Contributions to Formal Language Theory: Fixed Points, Complexity and Context-Free Sequences

by

Dave Hamm

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Masters of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 1998

Contents

Chapter 1

Introduction

The central issue in formal language theory is the finite specification of languages. A finite language can be finitely specified by simply enumerating the words in the language. The issue only becomes challenging when we specify infinite languages.

Let us be more precise about what is meant by "finite specification of a language". First we need to introduce some notation, some of which is standard and can be found in [27].

An *alphabet* is a set of symbols. The alphabets considered here are always finite nonempty sets. The elements of an alphabet Σ are called *letters*. A *word* over an alphabet Σ is a string consisting of zero of more letters of Σ. The string consisting of zero letters is called the *empty word*, written ϵ. The set of all nonempty finite words over an alphabet Σ is denoted Σ^+ and $\Sigma^* = \Sigma^+ \cup \{\epsilon\}$.

For single letters, that is, elements of Σ, we use the lower case letters a, b, c, d. For finite words, we use the lower case letters u, v, w, x, y, z. For infinite words, we use bold-face letters $\mathbf{t}, \mathbf{u}, \mathbf{v}, \mathbf{w}, \mathbf{x}, \mathbf{y}, \mathbf{z}$. If $w \in \Sigma^*$, then by $|w|$ we mean the length of, or number of symbols in w. If S is a set, then by Card S we mean the number of elements of S. We say $x \in \Sigma^*$ is a *subword* of $y \in \Sigma^*$ if there exist words $w, z \in \Sigma^*$ such that $y = wxz$.

Subsets of Σ^* are referred to as *languages* over Σ. Thus, for $\Sigma \subseteq \Sigma_1$, if L is a language over Σ, it is also a language over Σ_1. $\overline{L}$ is the *complement* of L. Given a word $w \in \Sigma^*$, $w \in \overline{L}$ if and only if $w \notin L$.

Any finite specification of a language must itself be a string over some alphabet Σ. Furthermore, different languages certainly must have different specifications, otherwise the term "specification" is not appropriate. These two restrictions severely limit the scope of finite specifications for the following reason. Assuming that all alphabets are subsets of some countably infinite set of symbols A, then all finite specifications are elements of A^*. However, the number of finite words over a countably infinite set is countably infinite. Therefore only a countably infinite number of languages can be finitely specified. Now note that the set of all possible languages over a given alphabet Σ (that is, 2^{Σ^*}) is uncountably infinite. With only a countable number of specifications and an uncountable number of languages to specify, there are an uncountably infinite number of languages that cannot be specified finitely.

Hence, the best we can hope for is that we can find finite specifications for at least the more interesting languages.

Naturally, it is not possible to give an example of a language that cannot be specified finitely. We can only exhibit languages that cannot be specified with a given representational scheme.

In this thesis, we examine several problems which are directly or indirectly related to representational schemes. The following is a description of the main results.

In chapter 2, we present examples of two *k-quasi-context-free sequences* which have exponential *subword complexity.* One of these sequences has maximum *subword complexity. Subword complexity* and *k-quasi-context-free sequences* are defined later.

In chapter 3, we present a characterization of *finite fixed points.* Theorem 3.0.4 shows that given a *homomorphism* h and a word w, $h(w) = w$ if and only if w belongs to a given set. Theorem 3.0.6 gives the best possible upper bound on the shortest nonempty finite fixed point of a homomorphism. *Finite fixed points* and *homomorphisms* are defined later.

In chapter 4, we present a characterization of *infinite fixed points*, due to Jeff Shallit. Theorem 4.0.8 shows that given a *homomorphism* h and an infinite word $\mathbf{w}$, $h(\mathbf{w}) = \mathbf{w}$ if and only certain conditions are met.

In chapter 5, we improve upon a result of Ehrenfeucht, Lee and Rozenberg [21, Lemma 1, p. 64] giving a smaller upper bound on the maximum period of the prefixes (and suffixes) of a *PD0L sequence* over Σ. Interestingly, this upper bound is related to the maximum order of permutations over $|\Sigma|$ elements. Furthermore we show that this new upper bound is the best possible.

In chapter 6, we present some open problems related to the results given in this thesis.

We now define the notions of subword complexity and homomorphism.

1.1 Subword Complexity

Let $\mathbb{N}$ denote the set of non-negative integers and let Σ be an alphabet. By a *sequence over* Σ we mean a function $\mathbf{x}$ from $\mathbb{N}$ into Σ; less formally, a sequence is a string of symbols, infinite to the right,

$$\mathbf{x} = \mathbf{x}(0)\mathbf{x}(1)\mathbf{x}(2)\cdots\mathbf{x}(i)\cdots$$

indexed by the non-negative integers. Sometimes we write $\mathbf{x}_i$ instead of $\mathbf{x}(i)$.

A sequence

$$\mathbf{x} = \mathbf{x}(0)\mathbf{x}(1)\mathbf{x}(2)\cdots\mathbf{x}(i)\cdots$$

is *ultimately periodic* if, for some $i \geq 0$ and some $k > 0$, $\mathbf{x}(i) = \mathbf{x}(i + ck)$ for all $c \geq 0$.

Denote by $\mathrm{Sub}_k(\mathbf{x})$ the set of all subwords of $\mathbf{x}$ of length k, so

$$\mathrm{Sub}_k(\mathbf{x}) = \{w \in \Sigma^k \mid \exists\, n \text{ such that } w = \mathbf{x}(n)\mathbf{x}(n+1)\cdots\mathbf{x}(n+k-1)\}.$$

We define $\mathrm{Sub}(\mathbf{x})$ to be the set of *all* subwords of $\mathbf{x}$, so

$$\mathrm{Sub}(\mathbf{x}) = \bigcup_{k \in \mathbb{N}} \mathrm{Sub}_k(\mathbf{x}).$$

Let $p_{\mathbf{x}}$ be a function which computes the number of subwords of $\mathbf{x}$ of a given length, so

$$p_{\mathbf{x}}(k) = \mathrm{Card}\ \mathrm{Sub}_k(\mathbf{x}).$$

As defined, $p_{\mathbf{x}}$ computes the *subword complexity* of $\mathbf{x}$.

1.2 Homomorphisms

> If there is one central idea which is common to all aspects of modern algebra it is the notion of homomorphism [26].

We now define the operation of *homomorphism*, or just *morphism* for short. A morphism is a map

$$h : \Sigma \to \Delta^*$$

that assigns a string $h(a)$ to each symbol a of a finite alphabet Σ. This map is extended to Σ^* using the rule

$$h(xy) = h(x)h(y).$$

If $\Delta \subseteq \Sigma$ then h is also an *endomorphism*.

If there exists an integer $j \geq 1$ such that $h^j(a) = \epsilon$, then the letter a is said to be *mortal.* The set of mortal letters associated with a homomorphism h is denoted by M_h. The *mortality exponent* of a homomorphism h is defined to be the least integer $t \geq 0$ such that $h^t(a) = \epsilon$ for all $a \in M_h$. (If $M_h = \emptyset$, we take $t = 0$.) We write the mortality exponent as $\exp(h) = t$.

If $h(a) \neq \epsilon$ for all $a \in \Sigma$, then h is *non-erasing.*

An upper bound on $\exp(h)$ is given in the following lemma.

Lemma 1.2.1 *Given any homomorphism h,* $\exp(h) \leq \text{Card } M_h$.

Proof. Define $S_0 = \emptyset$. For $i > 0$, let

$$S_i = \{a \in \Sigma \ : \ h(a) \in S_{i-1}^*\}.$$

We now show that

$$S_i = \{a \in \Sigma \ : \ h^i(a) = \epsilon\}.$$

Clearly this is true for $i = 0$. Assume that for some $k \geq 0$,

$$S_k = \{a \in \Sigma \ : \ h^k(a) = \epsilon\}.$$

Let $a \in S_{k+1}$. By definition, $h(a) \in S_k^*$. By induction, $h^k(h(a)) = \epsilon$. Therefore

$$S_{k+1} \subseteq \{a \in \Sigma \ : \ h^{k+1}(a) = \epsilon\}.$$

Now let $h^{k+1}(a) = \epsilon$, then $h^k(h(a)) = \epsilon$. Thus by induction, $h(a) \in S_k^*$, and $a \in S_{k+1}$. Hence,

$$S_{k+1} \supseteq \{a \in \Sigma \ : \ h^{k+1}(a) = \epsilon\}.$$

Clearly $S_{\exp(h)} = M_h$ and

$$0 = |S_0| < |S_1| < \cdots < |S_{\exp(h)}| = |M_h|.$$

Therefore $\exp(h) \leq \text{Card } M_h$. ■

We let Σ^ω denote the set of all (one-sided) infinite words over the alphabet Σ. Most of the definitions above extend to Σ^ω in the obvious way. For example, if $\mathbf{w} = a_1 a_2 a_3 \cdots$, then $h(\mathbf{w}) = h(a_1)h(a_2)h(a_3)\cdots$. If $L \subseteq \Sigma^+$ is a set of nonempty words, then we define

$$L^\omega = \{w_1 w_2 w_3 \cdots \ : \ w_i \in L \text{ for all } i \geq 1\}.$$

Perhaps slightly less obviously, we can also define the word $h^\omega(a)$ for a letter a, provided $h(a) = wax$ and $w \in M_h^*$. In this case, there exists $t \geq 0$ such that $h^t(w) = \epsilon$. Then we define

$$h^\omega(a) = h^{t-1}(w) \cdots h(w)\, w\, a\, x\, h(x)\, h^2(x) \cdots ,$$

which is infinite iff $x \notin M_h^*$.

Chapter 2

Subword Complexity of k-quasi-context-free sequences

In this chapter we introduce the notion of k-quasi-context-free sequences and give two examples of k-quasi-context-free sequences whose subword complexity is exponential, one of which is maximum.

2.1 Context-Free Grammars

The context-free grammar representation scheme was initiated by Chomsky in [12] in an attempt to find a reasonable mathematical model of natural languages such as English, French, etc. In the period 1958–1960, several papers developing the theory were written [14, 13, 16, 8, 9].

In late 1960, it was discovered that the languages defined by *Backus-Naur Form* [7, 38] were identical with the context-free languages. Backus-Naur form is a notation used by computer scientists to describe programming languages. This caused a flurry of activity in the theoretical development of context-free languages. Much of this work was concerned either with natural languages or with programming languages.

Important later writings by Chomsky on the subject appear in [15]. Applications of context-free grammar theory have been made to compiler design, see [1, 2, 3, 28]. A useful compilation of much of the theory of context free grammars, as of the time it was written, is found in [24]. A comprehensive study of languages from the point of view of grammars is given in [24]. Two excellent textbooks which cover the topic of context-free grammars are [27] and [29].

A context-free grammar G is a quadruple (N, T, P, S), where

N (the set of *non-terminals*) is a set of symbols,

T (the set of *terminals*) is a set of symbols disjoint from N,

P (the set of *productions*) is a finite subset of $N \times (N \cup T)^*$, and

S (the *start symbol*) is an element of N.

For $A \in N$ and $B \in (N \cup T)^*$ a production is written $A \to B$.

Let us define the relations $\Rightarrow$ and $\Rightarrow^*$ between two strings in $(N \cup T)^*$. If $A \to B$ is a production of P and α and γ are any two strings in $(N \cup T)^*$, then $\alpha A\gamma \Rightarrow \alpha B\gamma$. We say that $\alpha A\gamma$ *directly derives* $\alpha B\gamma$. Two strings are related by $\Rightarrow$ exactly when the second is directly derived by the first. Now let $\alpha_1, \alpha_2, \cdots, \alpha_n$ be strings in $(N \cup T)^*$, $n \geq 1$, and

$$\alpha_1 \Rightarrow \alpha_2, \alpha_2 \Rightarrow \alpha_3, \cdots, \alpha_{n-1} \Rightarrow \alpha_n. \tag{2.1}$$

Then we say $\alpha_1 \Rightarrow^* \alpha_n$ or α_1 *derives* α_n.

The *language generated* by a grammar G, denoted $L(G)$, is

$$\{w : w \in T^* \text{ and } S \Rightarrow^* w\}.$$

We call a language L a *context-free language (CFL)* if it is $L(G)$ for some context-free grammar G.

Example 2.1.1

We give a grammar G which generates the arithmetic expressions with operators $+$ and $*$ and operands represented by the symbol **id**.

Let

$$G = (\{E\}, \{+, *, (,), \mathbf{id}\}, P, E)$$

where P consists of the following productions.

$$\begin{aligned} E &\to E + E \\ E &\to E * E \\ E &\to (E) \\ E &\to \mathbf{id} \end{aligned}$$

Now we will show that $\mathbf{id} * (\mathbf{id} + \mathbf{id}) \in L(G)$ by giving a derivation.

$$\begin{aligned} E &\Rightarrow E * E \\ &\Rightarrow \mathbf{id} * E \\ &\Rightarrow \mathbf{id} * (E) \\ &\Rightarrow \mathbf{id} * (E + E) \\ &\Rightarrow \mathbf{id} * (\mathbf{id} + E) \\ &\Rightarrow \mathbf{id} * (\mathbf{id} + \mathbf{id}) \end{aligned}$$

2.2 Regular Grammars

A context-free grammar is *right-linear* if all of the productions are of the form $A \to wB$ or $A \to w$, where A and B are non-terminals and w is a (possibly empty) string of terminals. If all productions are of the form $A \to Bw$ or $A \to w$, we call it *left-linear.*

A right-linear or left-linear grammar is a *regular grammar.* The language generated by a regular grammar is a *regular language.*

2.3 k-Automatic Sequences

We introduce the notion of k-automatic sequences and give a brief overview of the results pertaining to these sequences. In the next chapter we introduce context-free sequences which are a natural extension of automatic sequences.

Let $w \in \{0, 1, \cdots, k\}^*$. Denote by $[w]_n$ the value of w when treated as an integer in base n, $n > k$. If w is the empty word then $[w]_n = 0$. Inversely, for a positive integer i, let $(i)_n$ represent the word w, not beginning with 0, such that $[w]_n = i$. Define $(0)_n = \epsilon$.

Given a sequence $(a_n)_{n\geq 0}$ define

$$G_c(a) = \{(n)_k : a_n = c\}.$$

The sequence $(a_n)_{n\geq 0}$ is defined to be k-automatic if and only if $G_c(a)$ is a regular language for all c.

A good overview of the complexity results pertaining to k-automatic sequences is given in [5]. We give a short summary of these results.

Theorem 2.3.1 *For any k-automatic sequence u, $p_u(n)$ is $O(n)$ [18]. Furthermore, $p_u(n) = \Omega(n)$ if and only if u is not ultimately periodic.*

Theorem 2.3.2 *If k is a prime number, and u is a sequence over the finite field $\mathbb{F}_q$, then u is k-automatic only if the formal series $\Sigma u_n x^n$ is algebraic over the field of rational fractions $\mathbb{F}_q(x)$ [35].*

For related results see [18, 20, 4].

Theorem 2.3.3 *If u is the Thue-Morse sequence then the sequence $(p_u(n+1) - p_u(n))_n$ is 2-automatic [37].*

Furthermore, if u is a k-automatic sequence that obeys a few extra conditions then the sequence $(p_u(n+1) - p_u(n))_n$ is k-automatic [42, 36, 6].

2.4 k-Context-Free Sequences

We introduce the notion of *k-context-free* sequences and *k-quasi-context-free* sequences.

Let Σ be a finite alphabet such that $|\Sigma| \geq 2$. Given a sequence $(a_n)_{n\geq 0}$ over Σ, let

$$H_c(a) = \{(n)_k : a_n = c\}.$$

The sequence $(a_n)_{n\geq 0}$ is *k-context-free* if and only if $H_c(a)$ is a context-free language for all $c \in \Sigma$.

The sequence $(a_n)_{n\geq 0}$ is *k-quasi-context-free* if and only if $H_c(a)$ is not a context-free language for at most one $c \in \Sigma$.

Note that since context-free languages are not closed under complement, a k-quasi-context-free sequence is not necessarily a k-context-free sequence. A k-quasi-context-free sequence a is not k-context-free if and only if, for some c, $H_c(a)$ is context-free and $\overline{H_c(a)}$ is not context-free.

Clearly, a k-context-free sequence is also k-quasi-context-free. Furthermore, a k-automatic sequence is both k-context-free and k-quasi-context-free.

We note that k-context-free and k-quasi-context-free sequences are generalizations of k-automatic sequences that have not been considered previously.

In what follows we give examples of two k-quasi-context-free sequences u for which $p_u(n)$ is exponential. In both examples we show that the *characteristic sequence* of a given language is 2-quasi-context-free.

We say that $\chi_L : \mathbb{N} \to \{0,1\}$ is the *characteristic sequence* of the subset L of $1\{0,1\}^* \cup \{\epsilon\}$ if $w \in L$ if and only if $\chi([w]_2) = 1$.

2.5 A 2-quasi-context-free sequence u for which $p_u(n)$ is $\Omega(2^{\frac{n}{2}-\epsilon})$

Theorem 2.1 states that for a k-automatic sequence u, $p_u(n)$ is $O(n)$. It would be interesting to show a similar result for context-free sequences. Currently we have been unable to show the existence or non-existence of a context-free sequence u such that $p_u(n)$ is exponential. However, we have found examples of quasi-context-free sequences u for which $p_u(n)$ is exponential.

The sequence considered here is the characteristic sequence of a relatively simple context-free language.

For $h \geq 1$, the language L_h is defined as follows.

$$L_h = \{w \in 1\{0,1\}^* : w \text{ ends in a non-empty palindrome } p \text{ such that } |p| \equiv 0 \pmod{2h}\} \tag{2.2}$$

In what follows, all palindromes are assumed to be non-empty.

Lemma 2.5.1 *For all $h \geq 1$, L_h is context-free.*

Proof. To see that L_h is context-free consider the following grammar G_h.

$$
\begin{aligned}
S &\rightarrow S_1 \,|\, S_2 \\
S_1 &\rightarrow 1A \\
S_2 &\rightarrow 1B_2 1 \\
A &\rightarrow 0A \,|\, 1A \,|\, B_1 \\
B_i &\rightarrow 0B_{i+1}0 \,|\, 1B_{i+1}1 \text{ for } 1 \leq i \leq h \\
&\vdots \\
B_{h+1} &\rightarrow B_1 \,|\, \lambda
\end{aligned}
$$

We will give a sketch of a proof that $L(G_h) = L_h$.

It can be seen that $B_1 \Rightarrow^* p \in \{0,1\}^*$ if and only p is a palindrome and $|p| \equiv 0 \pmod{2h}$. Similarly, $S_2 \Rightarrow^* p \in \{0,1\}^*$ if and only if p is a palindrome that begins with 1 and $|p| \equiv 0 \pmod{2h}$.

Let A^F be the set of sentential forms of the form $\{vB_1\}$ where $v \in \{0,1\}^*$. Then $A \Rightarrow^* w \in \{0,1\}^*$ if and only there exists $f \in A^F$ such that $f \Rightarrow^* w$. So A derives words that end in a palindrome p such that $|p| \equiv 0 \pmod{2h}$.

Therefore S derives words that start with 1 and end in a palindrome p such that $|p| \equiv 0 \pmod{2h}$. Therefore $S \Rightarrow^* w \in \{0,1\}^*$ if and only if $w \in L_h$. ■

Denote by $\gamma_h(n)$, $n \geq 1$, the number of occurrences of 1 in

$$\chi_{L_h}(2^{n-1}) \cdots \chi_{L_h}(2^n - 1).$$

In other words, $\gamma_h(n)$ counts the number of words in L_h of length n.

Lemma 2.5.2 *For $n \geq 1$ and $h \geq 1$, $\gamma_h(n) < \frac{2^n}{2^h-1}$.*

Proof. Let $S = \{(2^{n-1})_2, (2^{n-1}+1)_2, \cdots, (2^n-1)_2\}$. The number of words in S that end in palindromes of length $2i$ is 2^{n-i} when $0 \leq i \leq \lfloor \frac{n}{2} \rfloor$. Note that $w \in S$ may end in palindromes of more than one length.

Therefore

$$\begin{aligned} \gamma_h(n) &\leq 2^{n-h} + 2^{n-2h} + \cdots + 2^{n-\lfloor \frac{n}{2h} \rfloor \cdot h} \\ &= \frac{1}{2^h-1}\left(2^n - 2^{n-\lfloor \frac{n}{2h} \rfloor \cdot h}\right) \\ &< \frac{2^n}{2^h-1} \blacksquare \end{aligned}$$

Denote by $\alpha_h(n)$ the number of words w in

$$\{(2^{n-1})_2, (2^{n-1}+1)_2, \cdots, (2^n-1)_2\}$$

such that w ends in 1 and $w \in L_h$. In other words, $\alpha_h(n)$ counts the number of words, $w \in L_h$, of length n that end in 1.

Lemma 2.5.3 *For $n \geq 1$ and $h \geq 1$, $\alpha_h(n) \leq \frac{1}{2}(\gamma_h(n) + 2^{\frac{n}{2}})$.*

Proof. First we define two sets of words in L_h. Let

$$T_h = \{w : |w| = n \text{ and } w = vp, \text{ where } v \in 1\{0,1\}^* \\ \text{and } p \text{ is a palindrome where } |p| \equiv 0 \pmod{2h}\}.$$

$$S_h = \{p : |p| = n, \\ \text{and } p \text{ is a palindrome where } |p| \equiv 0 \pmod{2h}\}.$$

Clearly $ap \in T_h$ if and only if $a\overline{p} \in T_h$. Thus exactly half of the words in T_h end in 1 and the other half end in 0.

Now $|S_h| = 2^{\frac{n}{2}}$ for $n \equiv 0 \pmod{2h}$, and $|S_h| = 0$ otherwise. Note that a given w may be contained in both S_h and T_h. In this context,

$$\gamma_h(n) = |S_h| + |T_h| - |S_h \cap T_h|.$$

Therefore

$$\begin{aligned}
\alpha_h(n) &= |S_h| + \frac{1}{2}|T_h| - |S_h \cap T_h| \\
&= \frac{1}{2}|S_h| + \frac{1}{2}|T_h| - \frac{1}{2}|S_h \cap T_h| + \frac{1}{2}|S_h| - \frac{1}{2}|S_h \cap T_h| \\
&\leq \frac{1}{2}(|S_h| + |T_h| - |S_h \cap T_h| + |S_h|) \\
&\leq \frac{1}{2}(\gamma_h(n) + 2^{\frac{n}{2}}) \blacksquare
\end{aligned}$$

Denote by $\delta_h(n)$ the number of words w in

$$\{(2^{n-1})_2, (2^{n-1}+1)_2, \cdots, (2^n - 1)_2\},$$

such that w ends in 1 and $w \notin L_h$.

Lemma 2.5.4 $\delta_h(n) \geq 2^{n-2-\epsilon}$.

Proof. The total number of words w of length n such that $w \in 1\{0,1\}^*1$ is 2^{n-2}. Therefore

$$\begin{aligned}
\delta_h(n) &= 2^{n-2} - \alpha_h(n) \\
&\geq 2^{n-2} - \frac{1}{2}(\gamma_h(n) + 2^{\frac{n}{2}}) \\
&> 2^{n-2} - \frac{1}{2}(\frac{2^n}{2^h - 1} + 2^{\frac{n}{2}}) \\
&= 2^{n-2} - \frac{2^{n-1}}{2^h - 1} - 2^{\frac{n-2}{2}} \\
&= \frac{(2^h - 3)2^{n-2}}{2^h - 1} - 2^{\frac{n-2}{2}} \\
&= \left(\frac{2^h - 3}{2^h - 1}\right) 2^{n-2} - 2^{\frac{n-2}{2}} \\
&> 2^{n-2-\epsilon} \text{ for any } \epsilon > 0 \text{ and sufficiently large } h \text{ and } n. \blacksquare
\end{aligned}$$

Lemma 2.5.5 *For $h \geq 1$ and $k \equiv 0 \pmod{2h}$, let $R_{h,k} = \{w_1, w_2, \cdots, w_n\}$ be the set of all words such that for all i*

1. *w_i is of length k*
2. *$w_i \notin L_h$, and*
3. *$w_i \in 1\{0,1\}^*1$.*

Let $\{x_1, x_2, \cdots, x_n\}$ be a set of words such that $x_i = w_j^R 0^k w_i$ for some j. Then $x_i \notin L_h$ for all $i \neq j$. Furthermore, $x_j \in L_h$.

Proof. Clearly $x_j \in L_h$ since x_j is a palindrome and $|x_j| \equiv 0 \pmod{2h}$.

Assume, to the contrary, that there exists $i \neq j$ such that $x_i \in L_h$. Then $x_i = vp$ where $v \in 1\{0,1\}^* \bigcup \epsilon$ and p is a palindrome such that $|p| \equiv 0 \pmod{2h}$.

Now p cannot be a suffix of w_i since $w_i \notin L_h$, so $|p| > k$. Also, p ends with a 1 so it must begin with a 1, so $|p| > 2k$.

Let $p = p_0 p_1 \cdots p_{2l-1}$, thus $|p| = 2l$. Then p_{l-1} must be located in $x_i(k) \cdots x_i(2k-1)$, the block of zeroes. But since $x_i(k-1) = 1$ and $x_i(2k) = 1$, and p is a palindrome, p_{l-1} must be located at position $x_i(\frac{3k}{2} - 1)$.

Therefore $|p| = |x_i|$. But this implies that $w_j = w_i$. This gives the desired contradiction. ∎

Define the set $R_{h,k}$ as in Lemma 2.5.5.

We claim that for any mapping f from $R_{h,k}$ to $\{0,1\}$, there exists an index m such that $\chi_{L_h}(m + w_i) = f(w_i)$ for all i.

Consider the following algorithm.

FINDINDEX($R_{h,k} = \{w_1, w_2, \cdots, w_n\}$, f)

{ input: $R_{h,k}$ and f as defined above }

{ output: m as defined above }

set $x_i = w_i$ for $1 \leq i \leq n$

{invariant: x_i always begins and ends with a 1 for $1 \leq i \leq n$ }

{invariant: $x_i \in \{0,1\}^* w_i$ for $1 \leq i \leq n$ }

{invariant: $[x_i]_2 - [w_i]_2 = [x_j]_2 - [w_j]_2$ for $1 \leq i \leq n$ and $1 \leq j \leq n$ }

(1) for $i \leftarrow 1$ to n do

{after this iteration $\chi_{L_h}(x_i) = f(w_i)$}

{during this iteration x_j may change but $\chi_{L_h}(x_j)$ for $j \neq i$ will not change}

(2) if $f(w_i) = 1$ then

$c \leftarrow x_i$

(3) for $j \leftarrow 1$ to n do

(4) $x_j \leftarrow c^R 0^k x_j$

end for

(5) end if

end for

{output the required index into χ_{L_h}}

output $m = [x_1]_2 - [w_1]_2 + 2^{k-1}$

After the lth iteration of loop (1),

$$\chi_{L_h}(x_i) = f(w_i)$$

for $i \leq l$.

Consider the lth iteration of loop (1). Initially, by definition, $\chi_{L_h}(x_i) = 0$ for all i. If $f(w_l) = 0$ then we are done. Otherwise, for $1 \leq i \leq n$, we need to find a new x_i such that for $i \neq l$, $\chi_{L_h}(x_i)$ does not change, and

$$\chi_L(x_l) = f(w_l) = 1.$$

By Lemma 2.5.5, for $j \neq i$, if $x_j \notin L_h$ before execution of statement (4), then $x_j \notin L_h$ after execution of statement (4). If $x_j \in L_h$ before execution of statement (4) then $x_j \in L_h$ after execution of statement (4) since

$$\{1\{0,1\}^* \bigcup \epsilon\}x_j \in L_h.$$

Clearly, $x_i \in L_h$ after the execution of statement (4) since x_i becomes a palindrome and

$$|x_i| \equiv 0 \pmod{2h}.$$

By Lemma 2.5.5, $|R_{h,k}|$ is $2^{n-2-\epsilon}$. Therefore, the minimum number of subwords of χ_{L_h} of length 2^{n-1} is $2^{2^{n-2-\epsilon}}$, and thus

$$p_{\chi_{L_h}}(n) = \Omega(2^{\frac{n}{2}-\epsilon}).$$

Example 2.5.6

Let $h = 3$, $k = 6$, $n = 12$,

$$R_{3,6} = \{100011, 100101, 100111, 101001, 101011, 101111, \\ 110001, 110101, 110111, 111001, 111011, 111101\}$$

and

$$f = \{1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0\}.$$

The rows in the table below give variables' values immediately following each execution of (1). Note that only x_1 is displayed in the table. The value of x_a, for $1 \leq a \leq 12$, after (1) is

$$([x_1]_2 - [w_1]_2 + [w_a]_2)_2.$$

When displaying x_1, spaces have been introduced for clarity.

i	w_i	$f(w_i)$	x_1
1	100011	1	100011
2	100101	0	110001 0^6 100011
3	100111	0	110001 0^6 100011
4	101001	0	110001 0^6 100011
5	101011	1	110001 0^6 100011
6	101111	0	110101 0^6 100011 0^{18} 110001 0^6 100011
7	110001	0	110101 0^6 100011 0^{18} 110001 0^6 100011
8	110101	0	110101 0^6 100011 0^{18} 110001 0^6 100011
9	110111	1	110101 0^6 100011 0^{18} 110001 0^6 100011
10	111001	0	111011 0^6 100011 0^{18} 110001 0^6 101011 0^{36} 110101 0^6 100011 0^{18} 110001 0^6 100011
11	111011	0	111011 0^6 100011 0^{18} 110001 0^6 101011 0^{36} 110101 0^6 100011 0^{18} 110001 0^6 100011
12	111101	0	111011 0^6 100011 0^{18} 110001 0^6 101011 0^{36} 110101 0^6 100011 0^{18} 110001 0^6 100011

After completion of the algorithm,

$$\begin{aligned} m &= [x_1]_2 - [100011]_2 + 2^5 \\ &= [111011\,0^6\,100011\,0^{18}\,110001\,0^6\,101011\,0^{36}\,110101\,0^6\,100011\,0^{18}\,110001\,0^6\,100000]_2. \\ &= 20\,561\,476\,950\,854\,489\,169\,224\,967\,059\,378\,489\,673\,322\,528. \end{aligned}$$

2.6 A 2-quasi-context-free sequence u for which $p_u(n)$ is $\Omega(2^n)$

Here we give an example of a 2-quasi-context-free sequence u where $p_u(n)$ is maximum.

Let h be a homomorphism over $\{0,1\}$ such that

$$\begin{aligned} h(0) &= 00 \\ h(1) &= 01 \end{aligned}$$

Define the language L as follows.

$$\begin{aligned} L = \{ & 10h(w_1)10h(w_2)10\cdots 10h(w_n)11w^R : \text{ for } n \geq 1 \\ & \exists i \text{ such that } w = w_i \text{ and } w_1, w_2, \cdots, w_n \in (0+1)^* \} \end{aligned}$$

Lemma 2.6.1 *L is context-free.*

Proof. To see that L is context-free consider the following grammar G.

$$\begin{aligned} S &\to A10B_1 \\ A &\to 10CA \,|\, \epsilon \\ B_1 &\to 00B_20 \,|\, 01B_21 \,|\, B_2 \\ B_2 &\to B_1 \,|\, B_3 \\ B_3 &\to A11 \\ C &\to 00C \,|\, 01C \,|\, \epsilon \end{aligned}$$

We will sketch a proof that $L(G) = L$. Clearly $C \Rightarrow^* w \in \{0,1\}^*$ implies $w \in (00+01)^*$. Let A^F be the set of sentential forms of the form $\{10C\}^*$. Then $A \Rightarrow^* w \in \{0,1\}^*$ if and only if there exists $f \in A^F$ such that $f \Rightarrow^* w$.

Let B_1^F be the set of sentential forms of the form $h(w)A11w^R$. where $w \in \{0,1\}^*$. Then $B_1 \Rightarrow^* w$ if and only if there exists $f \in B_1^F$ such that $f \Rightarrow^* w$.

Let S^F be the set of sentential forms of the form

$$\{10C\}^* 10h(w)\{10C\}^* 11w^R.$$

where $w \in \{0,1\}^*$. Then $S \Rightarrow^* w$ if and only if there exists $f \in S^F$ such that $f \Rightarrow^* w$. ∎

For $k \geq 1$, let

$$R_k = \{w \,:\, w \in 1\{0+1\}^* \text{ and } |w| = k\}.$$

We claim that for any mapping f from R_k to $\{0,1\}$, there exists an index m such that $\chi_L(m + [w_i]_2) = f(w_i)$ for all i.

Consider the following algorithm.

```
FINDINDEX(k, R_k, f, h)
    { input: R_k = {w_1, w_2, ..., w_p} as defined above }
    { input: a mapping f from R_k to {0,1} }
    { input: the homomorphism h }
    { output: m as defined above }
    x ← ε
    for i ← 1 to p do
      if f(w_i) = 1 then
(1)     x ← x10h((w_i)^R)
      end if
    end for
(2) x ← x110^k
    {output the required index into χ_L}
```

(3) output $m = [x]_2$

Therefore $\mathcal{X}_L$ contains $2^{|R_k|}$ different subwords of length $|R_k|$.

Therefore $p_{\mathcal{X}_L}(n) = \Omega(2^n)$.

Example: Let $k = 3$, $p = 8$ and $R_3 = \{000, 001, 010, 011, 100, 101, 110, 111\}$. See the following table for the definition of f.

The rows in the table below give variables' values immediately following each execution of (1). When displaying x, spaces have been introduced for clarity.

i	w_i	$f(w_i)$	x
1	000	1	10 000000
2	001	1	10 000000 10 000001
3	010	0	10 000000 10 000001
4	011	0	10 000000 10 000001
5	100	1	10 000000 10 000001 10 010000
6	101	0	10 000000 10 000001 10 010000
7	110	0	10 000000 10 000001 10 010000
8	111	1	10 000000 10 000001 10 010000 10 010101

After (2), $x = 10\,000000\,10\,000001\,10\,010000\,10\,010101\,11\,000$

After (3), $m = [x]_2 = 68\,991\,193\,784$.

Chapter 3

A Characterization of Finite Fixed Points

In this chapter we are concerned with characterizing the *finite fixed points* of h, In light of the extensive literature on homomorphisms on free monoids [32, 11, 23], it seems remarkable that this has not yet been done, although there are some results (usually phrased in the language of D0L systems) that are vaguely related to ours. See, for example, [41, 19, 44].

Theorem 3.0.4 shows that given a *homomorphism* h and a word w, $h(w) = w$ if and only if w belongs to a given set. Theorem 3.0.6 gives the best possible upper bound on the shortest nonempty finite fixed point of a homomorphism. Much of this chapter is extracted verbatim from my joint paper [22].

Lemma 3.0.2 *Let $h : \Sigma^* \to \Sigma^*$ be a homomorphism. Let $w \in \Sigma^+$ be a finite nonempty word such that w is a subword of $h(w)$. Then there exists a letter $a \in \Sigma$ occurring in w such that a occurs in $h(a)$.*

Proof. Let $w = c_1 c_2 \cdots c_n$, where $c_i \in \Sigma$ for $1 \leq i \leq n$. For $0 \leq i \leq n$, define $s_w(i) =$

$|h(c_1c_2\cdots c_i)|$. (If the word w is clear, we omit the subscript.) In particular, $s(0) = 0$.

Let $h(w) = d_1d_2\cdots d_{s(n)}$, where $d_i \in \Sigma$ for $1 \leq i \leq s(n)$. Hence

$$h(c_i) = d_{s(i-1)+1}\cdots d_{s(i)}$$

for $1 \leq i \leq n$. Since w is a subword of $h(w)$, we know there must exist an integer t, $0 \leq t \leq s(n) - n$, such that $w = d_{t+1}\cdots d_{t+n}$. Hence $c_i = d_{t+i}$ for $1 \leq i \leq n$.

Consider the least index $j \geq 1$ for which $s(j) \geq t + j$. Such an index must exist, since the inequality holds for $j = n$. There are now two cases to consider.

Case 1: $j = 1$: Then $h(c_1) = d_1d_2\cdots d_{s(1)}$ and $s(1) \geq t + 1$. Hence $h(c_1)$ contains $d_{t+1} = c_1$. Let $a = c_1$.

Case 2: $j > 1$: Then by the definition of j we must have $s(j-1) < t + j - 1$. Hence $s(j-1)+1 < t+j$, and since $h(c_j) = d_{s(j-1)+1}\cdots d_{s(j)}$, we know $h(c_j)$ contains $d_{t+j-1}d_{t+j} = c_{j-1}c_j$ as a subword. Let $a = c_j$. ■

Corollary 3.0.3 *If $w \in \Sigma^+$ is a nonempty finite word with $h(w) = w$, then there exist words $w_1, w_2, w_3, w_4 \in \Sigma^*$ and a letter $a \in \Sigma$ such that $w = w_1w_2aw_3w_4$, $h(w_1w_2) = w_1$, $h(a) = w_2aw_3$, and $h(w_3w_4) = w_4$.*

If $h(w) = w$, then, using the notation in the proof of Lemma 3.0.2, we have $t = 0$ and

$s(n) = n$. Let

$$w_1 = d_1 \cdots d_{s(j-1)};$$
$$w_2 = d_{s(j-1)+1} \cdots d_{j-1};$$
$$a = d_j;$$
$$w_3 = d_{j+1} \cdots d_{s(j)};$$
$$w_4 = d_{s(j)+1} \cdots d_n.$$

It works. ■

Now define

$$A_h = \{a \in \Sigma \ : \ \exists\ x, y \in \Sigma^* \text{ such that } h(a) = xay \text{ and } xy \in M_h^*\}$$

and

$$F_h = \{h^t(a) \ : \ a \in A_h \text{ and } t = \exp(h)\}.$$

Note that there is at most one way to write $h(a)$ in the form xay with $xy \in M_h^*$.

Theorem 3.0.4 *Let $h : \Sigma^* \to \Sigma^*$ be a homomorphism. Then a finite word $w \in \Sigma^*$ has the property that $w = h(w)$ if and only if $w \in F_h^*$.*

($\Longleftarrow$): Suppose $w \in F_h^*$. Then we can write $w = w_1 w_2 \cdots w_r$, where each $w_i \in F_h$, and there exist letters $a_1, a_2, \ldots, a_r \in A_h$ such that $w_i = h^t(a_i)$, with $t = \exp(h)$.

Since $a_i \in A_h$, we know that there exist x_i, y_i with $x_i y_i \in M_h^*$ such that $h(a_i) = x_i a_i y_i$. Since $t = \exp(h)$, we have $h^t(x_i) = h^t(y_i) = \epsilon$. Hence

$$h^{t+1}(a_i) = h^t(x_i)\, h^t(a_i)\, h^t(y_i) = h^t(a_i).$$

Thus $h(w_i) = w_i$ for $1 \leq i \leq r$, and so $h(w) = w$.

($\Longrightarrow$): We prove the result by contradiction. Suppose $h(w) = w$, and assume w is the shortest such word with $w \notin F_h^*$. Clearly $w \neq \epsilon$.

By Corollary 3.0.3 there exist w_1, w_2, w_3, w_4, a such that $w = w_1 w_2 a w_3 w_4$, $h(w_1 w_2) = w_1$, $h(a) = w_2 a w_3$, and $h(w_3 w_4) = w_4$.

Now a is a subword of w, so $h(a)$ is a subword of $h(w) = w$, and hence by an easy induction, it follows that

$$h^i(a) \text{ is a subword of } w \text{ for all } i \geq 0. \tag{3.1}$$

Then we must have $w_2 w_3 \in M_h^*$, since otherwise the length of

$$h^i(a) = h^{i-1}(w_2) \cdots h(w_2)\, w_2\, a\, w_3\, h(w_3) \cdots h^{i-1}(w_3)$$

would grow without bound as $i \to \infty$, contradicting (3.1). It follows that $h^t(w_2 w_3) = \epsilon$, where $t = \exp(h)$.

Now we have $w_1 = h(w_1 w_2)$, so by applying h^t to both sides, we see

$$h^t(w_1) = h^{t+1}(w_1 w_2) = h^{t+1}(w_1)\, h^{t+1}(w_2) = h^{t+1}(w_1).$$

Hence, defining $y_1 = h^t(w_1)$, we have $h(y_1) = y_1$. In a similar fashion, if we set $y_2 = h^t(w_4)$, then $h(y_2) = y_2$. Since $|y_1|, |y_2| < |w|$, it follows by the minimality of w that $y_1, y_2 \in F_h^*$. Now

$$w = h^t(w) = h^t(w_1)\, h^t(w_2)\, h^t(a)\, h^t(w_3)\, h^t(w_4) = y_1\, h^t(a)\, y_2,$$

and hence $w \in F_h^*$, a contradiction. ■

Example 3.0.5

Consider the homomorphism h defined on $\Sigma = \{a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4\}$ as follows:

$$
\begin{aligned}
h(a_1) &= b_1\, b_3\, a_1\, b_4 \\
h(a_2) &= a_2\, b_2\, b_1 \\
h(a_3) &= a_1\, b_4 \\
h(a_4) &= a_4\, a_4 \\
h(b_1) &= b_2\, b_3 \\
h(b_2) &= b_3\, b_4\, b_3 \\
h(b_3) &= b_4\, b_4 \\
h(b_4) &= \epsilon
\end{aligned}
$$

Then $M_h = \{b_1, b_2, b_3, b_4\}$ and $A_h = \{a_1, a_2\}$. Since $h^4(b_1) = h^4(b_2) = h^4(b_3) = h^4(b_4) = \epsilon$ and $h^3(b_1) \neq \epsilon$, clearly $\exp(h) = 4$. Thus

$$
\begin{aligned}
F_h &= \{h^4(a_1), h^4(a_2)\} \\
&= \{b_4\, b_4\, b_4\, b_4\, b_3\, b_4\, b_3\, b_4\, b_4\, b_2\, b_3\, b_4\, b_4\, b_1\, b_3\, a_1\, b_4,\ a_2\, b_2\, b_1\, b_3\, b_4\, b_3\, b_2\, b_3\, b_4\, b_4\, b_4\, b_4\, b_3\, b_4\, b_3\, b_4\, b_4\, b_4\, b_4\}.
\end{aligned}
$$

Note that $h^4(a_1) = h(h^4(a_1))$ and $h^4(a_2) = h(h^4(a_2))$ since $a_1 \in A_h$ and $a_2 \in A_h$. Therefore the elements of F_h are finite fixed points as expected.

We now examine the following question. Suppose h possesses a nonempty finite fixed point w. How long can the shortest w be, as a function of the description of h?

Theorem 3.0.6 *If a homomorphism h possesses a nonempty finite fixed point, then there exists such a fixed point w with $|w| \leq m^{n-1}$, where $n = \text{Card}\ \Sigma$ and $m = \max_{a \in \Sigma} |h(a)|$. Furthermore, this bound is best possible.*

As we have seen in Theorem 3.0.4, a word w is a finite fixed point iff $w \in F_h^*$. Hence, if there exists a nonempty finite fixed point, the shortest such must lie in F_h. But

$$F_h = \{h^t(a) \ : \ a \in A_h \text{ and } t = \exp(h)\}.$$

Since $a \in A_h$, we have $h(a) = xay$ with $xy \in M_h^*$. Hence $a \notin M_h$ and so $\exp(h) \leq \text{Card } M_h \leq n-1$. If $m = \max_{a\in\Sigma} |h(a)|$, then clearly $|h^i(a)| \leq m^i$ for all $i \geq 0$. It follows that $|w| = |h^t(a)| \leq m^{n-1}$.

To see that the bound is best possible, consider the homomorphism h defined on $\Sigma = \{a_1, a_2, \ldots, a_n\}$ as follows:

$$h(a_1) = a_1 a_2^{m-1};$$
$$h(a_i) = a_{i+1}^m \text{ for } 2 \leq i \leq n-1;$$
$$h(a_n) = \epsilon.$$

Then

$$w = a_1 a_2^{m-1} a_3^{m(m-1)} \cdots a_n^{m^{n-2}(m-1)}$$

is a fixed point of h, and

$$|w| = 1 + (m-1) + m(m-1) + \cdots + m^{n-2}(m-1) = m^{n-1}. \blacksquare$$

Example 3.0.7

Consider the homomorphism h defined on $\Sigma = \{a_1, a_2, a_3, a_4, a_5\}$ as follows:

$$h(a_1) = a_1\, a_2$$
$$h(a_2) = a_3\, a_3$$
$$h(a_3) = a_4\, a_4$$
$$h(a_4) = a_5\, a_5$$
$$h(a_5) = \epsilon$$

Using the notation in the above theorem, $m = 2$ and $n = 5$. The following word w should be a finite fixed point such that $|w| \leq 2^4$.

$$w = a_1\, a_2\, a_3^2\, a_4^4\, a_5^8 \quad \text{and}$$
$$h(w) = a_1\, a_2\, a_3^2\, a_4^4\, a_5^8.$$

Chapter 4

A Characterization of Infinite Fixed Points

This chapter is due to Jeff Shallit and has been included because of its relevance to other results in this thesis. It has been extracted verbatim from our joint paper [22].

Infinite fixed points of homomorphisms have received a great deal of attention in the literature. The "usual way" to generate infinite fixed points is to take a homomorphism h and a letter a such that $h(a) = ax$ for some $x \notin M_h^*$. In this case, h is said to be "prolongable" on a [39], and

$$h^{\omega}(a) = a\,x\,h(x)\,h^2(x)\cdots$$

is clearly an infinite fixed point of h. However, this approach does not necessarily generate all the infinite fixed points of h.

As an example, consider the Thue-Morse word [43, 10]

$$\begin{aligned}\mathbf{t} &= t_0 t_1 t_2 \cdots \\ &= 0110100110010110\cdots\end{aligned}$$

where t_i is the sum of the bits in the binary representation of n, taken modulo 2. Then $\mathbf{t}$ is a fixed point of the homomorphism μ which sends $0 \to 01$ and $1 \to 10$; in fact, $\mathbf{t} = \mu^\omega(0)$. The infinite word $\mathbf{t}$ is of interest in part because it is cube-free, that is, it contains no nonempty subword of the form www. Similarly, the homomorphism $2 \to 210$, $1 \to 20$, and $0 \to 1$ has as a fixed point the infinite word

$$210201210120\cdots$$

which is square-free (contains no nonempty subword of the form ww).

Let $\mathbf{w} = c_1c_2c_3\cdots$ be an infinite (one-sided) word over Σ, and let h be a homomorphism. We are interested in characterizing those $\mathbf{w}$ for which $h(\mathbf{w}) = \mathbf{w}$.

Theorem 4.0.8 *The infinite word* $\mathbf{w}$ *is a fixed point of* h *if and only if at least one of the following two conditions holds:*

(a) $\mathbf{w} \in F_h^\omega$*; or*

(b) $\mathbf{w} \in F_h^* h^\omega(a)$ *for some* $a \in \Sigma$*, and there exist* $x \in M_h^*$ *and* $y \notin M_h^*$ *such that* $h(a) = xay$.

Note that there is at most one way to write $h(a) = xay$ with $x \in M_h^*$ and $y \notin M_h^*$.

($\Longleftarrow$): First, suppose condition (a) holds. Then we can write $\mathbf{w} = w_1w_2w_3\cdots$, where each $w_i \in F_h$. Then by Theorem 3.0.4 we have $h(w_i) = w_i$. It follows that $h(\mathbf{w}) = \mathbf{w}$.

Second, suppose condition (b) holds. Then we can write $\mathbf{w} = v\,\mathbf{z}$, where $v \in F_h^*$ and $\mathbf{z} = h^\omega(a)$, where $h(a) = xay$ for some $x \in M_h^*$, $y \notin M_h^*$. Then from Theorem 3.0.4, we have $h(v) = v$.

Since $x \in M_h^*$, we have $h^t(x) = \epsilon$, and hence

$$\mathbf{z} = h^\omega(a) = h^{t-1}(x)\cdots h(x)\,x\,a\,y\,h(y)\,h^2(y)\,h^3(y)\cdots.$$

Since $y \notin M_h^*$, it follows that $|h^i(y)| \geq 1$ for all $i \geq 0$, and hence $\mathbf{z}$ is indeed an infinite word. We then have

$$h(\mathbf{z}) = h^t(x) \cdots h(x)\, x\, a\, y\, h(y)\, h^2(y)\, h^3(y) \cdots = \mathbf{z}$$

and so $h(\mathbf{w}) = h(v\mathbf{z}) = v\mathbf{z} = \mathbf{w}$.

($\Longrightarrow$): Now suppose $\mathbf{w} = c_1 c_2 c_3 \cdots$ is an infinite word, with $c_i \in \Sigma$ for $i \geq 1$, and $h(\mathbf{w}) = \mathbf{w}$. As before, we define $s_{\mathbf{w}}(i) = |h(c_1 c_2 \cdots c_i)|$ for $i \geq 0$. There are several cases to consider.

Case 1: $s_{\mathbf{w}}(i) = i$ for infinitely many integers $i \geq 1$. Suppose $s(i) = i$ for $i = i_0, i_1, i_2, \ldots$. Clearly we may take $i_0 = 0$. Then we can write

$$\mathbf{w} = y_1 y_2 y_3 \cdots$$

where $y_j = c_{i_{j-1}+1} \cdots c_{i_j}$ and $h(y_j) = y_j$ for $j \geq 1$. It follows that $\mathbf{w} \in F_h^\omega$.

Case 2: $s_{\mathbf{w}}(i) = i$ for finitely many $i \geq 1$, and at least one such i. Let $s(i) = i$ for $i = i_0, i_1, \ldots, i_r$, and again take $i_0 = 0$. Then for some integer $r \geq 1$ we can write

$$\mathbf{w} = y_1 y_2 y_3 \cdots y_r\, \mathbf{x}$$

where $y_j = c_{i_{j-1}+1} \cdots c_{i_j}$ and $h(y_j) = y_j$ for $1 \leq j \leq r$, and $h(\mathbf{x}) = \mathbf{x}$. Furthermore, if we write $\mathbf{x} = d_1 d_2 d_3 \cdots$ for $d_i \in \Sigma$, $i \geq 1$, then

$$s_{\mathbf{x}}(i) \neq i \text{ for all } i \geq 1. \tag{4.1}$$

If we can show that (4.1) implies that $\mathbf{x} = h^\omega(a)$, where $h(a) = xay$ for some $x \in M_h^*$, $y \notin M_h^*$, we will be done. This leads to Case 3.

Case 3: $s_{\mathbf{w}}(i) \neq i$ for all $i \geq 1$. Suppose there exist i, j with $1 \leq i < j$ and

$$s(i) > i \text{ but } s(j) < j. \tag{4.2}$$

Among all pairs (i, j) with $1 \leq i < j$ satisfying (4.2), let j_0 be the smallest such j. Next, among all pairs (i, j_0) satisfying (4.2), let i_0 be the largest such i. Suppose there exists an integer k with $i_0 < k < j_0$. If $s(k) < k$, then j_0 is not minimal, while if $s(k) > k$, then i_0 is not maximal. It follows that $j_0 = i_0 + 1$. Then $s(i_0) > i_0$, but $s(i_0 + 1) < i_0 + 1$, a contradiction, since $s(i_0) \leq s(i_0 + 1)$.

It follows that either (a) $s(i) < i$ for all $i \geq 1$, or (b) there exists an integer $r \geq 1$ such that $s(i) < i$ for $1 \leq i < r$ and $s(i) > i$ for all $i \geq r$.

Case 3a: $s_{\mathbf{w}}(i) < i$ for all $i \geq 1$. Since this is true for $i = j_0 := 1$, in particular we see that $h(c_1) = \epsilon$. Now let j_1 be the least index such that

$$h(c_{j_1}) \text{ contains } c_1; \tag{4.3}$$

such an index must exist since $h(\mathbf{w}) = \mathbf{w}$. We then have $h(c_2) = h(c_3) = \cdots = h(c_{j_1 - 1}) = \epsilon$, so the first occurrence of c_{j_1} in $\mathbf{w}$ is at position j_1.

Now inductively assume that we have constructed a strictly increasing sequence $j_0 < j_1 < \cdots < j_t$ such that the first occurrence of c_{j_i} in $\mathbf{w}$ is at position j_i, for $1 \leq i \leq t$.

Let j_{t+1} be the least index such that $h(c_{j_{t+1}})$ contains c_{j_t}. Assume $j_t \geq j_{t+1}$. Since $s(i) < i$ for all i, we have $h(c_{j_{t+1}}) = c_k \cdots c_l$ with $l < j_{t+1} \leq j_t$. Since $h(c_{j_{t+1}})$ contains c_{j_t}, this implies that c_{j_t} occurs to the left of position j_t, a contradiction. Hence $j_t < j_{t+1}$.

Thus we can construct an infinite strictly increasing sequence $j_0 < j_1 < \cdots$ such that the first occurrence of c_{j_i} in $\mathbf{w}$ is at position j_i. It follows that the letters $c_{j_0}, c_{j_1}, \ldots$ in Σ are all distinct. But Σ is finite, a contradiction. Hence this case cannot occur.

Case 3b: There exists an integer $r \geq 1$ such that

$$s_{\mathbf{w}}(i) < i \text{ for } 1 \leq i < r \text{ and } s_{\mathbf{w}}(i) > i \text{ for all } i \geq r. \tag{4.4}$$

Put $a = c_r$. Then $h(a) = xay$ for some $x, y \in \Sigma^*$, and $|y| \geq 1$. Furthermore, the conditions (4.4) on s imply that we can write $\mathbf{w} = u\, a\, \mathbf{v}$ and $h(\mathbf{w}) = h(u)\, x\, a\, y\, h(\mathbf{v})$ such that $u = h(u)x$. An easy induction now gives

$$h^i(\mathbf{w}) = h^i(u)\, h^{i-1}(x) \cdots h(x)\, x\, a\, y\, h(y) \cdots h^{i-1}(y)\, h^i(\mathbf{v}) \tag{4.5}$$

and

$$u = h^i(u)\, h^{i-1}(x) \cdots h(x)\, x \tag{4.6}$$

for all $i \geq 0$. Since $|u| < \infty$, it follows from letting $i \to \infty$ in Eq. (4.6) that there exists an integer $j \geq 0$ such that $h^j(x) = \epsilon$. Hence $x \in M_h^*$, and so $h^t(x) = \epsilon$, where $t = \exp(h)$.

Now $u = h(u)x$, so $h^t(u) = h^{t+1}(u)h^t(x) = h^{t+1}(u)$. Define $u' = h^t(u)$; then $h(u') = u'$. Hence, putting $j = |u'|$, it follows that $s(j) = j$. Hence $j = 0$ and $u' = \epsilon$.

Now, to get a contradiction, suppose that $y \in M_h^*$. Then $h^t(y) = \epsilon$. Define $z = h^t(a)$. Then

$$h(z) = h^{t+1}(a) = h^t(h(a)) = h^t(xay) = h^t(x)\, h^t(a)\, h^t(y) = h^t(a) = z.$$

Hence, putting $j = |z|$, we see that $s(j) = j$, a contradiction since $|z| \geq 1$. Hence $y \notin M_h^*$.

Now, letting $i \to \infty$ in (4.5), we see that $\mathbf{w} = h^\omega(a)$. ∎

Call an infinite fixed point *trivial* if it is in F_h^ω. Our last result shows that, up to application of a coding, all non-trivial infinite fixed points can be generated in the "usual way", i.e., by iterating a homomorphism f on a letter b such that $f(b) = b\,u$ with $u \notin M_f^*$.

Theorem 4.0.9 *Suppose $h : \Sigma^* \to \Sigma^*$ is a homomorphism and $\mathbf{w} \in \Sigma^\omega$ is an infinite word such that $h(\mathbf{w}) = \mathbf{w}$ and $\mathbf{w} \notin F_h^\omega$. Then there exists an alphabet Δ, a non-erasing homomorphism $f : \Delta^* \to \Delta^*$, a coding (i.e., a letter-to-letter homomorphism) $g : \Delta \to \Sigma$, a nonempty word $u \in \Delta^+$ and a letter $b \in \Delta$ such that $f(b) = bu$ and $g(f^\omega(b)) = \mathbf{w}$.*

If $\mathbf{w} \notin F_h^\omega$, then by Theorem 4.0.8, there exists $a \in \Sigma$ such that $\mathbf{w} \in F_h^* h^\omega(a)$, and $h(a) = xay$ with $x \in M_h^*$ and $y \notin M_h^*$. Thus, if $t = \exp(h)$, there exists $v \in F_h^*$ such that

$$\mathbf{w} = v\, h^{t-1}(x) \cdots h(x)\, x\, a\, y\, h(y)\, h^2(y) \cdots .$$

Define $z = vh^{t-1}(x)h^{t-2}(x)\cdots h(x)x$, and let $r = |z|$. If $r = 0$, then $v = x = \epsilon$, and the desired result follows by taking $f = h$ and $g =$ the identity map.

Hence assume $r > 0$ and write $z = b_1 b_2 \cdots b_r$ for $b_i \in \Sigma$, $1 \le i \le r$. Introduce $r + 1$ new symbols $b, a_2, \ldots, a_r, a_{r+1}$, and set $\Delta = \Sigma \cup \{b, a_2, \ldots, a_r, a_{r+1}\}$.

For $d \in \Delta$ define

$$f(d) = \begin{cases} b\, a_2 & \text{if } d = b; \\ a_{i+1}, & \text{if } d = a_i \text{ with } 2 \le i \le r; \\ y, & \text{if } d = a_{r+1}; \\ h(d), & \text{if } d \in \Sigma. \end{cases}$$

Then we have

$$f^\omega(b) = b\, a_2 \cdots a_r\, a_{r+1}\, y\, h(y)\, h^2(y) \cdots .$$

Finally, define the coding $g : \Delta \to \Sigma$ as follows:

$$g(d) = \begin{cases} b_1, & \text{if } d = b; \\ b_i, & \text{if } d = a_i \text{ with } 2 \le i \le r; \\ a, & \text{if } d = a_{r+1}; \\ d, & \text{if } d \in \Sigma. \end{cases}$$

It follows that

$$g(f^\omega(b)) = b_1\, b_2 \cdots b_r\, a\, y\, h(y)\, h^2(y) \cdots = \mathbf{w},$$

as desired.

Note that f is non-erasing iff h is. The following theorem is from Cobham [17].

Theorem 4.0.10 *Let h be a homomorphism such that $h(a) = ax$, and $x \notin M_h^*$. Then there exists an alphabet Δ, a non-erasing homomorphism $f : \Delta^* \to \Delta^*$, a coding (i.e., a letter-to-letter homomorphism) $g : \Delta \to \Sigma$ and a letter $b \in \Delta$ such that $g(f^\omega(b)) = h^\omega(a)$.*

Example 4.0.11

Consider the homomorphism h defined on $\Sigma = \{a, c_1, c_2, c_3, c_4, c_5, c_6\}$ as follows:

$$h(c_1) = c_2\, c_2\, c_3$$

$$h(c_2) = c_3\, c_3$$

$$h(c_3) = \epsilon$$

$$h(c_4) = c_4\, c_1\, c_4$$

$$h(c_5) = c_2\, c_5\, c_3$$

$$h(c_6) = c_1\, c_6$$

$$h(a) = c_2\, c_1\, a\, c_4$$

Using the notation in the above theorem, $M_h = \{c_1, c_2, c_3\}$, $F_h = \{c_5, c_6\}$, $t = 3$, $x = c_2\, c_1$ and $y = c_5$. Let

$$\begin{aligned} v &= h^3(c_5)\, h^3(c_6) \\ &= c_3\, c_3\, c_2\, c_5\, c_3\, c_3\, c_3\, c_3\, c_3\, c_2\, c_2\, c_3\, c_1\, c_6 \end{aligned}$$

and

$$\begin{aligned} \mathbf{w} &= v\, h^{t-1}(x) \cdots h(x)\, x\, a\, y\, h(y)\, h^2(y) \cdots \\ &= c_3\, c_3\, c_2\, c_5\, c_3\, c_3\, c_3\, c_3\, c_3\, c_2\, c_2\, c_3\, c_1\, c_6\, c_3\, c_3\, c_3\, c_3\, c_3\, c_3\, c_2\, c_2\, c_3\, c_2\, c_1\, a\, c_5\, h(c_5)\, h^2(c_5) \cdots \\ &= z\, a\, c_5\, h(c_5)\, h^2(c_5) \cdots . \end{aligned}$$

Thus $r = 25$ and $z = b_1 b_2 \cdots b_{25}$ for $b_i \in \Sigma$, $1 \leq i \leq 25$. Set $\Delta = \Sigma \cup \{b, a_2, a_3, \cdots, a_{25}, a_{26}\}$.

Therefore the homomorphism f is defined as follows:

$$f(c_i) = h(c_i) \text{ with } 1 \leq i \leq 6$$

$$f(a) = c_2\, c_1\, a\, c_4$$

$$f(b) = b\, a_2$$

$$f(a_i) = a_{i+1} \text{ with } 2 \leq i \leq 25$$

$$f(a_{26}) = y$$

Then we have $f^\omega(b) = b\, a_2 \cdots a_{25}\, a_{26}\, y\, h(y)\, h^2(y) \cdots$.

The definition of g is as follows:

$$g(b) = b_1$$

$$g(a_i) = b_i \text{ with } 2 \leq i \leq 25$$

$$g(a_{26}) = a$$

$$g(c_i) = c_i \text{ with } 1 \leq i \leq 6$$

Finally,

$$\begin{aligned} g(f^\omega(b)) &= b_1\, b_2 \cdots b_{25}\, a\, y\, h(y)\, h^2(y) \cdots \\ &= z\, a\, y\, h(y)\, h^2(y) \cdots \\ &= v\, h^{t-1}(x) \cdots h(x)\, x\, a\, y\, h(y)\, h^2(y) \cdots \\ &= \mathbf{w}. \end{aligned}$$

Chapter 5

The periodicity of prefixes (and suffixes) of a PD0L sequence

In this chapter, we improve upon a result in [21, Lemma 1, p. 64] giving a smaller upper bound on the maximum period of the prefixes (and suffixes) of a *PD0L sequence* over Σ. Interestingly, this upper bound is related to the maximum order of a permutation over $|\Sigma|$ elements. Furthermore, we show that this new upper bound is the best possible. The proofs given in this chapter are similar to the proofs given in [21]. They only differ notationally and in the discussion of the improved upper bound.

A *deterministic L-system without interactions* (abbreviated as a D0L-system) is an ordered triple

$$G = (\Sigma, h, w),$$

where Σ is an alphabet, h is an endomorphism defined on Σ^*, and w, referred to as the *axiom*, is an element of Σ^*. The *(word) sequence* $E(G)$ generated by G consists of the words

$$h^0(w) = w, h(w), h^2(w), h^3(w), \ldots .$$

The *language* of G is defined by

$$L(G) = \{h^i(w) \mid i \geq 0\}.$$

Example 5.0.12

Consider the D0L-system

$$G = (\{a, b\}, h, ab)$$

with $h(a) = a$, $h(b) = ab$. Then

$$E(G) = ab, a^2b, \ldots, a^nb, \ldots$$

and so

$$L(G) = \{a^nb \mid n \geq 1\}.$$

A D0L-system (Σ, h, w) is termed *propagating* or, shortly, a *PD0L-system* if h is non-erasing. Thus Example 5.0.12 deals with a PD0L-system.

If $s = w_1, w_2, \cdots$ is a sequence of words over Σ, then

$$U(s) = \bigcup_{i \in \mathbb{N}} \{w_i\}$$

and $U(s)$ is called the *language generated by s*.

If s is an infinite sequence but $U(s)$ is finite, then s is called *singly infinite*. Otherwise s is called *doubly infinite*.

5.1 L-systems

L-systems are arguably one of the most prominent examples of the interdisciplinary nature of formal language theory. The *self-similarity* of many organisms allows their description by simple developmental algorithms. Self-similarity is characterized by Mandelbrot [33, p. 34] as follows:

> When each piece of a shape is geometrically similar to the whole, both the shape and the cascade that generate it are called self-similar.

Biological considerations were strongly connected with the origination of L-systems by Aristid Lindenmayer in [30] and [31]. A corresponding biological phenomenon is described by Herman, Lindenmayer and Rozenberg in [25]:

> In many growth processes of living organisms, especially of plants, regularly repeated appearances of certain multicellular structures are readily noticeable.... In the case of a compound leaf, for instance, some of the lobes (or leaflets), which are parts of a leaf at an advanced stage, have the same shape as the whole leaf has at an earlier stage.

Such a simple formalism for developmental processes clearly has applications to many areas, one of which is computer graphics. In [40] Prusinkiewicz and Lindenmayer write:

> Plant models expressed using L-systems became detailed enough to allow the use of computer graphics for realistic visualization of plant structures and developmental processes.

In this paper we deal only with simplest type of L-system which is the D0L-system.

A D0L-system is a string rewriting system, where each letter of a string can symbolize the presence in that position of a cell of a certain type or state. The whole string represents a filament of cells. Time is assumed to be discrete and, in between two consecutive moments of time, each letter of a string is rewritten as a string which may be empty. This rewriting depends only on the letter concerned, as opposed to other L-systems where the rewriting depends on the m left and n right neighbours of the letter concerned. The resultant string consists of the concatenation of the strings resulting from the rewriting of the individual letters. By repeating this process we obtain a sequence of strings representing the development of the modeled organism.

5.2 Permutations

A *permutation* is simply a bijection on a finite nonempty set. Recall that a bijection is a function $F : X \to Y$ such that for each element $y \in Y$, there is exactly on element $x \in X$ such that $y = f(x)$. More precisely, a permutation σ on S is a bijection $\sigma : S \to S$. Therefore we can think of a permutation simply as a rearrangement of the elements in a finite nonempty set.

Example 5.2.1

If $S = \{a, b, c, d\}$, then one permutation on S is defined by $\sigma(a) = b, \sigma(b) = d, \sigma(c) = a$ and $\sigma(d) = c$. Permutations are usually written in *array notation.* Array notation for σ is as follows.

$$\sigma = \begin{pmatrix} a & b & c & d \\ b & d & a & c \end{pmatrix}$$

where the elements of S are written in the top row and their corresponding images, under σ, are written below.

Composition of permutations expressed in array notation is carried out from right to left by going from top to bottom.

Example 5.2.2

Let

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 4 & 3 & 5 & 1 \end{pmatrix}$$

and

$$\gamma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 4 & 1 & 2 & 3 \end{pmatrix};$$

then

$$\gamma\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 4 & 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 4 & 3 & 5 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 2 & 1 & 3 & 5 \end{pmatrix}.$$

We can see that $(\gamma\sigma)(1) = \gamma(\sigma(1)) = \gamma(2) = 4$. The remainder of the bottom row of $\gamma\sigma$ is obtained in a similar fashion.

There is another notation commonly used to specify permutations. It is called *cycle notation.* Cycle notation has the advantage that it helps readily determine certain important properties of a permutation. Consider

$$\beta = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 5 & 3 & 1 & 6 & 2 & 4 \end{pmatrix}.$$

In cycle notation, β can be written $(2,3,1,5)(6,4)$ or $(4,6)(3,1,5,2)$. An expression of the form $(a_1, a_2, \cdots, a_m)$ is called a *cycle of length m* or an m-cycle. If we think of a cycle as a permutation that fixes any symbol not appearing in the cycle then a permutation can be thought of as a multiplication of cycles. Thus the cycle $(4,6)$ can be thought of as representing the permutation

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 2 & 3 & 6 & 5 & 4 \end{pmatrix}.$$

We now state two important properties of permutations.

1. *Every permutation of a finite set can be written as a product of disjoint cycles.* Let α be a permutation over a finite set $S = \{a_1, a_2, \cdots, a_m\}$. To express α as a product of

disjoint cycles start by choosing any member of A, say a_1, and let

$$\begin{aligned} a_2 &= \alpha(a_1), \\ a_3 &= \alpha(\alpha(a_1)) = \alpha^2(a_1), \\ &\vdots \\ a_1 &= \alpha^m(a_1) \quad \text{for some } m \end{aligned}$$

We know that such an m exists because the sequence $a_1, \alpha(a_1), \alpha^2(a_1), \cdots$ contains a finite number of different elements; so there must eventually be a repetition, say $\alpha^i(a_1) = \alpha^j(a_1)$ for $0 \le i < j$. Then $a_1 = \alpha^m(a_1)$, where $m = j - i$. If we have not exhausted all the elements of A in this process then choose an element $b_1 \in A$ not appearing in the first cycle and create a new cycle as before.

2. *The order of a permutation of a finite set written in disjoint cycle form is the least common multiple of the lengths of the cycles.* The order of a permutation α on $S = \{a_1, a_2, \cdots, a_m\}$ is defined to be the smallest value $n \ge 1$ such that $\alpha^n(a_i) = a_i$ for all $1 \le i \le m$. First, observe that a cycle of length n has order n. Next, suppose that α and β are disjoint cycles of lengths m and n, and let k be the least common multiple of m and n. Now let $\gamma = \alpha\beta$. The order of both α and β must divide the order of γ. The smallest such number is k by definition of least common multiple. Therefore, k is the order of γ. The general case involving more than two cycles can be handled similarly.

Now define $G(n)$ to be the maximum order of a permutation on n elements and define

$$f(a) = \sqrt{n \log(n)} \left(1 + \frac{\log \log n - a}{2 \log n}\right).$$

From [34] we know that

$$f(2) \le \log G(n) \le f(0.975), \quad n \ge 810.$$

Furthermore this can be improved to

$$f(1.16) \leq \log G(n) \leq f(0.975), \quad n \geq 1179568.$$

5.3 The maximum period of the prefixes (and suffixes) of a PDOL sequence is $G(|\Sigma|)$

We improve upon a result in [21, Lemma 1, p. 64] giving a smaller upper bound on the maximum period of the prefixes (and suffixes) of a PD0L sequence. Furthermore we show that this new upper bound is the best possible.

We say that a letter a is *growing* (under a homomorphism h) if

$$\lim_{i \geq 0} |h^i(a)| = +\infty.$$

Lemma 5.3.1 *Let h be a homomorphism over Σ and let $|\Sigma| = p$. If $b \in \Sigma$ is a growing letter under h then $h^{n \cdot p}(b)$ contains $\geq n+1$ non-erasing letters for all $n \geq 0$.*

Proof. If a is a growing letter then clearly $h(a)$ must contain a growing letter. Then for $p > j > i \geq 0$ there must exist $h^i(a)$ and $h^j(a)$ which contain the same growing letter, b. Furthermore, $h^j(a)$ must contain at least two non-erasing letters (at least one of which is b). To see this assume that the only non-erasing letter in $h^j(a)$ is b. Now define a morphism g over Σ so that for $a \in \Sigma$, $g(a) = h^{j-i}(a)$. Then $b \in A_g$ (recall the definition of A_g on page 27). Therefore b derives a finite fixed point over g, and $|g^k(b)| = |h^{k(j-i)}(b)|$ is bounded. So b is not growing. Therefore $h^{n \cdot j}(b)$ contains $n+1$ non-erasing lettters. Since $j < p$, $h^{n \cdot p}(b)$ contains at least $n+1$ non-erasing letters. ■

Corollary 5.3.2 *If $b \in \Sigma$ is a growing letter under h then $|h^{n \cdot p}(b)| \geq n+1$ for every positive integer n.*

Given a doubly infinite PD0L sequence,

$$S = w_0, w_1 = h(w_0), \cdots, w_i = h(w_{i-1}), \cdots,$$

the following lemma shows that for all i, the sequence of ith letters of the words of S,

$$w_0(i), w_1(i), w_2(i), \cdots$$

is ultimately periodic with period $\leq |\Sigma|$.

Lemma 5.3.3 *Let $w_0, w_1, \cdots$ be a doubly infinite PD0L sequence. Let $G = (\Sigma, h, w)$ be a PD0L system such that $E(G) = S$. Let $m = |\Sigma|$. There exist constants $f \leq m$ and $C \leq 2m$ such that for every $k \geq 1$ there exists N_k such that $N_k \leq C \cdot k$ and for every $j \geq N_k$ and $l \geq 0$,*

$$w_j(k) = w_{j+lf}(k).$$

Proof. Let $S = w_0, w_1, \cdots$ be a doubly infinite PD0L sequence. Let $G = (\Sigma, h, w)$ be a PD0L system such that $E(G) = S$. Let $m = |\Sigma|$ and $C = 2m$. We show the existence of a sequence of integers $f_1, f_2, \cdots$ such that $f_i \leq m$ for all $i \geq 1$ and the sequence $w_0(i), w_1(i) \cdots$ has period f_i.

We proceed by induction on k.

For $k = 1$ we need to show that the sequence $w_1(1), w_2(1), \cdots$ is ultimately periodic. Since this sequence is over only m different letters, there must be a repetition with the first $m + 1$ elements. Hence $w_r(1) = w_s(1)$ for some $1 \leq r < s \leq m + 1$. This implies that $h(w_r(1)) = h(w_s(1))$ and, since S is a PD0L sequence, $w_{r+1}(1) = w_{s+1}(1)$. Similarly, for $c \geq 1$, $w_{r+c}(1) = w_{s+c}(1)$. Since $N_1 \leq C \cdot 1$ and $f_1 = s - r \leq m$, the result is proven for $k = 1$.

Now assume that the result is true for $1, \ldots, k$. Let $g = 2pk + p$. From Lemma 5.3.1 we know that $|w_g| \geq k + 1$. There are two cases to consider.

Case 1. Suppose the direct ancestor of $w_g(k+1)$ is the qth letter of w_{g-1} where $q \leq k$. Since $g - 1 = 2pk + p - 1 \geq 2pk = C \cdot k \geq N_k$, then for every r such that $r \geq g - 1$, all $m \geq 1$ and all $q \leq i \leq k$. We have

$$w_r(i) = w_{r+mf_i}(q).$$

Hence by the induction hypothesis

$$\begin{aligned} w_g(k+1) &= w_{g+f_q}(k+1) \\ w_{g+1}(k+1) &= w_{g+f_q+1}(k+1) \\ &\vdots \\ w_{g+f_q-1}(k+1) &= w_{g+2f_q-1}(k+1). \end{aligned}$$

Let $N_{k+1} = g \leq C \cdot (k+1)$. The above equalities together with the induction hypothesis imply that

$$w_j(k+1) = w_{j+lf_{k+1}}(k+1)$$

for every $j \geq N_{k+1}$, $l \geq 1$ and $f_{k+1} = f_q \leq m$.

Case 2. Suppose the direct ancestor of $w_g(k+1)$ is not one of the k letters of w_{g-1}. Then there are two cases.

Case 2a. For some e such that $2pk + p < e \leq 2pk + 2p$, the direct ancestor of $w_e(k+1)$ is the qth letter w_{e-1} where $q \leq k$. We can take $N_{k+1} = e \leq C \cdot (k+1)$ and $f_{k+1} = f_q$. The remainder of the proof of this case is similar to Case 1.

Case 2b. For every e such that $2pk+p \leq e \leq 2pk+2p$, $w_{e-1}(k+1)$ is the direct ancestor of $w_e(k+1)$ (*). Then for some e_1 and e_2 such that $2pk+p \leq e_1 < e_2 \leq 2pk+2p$,

$w_{e_1}(k+1) = w_{e_2}(k+1)$. Now we can take $N_{k+1} = e_1 \le C\cdot(k+1)$. The remainder of the proof of this case is similar to the case for $k = 1$.

Thus the lemma is proved. ■

Lemma 5.3.4 *Let $w_0, w_1, \cdots$ be a doubly infinite PD0L sequence. Let $G = (\Sigma, h, w)$ be a PD0L system such that $E(G) = S$. Let $m = |\Sigma|$. For $i \ge 1$ let f_i be the period of the ultimately periodic sequence $w_0(i), w_1(i), \ldots$.*

$$\mathrm{lcm}(f_1, f_2, \ldots) \le G(m).$$

Proof. Let L_i represent the sequence of letters $w_0(i), w_1(i), \cdots$. From Lemma 5.3.3 we know that $f_i \le m$ for all $i \ge 1$. First note that if $i > j$ then $\mathrm{lcm}(f_i, f_j) = f_j$ unless L_i has property (*) from Lemma 5.3.3. Now consider $i > j$ such that L_i and L_j both have the property (*). For all $k \ge 1$, let $A_k = \{a_0^k, a_1^k, \cdots, a_{f_k-1}^k\}$ be the set of letters contained in the periodic portion of L_k. Assume that there exist u and v such that $a_u^k = a_v^k$. Since $h(a_u^i) = h(a_v^j)$, then $a^i_{u+1 \pmod{f_i}} = a^j_{v+1 \pmod{f_j}}$. By repeating this argument we see that $A_i = A_j$. Hence either A_i and A_j are disjoint or $A_i = A_j$.

Let $I = i_1, i_2 \cdots, i_j$ be a sequence of integers of maximum length such that for all $1 \le k \le j$, L_{i_k} has the property (*) and $A_{i_k} \ne A_{i_l}$ for all $1 \le l < k$. Furthermore there exists no k with $1 \le k < i_j$ and $k \notin I$ such that L_k has the property (*) and $A_k \ne A_{i_l}$ for all $1 \le i_l \le k$. This implies that $A_{i_1}, A_{i_2}, \cdots A_{i_j}$ are all mutually disjoint. We know that such a finite j exists since $i_k > 0$ for all $1 \le k \le j$ and $\sum_{k=1}^{j} i_k \le m$. Therefore $j \le m$.

Now $\mathrm{lcm}(f_{i_1}, f_{i_2}, \cdots, f_{i_j}) = \mathrm{lcm}(f_1, f_2, \cdots)$. To see this let $q = \sum_{k=1}^{j} i_k$. Since $A_{i_1}, A_{i_2}, \cdots A_{i_j}$ are all mutually disjoint, $\mathrm{lcm}(f_{i_1}, f_{i_2}, \cdots, f_{i_j})$ is the order of a permutation which is a product of disjoint cycles whose lengths are $f_{i_1}, f_{i_2}, \cdots, f_{i_j}$. Therefore $\mathrm{lcm}(f_{i_1}, f_{i_2}, \cdots, f_{i_j}) \leq G(q) \leq G(m)$. ∎

The following lemma states that in a PD0L sequence the sequence of prefixes (suffixes) of length k, for any $k \geq 1$, is ultimately periodic with a period f which is independent of k. Let $m = |\Sigma|$. The following lemma is proved by Ehrenfeucht, Lee and Rozenberg [21] with $f \leq m!$. We reiterate the proof of the lemma showing that in fact $f \leq G(m)$.

Lemma 5.3.5 *Let $w_0, w_1, \cdots$ be a doubly infinite PD0L sequence. Let $G = (\Sigma, h, w)$ be a PD0L system such that $E(G) = S$. Let $m = |\Sigma|$. There exist constants $f \leq G(m)$ and $C \leq 2m$ such that for every $k \geq 1$ there exists N_k such that $N_k \leq C \cdot k$ and for every $j \geq N_k$ and every $l \geq 0$,*

$$\mathrm{Pref}_k(w_j) = \mathrm{Pref}_k(w_{j+lf}).$$

and

$$\mathrm{Suff}_k(w_j) = \mathrm{Suff}_k(w_{j+lf}).$$

Proof. We prove the result for prefixes only. The proof for suffixes is similar.

Let $S = w_0, w_1, \cdots$ be a doubly infinite PD0L sequence. Let $G = (\Sigma, h, w)$ be a PD0L system such that $E(G) = S$. Let $m = |\Sigma|$ and $C = 2m$. Let f_i be the period of the ultimately periodic sequence $L_i(w_0), L_i(w_1), \ldots$. Let $f = \mathrm{lcm}(f_1, f_2, \ldots)$. We know from Lemma 5.3.4 that f exists and $f \leq G(m)$. Let N_k be such that $N_k \leq C \cdot k$ and for every $j \geq N_k$ and every $l \geq 0$

$$L_i(w_j) = L_i(w_{j+lf_i}).$$

We have already shown, in Lemma 5.3.3, the existence of a sequence $f_1, f_2, \cdots$ with $1 \leq f_i \leq m$ for all i, such that for some $C \leq 2m$ and every $k \geq 1$ there exists $N_k \leq C \cdot k$ and for every $j \geq N_k$ and every $l \geq 0$.

In Lemma 5.3.4 we have shown that for all i, $\text{lcm}(f_1, f_2, \cdots, f_i) \leq G(m)$.

We proceed by induction on k.

For $k = 1$ the result follows directly from Lemma 5.3.3. Now assume that the result is true for $1, 2, \cdots, k$. From Lemma 5.3.3, we know that for every $j \geq N_{k+1}$ and every $l \geq 0$

$$L_{k+1}(w_j) = L_{k+1}(w_{j+lf_{k+1}}).$$

Since $f_k | f$ this implies

$$L_{k+1}(w_j) = L_{k+1}(w_{j+lf}).$$

Now since $N_{k+1} > N_k$

$$\text{Pref}_{k+1}(w_j) = \text{Pref}_{k+1}(w_{j+lf}). \blacksquare$$

For $m \geq 1$, we show the existence of PD0L sequences P_m in which the period of the sequence of prefixes is $G(m)$.

Let p_k be a permutation on k elements whose order is $G(k)$. Let p_k be a product of $c(k)$ disjoint cycles of lengths $p_k^1, p_k^2, \cdots, p_k^{c(k)}$. Note that $c(k)$ is well defined since all decompositions of p_k into disjoint cycles contain the same number of cycles. Hence $\sum_{i=1}^{c(k)} p_k^i = k$. Define a set of PD0L-systems P_m, for $m \geq 1$, as follows.

$$P_m = (\Sigma, h, a_0^1)$$

where, $a_j^i \in \Sigma$ for $1 \leq i \leq c(m)$, $0 \leq j < p_m^i$ and $(i, j) = (0, 0)$. So $|\Sigma| = m + 1$. Define h as follows.

$$\begin{aligned} h(a_0^0) &= a_0^1 a_0^2 \cdots a_0^{c(m)} a_0^0 \\ h(a_j^i) &= a_{j+1 \pmod{p_m^i}}^i \quad \text{for all } (i, j) \neq (0, 0) \ . \end{aligned}$$

Lemma 5.3.6 *The sequence of prefixes of length $\geq c(m)$ of P_m has minimum period length of $f = G(m)$.*

Proof. First we prove by induction that, for $i > j \geq 1$, the first $n \leq c(m)$ letters of $h^i(a_0^0)$ and $h^j(a_0^0)$ are the same if and only if $i \equiv j \pmod{\operatorname{lcm}(p_m^1, p_m^2, \cdots, p_m^n)}$. Clearly, for $j \geq 0$, $h^j(a_0^1) = a^1_{j \pmod{p_m^1}}$. Therefore the first letter of $h^i(a_0^0)$ and $h^j(a_0^0)$ are the same if and only if $i \equiv j \pmod{p_m^1}$. Now assume that the result holds for $n \leq k < c(m)$. Clearly, for $j \geq 0$, $h^j(a_0^{k+1}) = a^{k+1}_{j \pmod{p_m^{k+1}}}$. Therefore by induction we know that for $i > j \geq 1$ the first $k+1$ letters of $h^i(a_0^0)$ and $h^j(a_0^0)$ are the same if and only if $i \equiv j \pmod{\operatorname{lcm}(p_m^1, p_m^2, \cdots, p_m^k)}$ and $i \equiv j \pmod{p_m^{k+1}}$. Therefore $i \equiv j \pmod{\operatorname{lcm}(p_m^1, p_m^2, \cdots, p_m^k, p_m^{k+1})}$. But $G(m) = \operatorname{lcm}(p_m^1, p_m^2, \cdots, p_m^{c(m)})$. Therefore the result is proven. ■

Example 5.3.7

Let us construct P_5. $G(5) = 6$ so let $p_5 = (21453) = (21)(453)$. Then $c(5) = 2$, $p_5^1 = 2$, and $p_5^2 = 3$. $P_5 = (\{a_0^0, a_0^1, a_1^1, a_0^2, a_1^2, a_2^2\}, h, a_0^0)$ and h is defined as follows.

$$\begin{aligned}
h(a_0^0) &= a_0^1 a_0^2 a_0^0 \\
h(a_0^1) &= a_1^1 \\
h(a_1^1) &= a_0^1 \\
h(a_0^2) &= a_1^2 \\
h(a_1^2) &= a_2^2 \\
h(a_2^2) &= a_0^2
\end{aligned}$$

Thus $E(P_5)$ is the following sequence.

$$
\begin{array}{l}
a_0^0 \\
a_0^1 a_0^2 a_0^0 \\
a_1^1 a_1^2 \cdots \\
a_0^1 a_2^2 \cdots \\
a_1^1 a_0^2 \cdots \\
a_0^1 a_1^2 \cdots \\
a_1^1 a_2^2 \cdots \\
a_0^1 a_0^2 \cdots \\
\vdots
\end{array}
$$

Lemma 5.3.5 implies that the period of the prefixes should be $\leq G(5) = 6$. It is easily seen that the period is exactly 6. ■

Chapter 6

Open Problems

This chapter is dedicated to enumerating the interesting problems which we have been unsuccessful at solving.

As Theorem 2.3.1 states, for a k-automatic sequence u, $p_u(n)$ is $O(n)$. As of yet, we have been unable to show the existence or non-existence of a context-free sequence u such that $p_u(n)$ is exponential.

Open Problem 6.0.8 *Show the existence or non-existence of a context-free sequence u such that $p_u(n)$ is exponential.*

Let $\mathbf{x}$ be an infinite sequence over $\{0,1\}$. The sequence $\mathbf{x}$ has *bounded gaps* if for all $w \in \text{Sub}(\mathbf{x})$ the following conditions are met.

1. The word w occurs infinitely often as a subword in $\mathbf{x}$.
2. Define $n_1^w, n_2^w, \cdots$ such that $\mathbf{x}(n_i^w)$ is the position of first character of the ith occurrence of w in $\mathbf{x}$. For some constant C and all $i \geq 1$, $n_{i+1}^w - n_i^w \leq C|w|$.

The Thue-Morse sequence is known to have bounded gaps. The characteristic sequence of $\{10^*\}$ does not have bounded gaps.

Open Problem 6.0.9 *Show the existence of non-existence of a context-free sequence or quasi-context-free sequence that is not automatic but has bounded gaps.*

While attempting to find context-free sequences with exponential subword complexity, it seems logical to consider writing an algorithm to compute the number of subwords of length $n \leq C$, for some constant C. This would be a good method for finding candidate sequences. However, in order to compute $p_{\mathbf{x}}(n)$ exactly for a given context-free sequence $\mathbf{x}$, it is necessary to find an upper bound on the position $f(\mathbf{x}, n)$ in $\mathbf{x}$ after which no new subwords of length n exist.

Open Problem 6.0.10 *Given a context-free sequence* $\mathbf{x}$ *compute* $f(\mathbf{x}, n)$ *such that* $w \in \mathrm{Sub}(\mathbf{x})$ *is a subword of* $\mathbf{x}(0)\mathbf{x}(1)\cdots\mathbf{x}(f(\mathbf{x}, |w|))$.

The density of a language L is defined to be

$$\lim_{n\to\infty} \frac{|L \bigcap \Sigma^{\leq n}|}{|\Sigma^{\leq n}|}.$$

Intuitively, one would think that subword complexity would tend to be less for languages whose density approaches 0 or 1.

Open Problem 6.0.11 *What is the exact density of the following context-free languages.*

$$L_h = \{w \in 1\{0,1\}^* \ : \ w \text{ ends in a non-empty palindrome } p \text{ such that } |p| \equiv 0 \pmod{2h}\}$$

and

$$\begin{aligned} L \ &= \ \{10h(w_1)10h(w_2)10\cdots 10h(w_n)11w^R : \text{ for } n \geq 1 \\ &\qquad \exists i \text{ such that } w = w_i \text{ and } w_1, w_2, \cdots, w_n \in (0+1)^*\} \end{aligned}$$

Open Problem 6.0.12 *Decide whether or not the following sequence is a 2-CFL-sequence (spaces have been introduced for clarity).*

$$1\ 10\ 11\ 100\ 101\ 110\ 111\ldots$$

Open Problem 6.0.13 *Decide whether or not the following language is context free.*

$$L = \{(n)_2 : n \textit{ is composite } \}.$$

Chapter 7

Acknowledgements

I am pleased to thank Steve Astels for a suggestion which simplified the proof of Theorem 3.0.4, and Jean-Paul Allouche for reading the parts of this manuscript carefully and pointing out several errors.

I am ever grateful for all of the hours that Jeff Shallit has spent reading and re-reading many drafts of this thesis. His commitment is greatly appreciated.

Bibliography

[1] A. V. Aho and J. D. Ullman. *Parsing*, volume 1 of *The Theory of Parsing, Translation and Compiling*. Prentice Hall, Englewood Cliffs, New Jersey, 1972.

[2] A. V. Aho and J. D. Ullman. *Compiling*, volume 2 of *The Theory of Parsing, Translation and Compiling*. Prentice Hall, Englewood Cliffs, New Jersey, 1973.

[3] A. V. Aho and J. D. Ullman. *Principles of Compiler Design*. Addison-Wesley, Reading, Mass., 1977.

[4] J.-P. Allouche. Automates finis en théorie des nombres. *Expo. Math.*, 5:239–266, 1987.

[5] J.-P. Allouche. Sur la complexité des suites infinies. *Bulletin of the Belgian Math. Soc.*, 1:133–143, 1994.

[6] J.-P. Allouche and J. Shallit. The ring of k-regular sequences. *Theor. Comp. Sci.*, 98:163–187, 1992.

[7] J. W. Backus. The syntax and semantics of the proposed international algebraic language of the Zurich ACM-GAMM conference. *Proc. Intl. Conf. on Information Processing*, pages 125–132, 1959.

[8] Y. Bar-Hillel, C. Gaifman, and E. Shamir. On categorial and phrase-structure grammars. *Bull. Res. Council Israel*, 9:155–166, 1960. Section F.

[9] Y. Bar-Hillel, M. Perles, and E. Shamir. On formal properties of simple phrase structure grammars. *Z. Phonetik Sprach. Kommunikationsforsch.*, 14:143–172, 1961.

[10] J. Berstel. Axel Thue's papers on repetitions in words: a translation. Technical Report 20, Publications du Laboratoire de Combinatoire et d'Informatique Mathématique, Université du Québec à Montréal, February 1995.

[11] C. Choffrut and J. Karhumäki. Combinatorics of words. In G. Rozenberg and A. Salomaa, editors, *Handbook of Formal Languages*, volume 1, pages 329–438. Springer-Verlag, 1997.

[12] N. Chomsky. Three models for the description of language. *IRE Trans. on Information Theory*, 2:113–124, 1956.

[13] N. Chomsky. A note on phrase structure grammars. *Information and Control*, 2:393–395, 1959.

[14] N. Chomsky. On certain formal properties of grammars. *Information and Control*, 2:137–167, 1959.

[15] N. Chomsky. Formal properties of grammars. *Handbook of Math. Psych.*, 2:323–418, 1963.

[16] N. Chomsky and G. A. Miller. Finite state languages. *Information and Control*, 1:91–112, 1958.

[17] A. Cobham. On the Hartmanis-Stearns problem for a class of tag machines. In *IEEE Conference Record of 1968 Ninth Annual Symposium on Switching and Automata Theory*, pages 51–60, 1968. Also appeared as IBM Research Technical Report RC-2178, August 23 1968.

[18] A. Cobham. Uniform tag sequences. *Math. Systems Theory*, 6:164–192, 1972.

[19] K. Culik II and A. Salomaa. On infinite words obtained by iterating morphisms. *Theoret. Comput. Sci.*, 19:29–38, 1982.

[20] M. Dekking, M. Mendès France, and A. J. van der Poorten. Folds! *Math. Intell.*, 4:130–138, 1982.

[21] A. Ehrenfeucht, K. P. Lee, and G. Rozenberg. Subword complexities of various classes of deterministic developmental languages without interactions. *Theoretical Computer Science*, 1:59–75, 1975.

[22] D. Hamm and J. Shallit. On fixed points of homomorphisms. submitted, 1998.

[23] T. Harju and J. Karhumäki. Morphisms. In G. Rozenberg and A. Salomaa, editors, *Handbook of Formal Languages*, volume 1, pages 439–510. Springer-Verlag, 1997.

[24] M. A. Harrison. *Introduction to Formal Language Theory.* Addison-Wesley, Reading, Mass., 1978.

[25] G. Herman, A. Lindenmayer, and G. Rozenberg. Description of developmental languages using recurrence systems. *Mathematical Systems Theory*, 8:316–341, 1975.

[26] I. N. Herstein. *Topics in Algebra.* Academic Press, London, 1964.

[27] J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation.* Addison-Wesley, 1979.

[28] P. M. Lewis II, D. J. Rosenkrantz, and R. E. Stearns. *Compiler Design Theory.* Addison-Wesley, Reading, Mass., 1976.

[29] H. R. Lewis and C. Papadimitriou. *Elements of the Theory of Computation.* Prentice Hall, Englewood Cliffs, New Jersey, 1981.

[30] A. Lindenmayer. Mathematical models for cellular interactions in development, part I. *Journal of Theoretical Biology*, 18:280–299, 1968.

[31] A. Lindenmayer. Mathematical models for cellular interactions in development, part II. *Journal of Theoretical Biology*, 18:300–315, 1968.

[32] M. Lothaire. *Combinatorics on Words*, volume 17 of *Encyclopedia of Mathematics and Its Applications.* Addison-Wesley, 1983.

[33] B. B. Mandelbrot. *The fractal geometry of nature.* W. H. Freeman, San Francisco, 1982.

[34] Jean-Pierre Massias, Jean-Louis Nicolas, and Guy Robin. Effective bounds of an element in the symmetric group. *Math. Comp.*, 53:665–678, 1989.

[35] M. Mendès, G. Rauzy, G. Christol, and T. Kamae. Suites algébriques, automates et substitutions. *Bull. Soc. Math. France*, 108:401–419, 1980.

[36] B. Mossé. Notions de reconnaissabilité pour les substitutions et complexité des suites automatiques. *Bull. Soc. Math. France*, 124(2):329–346, 1996.

[37] J. Mouline. *Contribution à l'étude de la complexité des suites substitutives.* PhD thesis, Université de Provence, 1989.

[38] P. Naur. Report on the algorithmic language ALGOL 60. *Comm. ACM*, 3:299–314, 1960. Revised in *Comm. ACM* 6: 1, 1-17.

[39] J.-J. Pansiot. Hiérarchie et fermeture de certaines classes de tag-systèmes. *Acta Informatica*, 20:179–196, 1983.

[40] P. Prusinkiewicz and A. Lindenmayer. *The Algorithmic Beauty of Plants.* Springer-Verlag, New York, 1990.

[41] A. Salomaa. Morphisms on free monoids and language theory. In R. Book, editor, *Formal Language Theory: Perspectives and Open Problems*, pages 141–166. Academic Press, 1980.

[42] T. Tapsoba. *Complexité de suites automatiques.* PhD thesis, Universite Aix-Marseille II, 1987.

[43] A. Thue. Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. *Norske vid. Selsk. Skr. Mat. Nat. Kl.*, 1:1–67, 1912. Reprinted in *Selected Mathematical Papers of Axel Thue*, T. Nagell, editor, Universitetsforlaget, Oslo, 1977, pp. 413–478.

[44] P. M. B. Vitanyi. *Lindenmayer Systems: Structurei, Languages and Growth Functions.* PhD thesis, University of Amsterdam, 1980.