s)]_2 = p1$. By definition we have that

$$p1 = [f_i(s)]_2 = [2^{k-1}f_{i-1}(s) + 1]_2.$$ 

Therefore,

$$[2^{k-1}f_{i-1}(s)]_2 = p0,$$

and so

$$[f_1(f_{i-1}(s))]_2 = [2^k(f_{i-1}(s)) + 1]_2 = [2(2^{k-1}f_{i-1}(s)) + 1]_2 = p01.$$ 

Therefore $p01 \in \mathcal{S}$ and is produced by $f_1$. Thus, by Lemma 5, it follows that $v = p0M_{|w|-|p|-1}$. Since $\mathcal{S}$ is right extendable, it is clear that $w = pm_{|w|-|p|}$. Thus, by Theorem 1, it follows that $\mathcal{S}$ is based. 

\[\Box\]
We conclude this section with a few examples. Note that in each case the elements in the base sequence are obtained from the indices of the powers of 2 in the sequence of ordered elements of $S_F$.

**Example 1.** Let $F = \{2x, 8x + 1\}$. Then

\[ S = \{0, 1, 2, 4, 8, 9, 10, 16, 17, 18, 32, 33, 34, 36, \ldots\} . \]

The numeration system $S$ is defined as the set of base two representations of the elements of $S_F$. The elements of $S$ correspond to the greedy expansions of the natural numbers with respect to the base $b_0 = 1$, $b_1 = 2$, $b_2 = 3$, and $b_n = b_{n-1} + b_{n-3}$ for $n \geq 3$.

**Example 2.** In general, if $F = \{2x, 2^k x + 1\}$, then $S_F$ is the set of nonnegative integers whose base two expansions correspond to the greedy expansions of the natural numbers with respect to the base sequence $b_0 = 1$, $b_1 = 2$, \ldots, $b_{k-1} = k-1$, and $b_n = b_{n-1} + b_{n-k}$ for $n \geq k$.

**Example 3.** Let $F = \{2x, 8x + 1, 32x + 5\}$. Then

\[ S_F = \{0, 1, 2, 4, 5, 8, 9, 10, 16, 17, 18, 20, 32, 33, 34, 36, 37, 40, 41, \ldots\} . \]

The numeration system $S$ consists of the greedy expansions of the natural numbers with respect to the base $b_0 = 1$, $b_1 = 2$, $b_2 = 3$, $b_3 = 5$, $b_4 = 8$, and $b_n = b_{n-1} + b_{n-3} + b_{n-5}$ for $n \geq 5$.

The generating functions used in Theorem 2 generate a numeration system that is treelike. It is possible to construct examples of based self-generating numeration systems for which this is not the case. The following example is perhaps the simplest such example.

**Example 4.** Let $F = \{2x, 4x + 1, 4x + 2\}$. Then

\[ S_F = \{0, 1, 2, 4, 5, 6, 8, 9, 10, 12, 16, 17, 18, 20, 21, 22, 24, 25, 26, \ldots\} . \]

The numeration system $S$ is defined as the set of base two representations of the elements of $S_F$. The first 18 elements of $S$ are shown in Figure 2, where we leave out the root vertex at 0 for the sake of brevity.

We have listed the elements of $S$ in rows, where the elements of a given row $k$ are all the elements of $S$ of length $k$. The tree connections between vertices are shown. The elements in the last row that are not connected to anything are roots of new trees in $T(S)$. It is interesting to note that the rooted trees in Figure 2 are all isomorphic. It is a routine exercise to prove a result similar to Lemma 5 for this $S$, and therefore the proof that $S$ is based is similar to the proof of Theorem 2. We can also compute the base and see that the elements of $S$ correspond to the greedy expansions of the natural numbers with respect to the base $b_0 = 1$, $b_1 = 2$, $b_2 = 3$, and $b_n = b_{n-1} + 2b_{n-2} - b_{n-3}$ for $n \geq 3$. 


4. Linearly Recurrent Base Sequences

We now show that the base sequence in any based self generating numeration system satisfies a linear recurrence. In [5] it was shown that the base sequence in any based treelike numeration system satisfies a linear recurrence if and only if the maximal sequence $M$ defined in (2) is periodic. As example 4 shows, a self generating numeration system need not be treelike. Shallit [15] also established some rather general conditions which guarantee that the base sequence in a based numeration system satisfies a linear recurrence. It can be shown that the self generating numeration systems we are considering satisfy these conditions. In this section we give a different proof which gives us a relatively simple means for constructing the linear recurrence. Note also that in the following theorem, we are no longer restricting our attention to the family $F$ given by (4).

**Theorem 3.** Let $F = \{f_0, f_1, \ldots, f_n\}$ be a family of functions defined as follows. Let $f_0(x) = 2x$, and for $1 \leq i \leq m$, let $f_i(x) = 2^{k_i}x + c_i$, where $k_i \geq 1$ and $0 \leq c_i < 2^{k_i}$. Let $S$ be the numeration system generated by $F$. If $S$ is based, then the base sequence $\{b_i\}_{i=0}^\infty$ satisfies a linear recurrence.

**Proof.** Let $\{s_i\}_{i=0}^\infty$ be the sequence of elements of $S$ listed in order. Since $S$ is based, by Lemma 3 it follows that $c_i = 1$ for some $i$ with $1 \leq i \leq n$. By definition it follows that for $n \geq 0$, $b_n$ is the index of $2^n$ in $\{s_i\}$. We need to show that $\{b_i\}$ satisfies a linear recurrence.

Let $\{u_i\}_{i=0}^\infty$ be the characteristic sequence of $S_F$. That is, for $i \geq 0$, $u_i = 1$ if $i \in S_F$ and $u_i = 0$ otherwise. It was shown in [8] that the characteristic sequence is 2-automatic (for a definition of automatic sequences see [1]). By Cobham’s Theorem [6], (see also Theorem 6.3.2 of [1]), $\{u_i\}$ is the image under a coding of the fixed point of a morphism $\sigma$ of constant length 2 over a finite alphabet $\{a_1, \ldots, a_k\}$. Let $\sigma : \{a_1, a_2, \ldots, a_k\}^* \rightarrow \{a_1, a_2, \ldots, a_k\}^*$ be this morphism, and suppose the fixed point is generated by iteration on $a_1$. Recall that the incidence matrix of $\sigma$ is defined as the $n$ by $n$ matrix $A = (m_{i,j})$, where $m_{ij} = |\sigma(a_j)|_{a_i}$, the number of occurrences
of $a_i$ in $\sigma(a_j)$ (see [1], chapter 8). Let $A$ be the incidence matrix of $\sigma$ and let

$$P(x) = x^k - c_1x^{k-1} - c_2x^{k-2} - \cdots - c_{k-1}x - c_k$$

be the characteristic polynomial of $A$. Finally, let $\{r_n\}$ be the sequence defined by the recurrence relation

$$r_n = c_1r_{n-1} + c_2r_{n-2} + \cdots + c_{k-1}r_{n-k}. \quad (6)$$

We show that $\{b_i\}$ satisfies this recurrence.

Let $e_1 = [1\ 0\ \cdots\ 0]^T$. We first show that for $1 \leq i \leq k$ the $i^{th}$ coordinate of $A^ne_1$ satisfies (6). By the Cayley-Hamilton Theorem $P(A) = 0$, and so

$$A^k - c_1A^{k-1} - c_2A^{k-2} - \cdots - c_{k-1}A - c_kI = 0.$$ 

Therefore $P(A) \cdot e_1 = 0$, and so

$$A^k e_1 - c_1A^{k-1} e_1 - c_2A^{k-2} e_1 - \cdots - c_{k-1}A e_1 - c_k I e_1 = 0.$$ 

For $n \geq k$, multiplying both sides of this last equation by $A^{n-k}$ gives

$$A^n e_1 - c_1A^{n-1} e_1 - c_2A^{n-2} e_1 - \cdots - c_{k-1} A^{n-k+1} e_1 - c_k A^{n-k} e_1 = 0.$$ 

Now, by definition of $A$, $|\sigma^n(a_i)|_{a_i}$, the number of occurrences of $a_i$ in $\sigma^n(a_1)$, is the $i^{th}$ coordinate of $A^ne_1$. Thus, for $n \geq 0$, $|\sigma^n(a_1)|_{a_i}$ satisfies the recurrence (6) with initial conditions $|\sigma(a_1)|_{a_i}, |\sigma^2(a_1)|_{a_i}, \ldots, |\sigma^k(a_1)|_{a_i}$.

Now, for $n \geq 0$ we have that $b_n$ is the index of $2^n$ in $\{s_i\}$. Equivalently, since $\{b_i\}$ is the base system for the abstract numeration system $\mathcal{S}$, $b_n$ is the number of ones in the sequence $\{u_n\}$ between $u_0$ and $u_{2^n-1}$. On the other hand, since $\sigma$ has constant length 2, it follows that

$$\sum_{i=1}^{k} |\sigma^n(a_1)|_{a_i} = |\sigma^n(a_1)| = 2^n.$$ 

Now let $a_{i_1}, a_{i_2}, \ldots, a_{i_l}$, where $1 \leq l \leq k$, be the characters that map to 1 under the coding that maps the fixed point of $\sigma$ to $\{u_n\}$. We therefore have that

$$b_n = |\sigma^n(a_1)|_{a_{i_1}} + \cdots + |\sigma^n(a_1)|_{a_{i_l}}.$$ 

Since each of $|\sigma^n(a_1)|_{a_{i_j}}$ satisfies (6), it follows that $b_n$ satisfies (6) as well. \hfill \square

5. Further Considerations

We close with some suggestions for further study. There are many examples of sets of generating functions that do not satisfy the conditions of Theorem 2 that seem
to generate based numeration systems. For example, numerical evidence seems to suggest that if $F = \{2x, 4x + 1, 16x + 3, 32x + 11\}$ then the resulting numeration system is based. In particular, the base two expansions of the elements of $S_F$ seem to correspond to the set of greedy representations of $\mathbb{N}$ with respect to the sequence $b_0 = 1, b_1 = 2, b_2 = 4, b_3 = 7, b_4 = 13,$ and $b_n = b_{n-1} + b_{n-2} + b_{n-4} + b_{n-5}.$ On the other hand, not every self generating numeration system is based. For example, it follows from Lemma 2 that the numeration system generated by $F = \{2x, 8x + 3\}$ is not based. In light of these examples we see that the characterization of all self generating based numeration systems remains an open problem.

It is also natural to try to extend the results of this paper to numeration systems with digit sets $\Sigma_k = \{0, 1, 2, \ldots, k - 1\}.$ Some examples were considered in [8]. For example, if $F = \{3x + 1, 3x + 2, 9x + 3, 9x + 6\}$ then

$$S_F = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 17, 19, 20, 21 \ldots \}.$$ 

This is the set of all nonnegative integers whose base 3 expansion does not contain the block 00. If $S = \{[s]_3 : s \in S_F\},$ where $[s]_3$ denotes the base 3 expansion of $s,$ then $S$ is a numeration system with digit set $\Sigma_3.$ It was noted in [8] that $S$ corresponds to the set of lazy representations of the natural numbers with respect to the base sequence $b_0 = 1, b_2 = 3,$ and $b_n = 2b_{n-1} + 2b_{n-2}$ for $n \geq 2.$

Finally, the reader may have noticed that Theorem 2 does not guarantee that the numeration systems are greedy with respect to their bases. It is not hard to verify this for specific examples. However, the question of which based numeration systems are greedy and which are lazy also remains open.

Acknowledgments The work of the first and second authors was supported in part by NSF Grant No. 0431664. We also would like to thank the referee for a thorough reading of the manuscript and for many helpful suggestions.

References


