University of California

Los Angeles

Avoidable Formulas in Combinatorics on Words

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Mathematics

by

Ronald James Clark

2001

The dissertation of Ronald James Clark is approved.

D. Stott Parker

Robert Brown

Bruce Rothschild

Kirby Baker, Committee Chair

University of California, Los Angeles

2001

For Azusa, partner-in-crime and partner-in-life

Table of Contents

1 Preliminaries 1

1.1 Introduction 1

1.2 Basic Definitions 2

1.3 Formulas and Divisibility 3

1.4 Avoidability 6

1.5 A Known Algorithm to Determine Avoidability 10

1.6 Finding the Index of Avoidability 12

1.7 Locked Formulas 18

2 Avoidance Bases 20

2.1 Introduction 20

2.2 Avoidance Bases 20

2.3 Finding Avoidance Bases 23

2.4 Avoidance Bases on Small Alphabets 27

2.5 Formulas on Three Letters Are 4-Avoidable 28

3 Minimally Locked Formulas 44

3.1 Introduction 44

3.2 Indices of MLF's on Small Alphabets 45

3.3 Growth Rates 48

4 Index 5 Formulas 53

4.1 Introduction 53

4.2 The Formula ρ Is 4-Unavoidable 54

4.3 5-Special Formulas 54

4.4 5-Special Formulas Are 5-Avoidable 56

4.5 Index 6? 65

A Cubefree Words on Two Letters 66

A.1 Introduction 66

A.2 Fixed Points of Endomorphisms 67

A.3 A Bound for the Density Function 71

A.4 Full Density 73

A.5 Growth Rates 77

References 78

List of Tables

1.1 Cassaigne's Backtracking Algorithm 13

2.1 A 3-Avoidance Basis 29

2.2 Possibilities for X if $XYX \leq \mathbf{G}$ 36

4.1 Known Index 5 Formulas 56

Acknowledgments

Foremost, I thank Kirby Baker, who introduced me to the subject of combinatorics on words. The successful completion of my thesis is largely due to his encouragement and patience.

I thank Bruce Rothschild, who instilled in me a love of combinatorics and afternoon naps. I greatly appreciate our time discussing Ramsey theory, and I thank him for serving on my committee.

I thank Robert Brown and Stott Parker for taking time out of their busy schedules to serve on my committee. Both men have been cordial and professional.

Finally, I thank Roberto Mena, who always had time for me.

Vita

1971	Born, Downey, California
1993	B. S., Mathematics California State University, Long Beach Long Beach, California
1995	M. A., Mathematics University of California, Los Angeles Los Angeles, California
1995-1997	Teaching Assistant Department of Mathematics University of California, Los Angeles Los Angeles, California
1997-present	Instructor Department of Mathematics and Sciences Rio Hondo Community College Whittier, California
1998-1999, 2001	Lecturer Department of Mathematics and Computer Science California State University, Los Angeles Los Angeles, California

Abstract of the Dissertation

Avoidable Formulas in Combinatorics on Words

by

Ronald James Clark

Doctor of Philosophy in Mathematics

University of California, Los Angeles, 2001

Professor Kirby Baker, Chair

The study of regularities in infinite words is a common theme in combinatorics on words. One class of regularities concerns pattern words. A *pattern word*, or *pattern* for short, is a finite word of variables. An infinite word $\mathbf{W}$ *avoids* a pattern P if for any substitution h of the variables of P with nonempty words, $h(P)$ is not a subword of $\mathbf{W}$. A pattern is *n-unavoidable* if no infinite word with n symbols avoids it.

A *formula* is a set of patterns. An infinite word $\mathbf{W}$ *avoids* a formula $f = \{ P_1, P_2, \ldots, P_n \}$ with the variable set Σ if for every substitution h of the variables of Σ, some $h(P_i)$ is not a subword of $\mathbf{W}$. Thus, the formula f is a logical "and" of patterns: any infinite word that fails to avoid f fails to avoid *every* pattern in f. A formula is *n-unavoidable* if no infinite word with n symbols avoids it. A formula is *avoidable* if it is not n-unavoidable for some $n > 0$.

The pattern $P = P_1 x_1 P_2 x_2 \ldots x_{n-1} P_n$, where each x_i is a variable which occurs only once, is related to the formula $f = \{ P_1, P_2, \ldots, P_n \}$. It is known that P is n-unavoidable if and only if f is n-unavoidable, a condition which is easier to verify. The formula f can be viewed as a "fragmentation" of P: if $\mathbf{W}$ does not avoid P for the substitution h, then the subwords $h(P_1)$, $h(P_2)$, $\ldots$, and $h(P_n)$

must occur *in order* in $\mathbf{W}$; if $\mathbf{W}$ does not avoid f for the substitution h, then the subwords $h(P_1)$, $h(P_2)$, $\ldots$, and $h(P_n)$ can occur *arbitrarily* in $\mathbf{W}$.

This thesis answers various questions about avoidable formulas. The index of a formula f is the smallest n for which f is not n-unavoidable. Of primary importance is the search for avoidable formulas having index 5 or higher, none of which have ever been found. Through our use of avoidance bases, we show that any such formula has at least four letters. A certain class of formulas, called locked formulas, have been proven to have index at most 4. We show that any minimally locked formula of over four letters has index equal to 4. We finally show the existence of an index 5 formula with an explicit example.

CHAPTER 1

Preliminaries

1.1 Introduction

In the early 1900's, the Norwegian mathematician Axel Thue [13, 14] showed that there exists an infinite word on a three-letter alphabet which does not contain two consecutive occurrences of the same subword. He also proved the existence of an infinite word over a two-letter alphabet which does not contain three consecutive occurrences of the same subword. Possibly because of the restricted availability of the journal in which his papers were published, Thue's results were mostly ignored. They have been rediscovered several times.

In 1979, Zimin [15] and Bean, Ehrenfeucht, and McNulty [2] independently generalized this idea to arbitrary patterns. They say that a word W avoids a pattern such as $aabba$ if for all possible substitutions A for a and B for b, the word $AABBA$ is not a subword of W. In this light, we can view Thue's first result as the following: there exists an infinite word on three letters avoiding aa, or alternatively, aa is 3-avoidable. Bean et al. and Zimin went on to give an algorithm for determining whether a given pattern can be avoided by an infinite word from some finite alphabet.

Of obvious interest is determining the smallest n for which an avoidable pattern P is n-avoidable. Such n is called the index of P. For example, aa has index 3, while aaa has index 2. In fact, most avoidable patterns have index 3 or smaller. In 1989, Baker, McNulty, and Taylor [1] showed that there exists a pattern having index 4. It is still an open problem whether for all N, there exists $n \geq N$ and a pattern P such that P has index n.

In 1994, Cassaigne [8] introduced formulas, which are finite sets of patterns. To him, a word W avoids a formula f such as $\{\, aba, bab \,\}$ if for all possible substitutions A for a and B for b, either ABA or BAB is not a subword of W. Thus, f is a logical "and" of aba and bab: a word fails to avoid f if and only if it fails to avoid both aba and bab (under the same substitution). Cassaigne showed that every formula can be associated in a natural way to a pattern with the same index. Because of this fact, formulas can be used to find patterns with index higher than 4.

We view this problem as the impetus for this thesis. Namely, is there an index 5 formula? How rare are index 4 formulas? This chapter discusses the basic notions and summarizes some previously discovered results.

1.2 Basic Definitions

An *alphabet* is a finite set. The elements of an alphabet are called *letters*. A *word* is a finite sequence of letters from some alphabet. The *length* of word W is the length of its sequence of letters, denoted $|W|$. If Σ is an alphabet, the set of all words over Σ is denoted Σ^*. Included in Σ^* is the empty word, ε. We use Σ^+ to mean $\Sigma^* \setminus \{\, \varepsilon \,\}$. The *product* of two words U and V, written UV, is the word consisting of the letters of U followed by the letters of V. For example, the product of *abcba* and *bc* is *abcbabc*. Under this operation, we can view Σ^* (resp. Σ^+) as the free monoid (resp. semigroup) generated by Σ. If S is a set of words, then S^* (resp. S^+) is the set of words which are products (resp. nonempty products) of words from S.

If Σ and Δ are alphabets, a map h from Σ^* to Δ^* is called a *homomorphism* if $h(W) = h(a_1)h(a_2)\dots h(a_n)$ for all words $W = a_1a_2\dots a_n$ in Σ^*. A homomorphism is determined by the images $h(a)$, $a \in \Sigma$. A homomorphism is *nonerasing* if $h(a) \neq \varepsilon$ for all $a \in \Sigma$. A homomorphism is *uniform* of length k if $|h(a)| = k$ for all $a \in \Sigma$. An *endomorphism* is a homomorphism from Σ^* to Σ^*. Given $U, V \in \Sigma^*$, U is a *subword* of V if $V = XUY$ for some $X, Y \in \Sigma^*$. If $X = \varepsilon$ (resp. $Y = \varepsilon$), we say that U is a *prefix* (resp. *suffix*) of V. We write $U \leq V$ to indicate that U is a subword of V and $U < V$ to mean $U \leq V$ but $U \neq V$. For example, $ab < babc$ and $\varepsilon \leq V$ for any word V. The relation $\leq$ defines a partial order on Σ^*.

An *infinite word* (over Σ) is an infinite sequence of letters from Σ. The set of all infinite words over Σ is denoted by Σ^ω. If $\mathbf{U} \in \Sigma^\omega$, we say that word P is a *prefix* of $\mathbf{U}$ if $\mathbf{U} = P\mathbf{V}$ for some $\mathbf{V} \in \Sigma^\omega$. Similarly, a word X is a *subword* of $\mathbf{U}$ if $\mathbf{U} = PX\mathbf{V}$ for some $P \in \Sigma^*$ and $\mathbf{V} \in \Sigma^\omega$. Finally, we say that $\mathbf{V} \in \Sigma^\omega$ is a *suffix* of $\mathbf{U}$ if $\mathbf{U} = P\mathbf{V}$ for some $P \in \Sigma^*$. If S is a set of nonempty words (over Σ), then S^ω is the set of infinite products of words from S, that is, the subset of Σ^ω consisting of those infinite words which, for all n, have an element of S^n as a prefix.

1.3 Formulas and Divisibility

Let Δ be an alphabet. A *formula* (over Δ) is a finite set of words over Δ, usually written in arbitrary order separated by dots, although sometimes we revert to set notation for convenience. Some examples are *abab.bba*, *abccba*, and *ab.ba.ac.ca.bc.cb*. The elements of a formula are called *fragments*. Since the "." is suggestive of a multiplicative operation, we sometimes write a formula as $f = \Pi P_i$, where each P_i is a fragment. Given formulas $f = \Pi_{i=1}^{m} P_i$ and $g = \Pi_{j=1}^{n} Q_j$, define $f.g$ to be $f \cup g$, that is, the formula having fragments $P_1, P_2, \ldots P_m, Q_1, Q_2, \ldots Q_n$. A *pattern* is a formula with one fragment. It will be convenient to identify a word P with the pattern $\{ P \}$.

For any formula f, let $\mathrm{alph}(f)$ be the set of letters occurring in f. For example, $\mathrm{alph}(abcba.bcbc) = \{ a, b, c \}$.

Definition 1.3.1. *Let f and g be formulas.*

1. *We say that f* divides *g, written $f \,|\, g$, if there exists a nonerasing homomorphism $\varphi : \mathrm{alph}(f)^* \to \mathrm{alph}(g)^*$ such that for all $P \in f$, there exists $Q \in g$ such that $\varphi(P) \le Q$. We say that f divides g via φ or that φ* shows *the divisibility of f into g.*

2. *We say that f e-*divides *g, written $f \,|_e\, g$, if there exists an injective φ showing $f \,|\, g$ with $|\varphi(a)| = 1$ for all $a \in \mathrm{alph}(f)$. (The e stands for embeds.)*

3. *We say that f i-*divides *g, written $f \,|_i\, g$, if $\mathrm{alph}(f) \subset \mathrm{alph}(g)$ and there exists φ showing $f \,|\, g$ with $\varphi(a) = a$ for all $a \in \mathrm{alph}(f)$. (The i stands for inclusion.)*

For example, $aba.bab \,|\, 011011$ (via $\varphi(a) = 0$ and $\varphi(b) = 11$), $aba.bab \,|_e\, cdcd$ (via $\varphi(a) = c$ and $\varphi(b) = d$), and $aba.bab \,|_i\, abab$. Note that $f \,|_i\, g$ implies $f \,|_e\, g$, which in turn implies $f \,|\, g$.

Two formulas are *equivalent* (resp. e-equivalent, i-equivalent) if they divide (resp. e-divide, i-divide) each other. We may indicate this relation by $\sim$ (resp. $\sim_e$, $\sim_i$). For example, $ab.ba$ and $ab.ba.ac.ca$ are equivalent, while ab and ac are e-equivalent.

If a formula f divides a word W, then we say that W *encounters* f or that W *contains* f.

Definition 1.3.2. *A formula f is* irredundant *if, for all fragments P and Q, $P \leq Q$ implies $P = Q$.*

For a formula f, let $\mathrm{irr}(f) = \{\ P \in f \mid Q \in f \text{ and } P \leq Q \text{ implies } P = Q\ \}$. For example, if $f = aba.ab.baab$, then $\mathrm{irr}(f) = aba.baab$.

Proposition 1.3.3. $\mathrm{irr}(f)$ *is irredundant and i-equivalent to f.*

Proof. Let $P, Q \in \mathrm{irr}(f)$, and suppose $P \leq Q$. Since $\mathrm{irr}(f) \subset f$, we have $Q \in f$. By the definition of $\mathrm{irr}(f)$, we have $P = Q$.

If $P \in f$, then $P \leq Q$ for some $Q \in \mathrm{irr}(f)$, so $f \,|_i\, \mathrm{irr}(f)$. Clearly, $\mathrm{irr}(f) \,|_i\, f$. This means that $\mathrm{irr}(f) \sim_i f$. □

Lemma 1.3.4. *Let f and g be irredundant formulas. Then, f and g are i-equivalent if and only if $f = g$.*

Proof. Suppose $f \,|_i\, g$ and $g \,|_i\, f$. Let $P \in f$. Since $f \,|_i\, g$, there exists $Q \in g$ such that $P \leq Q$. Since $g \,|_i\, f$, there exists $P' \in f$ such that $Q \leq P'$. The relation $\leq$ is transitive, so $P \leq P'$. By irredundancy, we have $P = P'$, which implies $P = Q$. Thus, $f \subset g$. A symmetric argument shows $g \subset f$.

The other direction is obvious. □

For a formula f, let $\mathrm{fact}(f) = \{\ P \in \mathrm{alph}(f)^* \mid P \leq Q \text{ for some } Q \in f\ \}$.

Proposition 1.3.5. *The set of all irredundant formulas over Δ forms a lattice under i-divisibility.*

Proof. Reflexivity and transitivity are clear. Lemma 1.3.4 shows antisymmetry. The sup of two formulas f and g is $\mathrm{irr}(f \cup g)$. The inf of f and g is $\mathrm{irr}(\mathrm{fact}(f) \cap \mathrm{fact}(g))$. □

Let f be a formula and let $P \in f$ have length greater than one. Let P_1 (resp. P_2) be the prefix (resp. suffix) of P of length $|P| - 1$. The formula $\mathrm{Simp}(f, P) = \{ Q \in f \mid Q \neq P \} \cup \{ P_1, P_2 \}$ is the *P-simplification* of f. A formula g is a simplification of f if $g = \mathrm{Simp}(f, P)$ for some $P \in f$. For example, $ab.ba.acb$ is a simplification of $aba.acb$.

Clearly, $\mathrm{Simp}(f, P) \,|_i\, f$, while a little thought shows $f \,|_i\, \mathrm{Simp}(f, P)$ if and only if $P \leq Q$ for some $Q \in f$, $Q \neq P$. In particular, if f is irredundant, then every simplification of f properly i-divides it.

Proposition 1.3.6. *Let f and g be formulas. Suppose g has no fragments of length 1. Then, $f \,|_i\, g$ if and only if $f \sim_i g$ or f i-divides a simplification of g.*

Proof. Suppose $f \,|_i\, g$. If, for all $Q \in g$, there exists $P \in f$ such that $Q \leq P$, then $g \,|_i\, f$, and so $f \sim_i g$. If this fails for some Q, then Q is not a subword of any $P \in f$. In particular, $Q \notin f$, which implies $f \,|_i\, \mathrm{Simp}(g, Q)$.

The reverse implication is obvious. □

1.4 Avoidability

Avoidability measures the complexity of a formula. Generally speaking, formulas built from small alphabets, having many fragments, or having long fragments are easier to avoid, while formulas from large alphabets, having few fragments, or having only short fragments are harder to avoid.

A word W *avoids* formula f if f does not divide W. An infinite word avoids

f if every prefix of it avoids f. For example, $abcab$ avoids aa but not $ab.ba$ ($\varphi(a) = ab$ and $\varphi(b) = c$ gives an encounter).

Definition 1.4.1. *Let Σ be an alphabet. A formula f is Σ-avoidable if, for any $M > 0$, there exists $W \in \Sigma^*$ with $|W| > M$ such that W avoids f.*

The following proposition is proven in Cassaigne [8]. A variation appears in Bean et al. [2].

Proposition 1.4.2. *Let f be a formula. The following are equivalent:*

1. *f is Σ-avoidable.*

2. *The set $\{ W \in \Sigma^* \mid W$ avoids $f \}$ is infinite.*

3. *There exists an infinite word $\mathbf{U}$ over Σ avoiding f.*

Proof. For 1 implies 2: let f be Σ-avoidable. We set $M_1 = 1$. There exists $W_1 \in \Sigma^*$ avoiding f with $|W_1| > M_1$. Having defined $M_1, M_2, \dots, M_k$ and $W_1, W_2, \dots, W_k$, we set $M_{k+1} = |W_k|$ and choose $W_{k+1} \in \Sigma^*$ avoiding f with $|W_{k+1}| > M_{k+1}$. By construction, $\{ W_1, W_2, \dots \}$ is an infinite subset of $\{ W \in \Sigma^* \mid W$ avoids $f \}$.

For 2 implies 3: we show inductively that there exists a sequence of words $\{ W_i \}$ such that (1) each W_i is a prefix of infinitely many words in $\{ W \in \Sigma^* \mid W$ avoids $f \}$, (2) each W_i has length i, and (3) each W_i is a prefix of W_{i+1}. Let $\Sigma = \{ a_1, a_2, \dots, a_n \}$. We may take $W_1 = a_1$. Having defined W_k, we claim that W_{k+1} can be chosen from among $W_k a_1, W_k a_2, \dots, W_k a_n$. Indeed, properties (2) and (3) clearly hold no matter which choice is made. Observing that

$$\{ W \in \Sigma^* \mid W \text{ avoids } f \text{ and } W_k \text{ is a prefix of } W \}$$

is the disjoint union

$$\{\, W_k \,\} \cup \bigcup_{1 \le i \le n} \{\, W \in \Sigma^* \mid W \text{ avoids } f \text{ and } W_k a_i \text{ is a prefix of } W \,\},$$

we see that one of the sets on the right must be infinite.

For 3 implies 1: choose $M > 0$. The prefix of $\mathbf{U}$ of length $M + 1$ works. □

If Σ and Σ' are two alphabets of the same cardinality, then clearly f is Σ-avoidable if and only if f is Σ'-avoidable. Because of this, we say that f is *n-avoidable* if f is Σ-avoidable for *any* alphabet Σ of size n.

If f is not n-avoidable, we say f is *n-unavoidable.* We say that f is *avoidable* if f is n-avoidable for some n. Otherwise, if f is n-unavoidable for all n, then f is *unavoidable.* If f is n-avoidable, then f is $(n + 1)$-avoidable. Taking the contrapositive, if f is $(n+1)$-unavoidable, then f is n-unavoidable. We define the *index* (of avoidability) of f, written $\mathrm{ind}(f)$, to be the smallest n such that f is n-avoidable if f is avoidable, or ∞ if f is unavoidable. If $f \mid g$, then $\mathrm{ind}(f) \ge \mathrm{ind}(g)$. In particular, if two formulas are equivalent, they have the same index.

Every formula is 1-unavoidable. Indeed, let f be a formula, and let $k = \max_{P \in f} |P|$. Every word of length k or larger in $\{\, a \,\}^*$ contains an encounter of f via $\varphi(x) = a$ for all $x \in \mathrm{alph}(f)$.

Many formulas have index 2. It is known that *aaa* and *ababa* do [Thue [14]]. Some formulas have index 3: the classic example is *aa* [Thue [13]]. Few known formulas have index 4. The first example discovered was *abxbayaczcawbc* by Baker, McNulty, and Taylor [1] in 1989.

Until now, no known formulas have been discovered having finite index higher than 4. We prove their existence in Chapter 4.

Formulas having infinite index—that is, unavoidable formulas—are well understood [Zimin [15] and Bean et al. [2]]. An easy example is *aba*: for any n, a word of length $2n+1$ over an alphabet of size n has two nonconsecutive occurrences of the same letter. Sending a to that letter and b to the subword separating them yields an encounter with aba. An algorithm that determines which formulas have infinite index will be discussed in the next section.

A simple corollary of Proposition 1.4.2 is the following.

Corollary 1.4.3. *Let f be a formula, and let $\tilde{f}$ be the formula whose fragments are the reversals of those of f. Then,* $\operatorname{ind}(f) = \operatorname{ind}(\tilde{f})$.

Proof. Suppose f is n-avoidable. Let Σ be an alphabet of size n. By the proposition, the set S of words over Σ^* avoiding f is infinite. Let $\tilde{S}$ be the set of words of S reversed. The words of $\tilde{S}$ avoid $\tilde{f}$, so $\tilde{f}$ is n-avoidable. The result follows by reversing this argument. □

Formulas were first introduced in Cassaigne's thesis [8] based on an idea of Kirby Baker. The connection with the historical notion of a pattern is based on the following theorem. An *isolated letter* of a formula is a letter that appears only once.

Theorem 1.4.4 (Cassaigne [8]). *Let P be a pattern. Let f be the formula obtained by replacing all isolated letters by "." (That is, let f be the set of subwords of P between isolated letters.) Then,* $\operatorname{ind}(P) = \operatorname{ind}(f)$.

Proof. If P is k-unavoidable, then every sufficiently long word on k letters is divisible by P. Since f divides P, by transitivity we have that f divides every sufficiently long word on k letters. This means that f is k-unavoidable. Hence, $\operatorname{ind}(f) \geq \operatorname{ind}(P)$.

Now, suppose $f = \Pi Q_j$ is k-unavoidable. Let $n = |\operatorname{alph}(f)|$. Every sufficiently long word, say of length M, on k letters is divisible by f. Let $\mathbf{W}$ be an infinite word on k letters. We write $\mathbf{W} = X_1Y_1X_2Y_2\dots$, where $|X_i| = M$ and $|Y_i| = |P|$ for all i. For every i, there exists a homomorphism φ_i which shows f dividing X_i. The number of such homomorphisms is bounded by k^{Mn}. By the pigeonhole principle, there must be an infinite sequence $\varphi_{i_1}, \varphi_{i_2}, \dots$ of identical homomorphisms, which we denote simply by φ. This means that $\varphi(Q_1)$ is a subword of X_{i_1}, $\varphi(Q_2)$ is a subword of X_{i_2}, and so on. If we send the isolated letters of P to the spaces between these subwords, we can extend φ to show an encounter of P in $\mathbf{W}$. Thus, $\text{ind}(P) \geq \text{ind}(f)$. □

As a simple example, we show that since *aba.bab* is 2-unavoidable, so is *abacbab*. It is not hard to verify that every word having length 9 over $\{0,1\}$ encounters *aba.bab*. The set of all such encounters can be described as pairs $(\varphi(a), \varphi(b))$: $(01, 1), (0, 0), (0, 11)$, etc. Only finitely many pairs exist, say N. Let W be a word of length $16N + 9$, and write $W = U_1V_1U_2V_2 \dots V_NU_{N+1}$, with $|U_i| = 9$ and $|V_j| = 7$ for all i, j. There are two U_i's that f divides via the same homomorphism φ, say U_j and U_k, $j < k$. In particular, U_j contains $\varphi(aba)$ and U_k contains $\varphi(bab)$. Letting $\varphi(c)$ be the subword of W between these two encounters yields an encounter of *abacbab*.

1.5 A Known Algorithm to Determine Avoidability

As stated in the introduction, Zimin [15] and Bean, Ehrenfeucht, and McNulty [2] proved a result that effectively determines whether a formula is avoidable. In this section, we summarize their findings. We begin with some definitions.

Let f be a formula. For every $a \in \text{alph}(f)$, we make two copies a_L and a_R, and let $V = \{\, a_L \mid a \in \text{alph}(f) \,\} \cup \{\, a_R \mid a \in \text{alph}(f) \,\}$. The *adjacency graph* of f, denoted $\text{AG}(f)$, is the bipartite graph with V as the vertex set such that $\{\, a_L, b_R \,\}$

is an edge if and only if ab is a *transition* of f, that is, ab is a subword of some fragment of f. We may indicate an adjacency graph by listing the associated transitions. For example, $\{ab, ac, bc\}$ refers to the adjacency graph with the edge set $\{\{a_L, b_R\}, \{a_L, c_R\}, \{b_L, c_R\}\}$.

A set $S \subset \mathrm{alph}(f)$ is *free* in formula f if for all $a, b \in S$, a_L and b_R lie in different components of $\mathrm{AG}(f)$. We say that formula f reduces to f' if f' is the formula obtained by deleting all the variables of a free set from f, discarding any empty word fragments. We denote this deletion by $\sigma_S(f) = f'$ (or $\sigma_x(f)$ if $S = \{x\}$). A formula is *reducible* if there is a sequence $f = f_1, f_2, \ldots, f_n = \emptyset$, where for $i < n$, f_i reduces to f_{i+1}.

Theorem 1.5.1 (Zimin [15] or Bean et al. [2]). *A formula is unavoidable if and only if it is reducible.*

Hence, a formula f is avoidable if and only if the deletion of every free set from f is also avoidable. For example, $f = abca.bcab.cac$ is avoidable. The free sets of f are $\{a\}$ and $\{c\}$. Deleting $\{a\}$, we get *bc.bcb.cc*, which is equivalent to *bcb.cc*. This formula has no free sets and so by Proposition 1.5.1 it must be avoidable. Deleting $\{c\}$ from *abca.bcab.cac* yields *aba.bab.a*, which is equivalent to *aba.bab*. This formula has two free sets: $\{a\}$ and $\{b\}$. Deleting $\{a\}$ gives $b.bb \sim bb$, which has no free sets and so must be avoidable. Deleting $\{b\}$ is analogous.

An example of an unavoidable formula is *abca.cabc.cbc*. Deleting the free set $\{c\}$ gives us $aba.ab.b \sim aba$. The formula *aba* has $\{a\}$ as a free set, so deleting it yields b. Finally, deleting free set $\{b\}$ from b produces $\emptyset$.

Theorem 1.5.2 (see [15] or [2]). *Every unavoidable pattern has an isolated letter.*

Proof. Let P be a pattern. We prove the contrapositive: every pattern with no isolated letters is avoidable. The proof is by induction on $n = |\operatorname{alph}(P)|$. If $n = 1$, then $aa \mid P$, and so P is avoidable. We assume the result is true for $n = k$ and that P is a pattern on $k + 1$ letters. If P has no free sets, we are done. Otherwise, let S be any free set of P. We note that $S \neq \operatorname{alph}(P)$ since $|P| > 1$. (If ab is a subword of P, then a and b cannot be in the same free set of P.) Deleting S, we get $\sigma_S(P)$, which by induction is avoidable. □

Theorem 1.5.3 (see [15] or [2]). *Every pattern on n letters of length 2^n or more is avoidable.*

Proof. The proof is by induction on n. If $n = 1$, the result is clear. We assume the theorem is true for $n = k$. Let P be a pattern on $k+1$ letters of length 2^{k+1}. If P has no isolated letters, then P is avoidable by Theorem 1.5.2. Otherwise, P must have an isolated letter, say a. We can write $P = RaS$, with R and S patterns on k letters. At least one of them must have length at least 2^k, say R. By induction, R is avoidable. Since $R \mid P$, P is avoidable. □

Corollary 1.5.4 (Cassaigne [8]). *If f is an unavoidable formula, then every fragment must have an isolated letter (for that fragment). If any fragment of a formula f has only n letters but has length 2^n or larger, then f is avoidable.*

1.6 Finding the Index of Avoidability

The preceding section gave a test for determining whether a formula is avoidable, but it offers no help finding the actual index of avoidability. In general, given an avoidable formula f, there are two things to prove to show $\operatorname{ind}(f) = n$:

1. $\operatorname{ind}(f) < n$: every sufficiently long word on $n - 1$ letters contains f.

```
Input: a formula $f$
Output: if $f$ is $n$-unavoidable, the words over $\{ a_1, \dots, a_n \}$ which avoid $f$.

  $W \leftarrow \varepsilon$.
  Repeat:
    While $f$ divides $W$
      While the last letter of $W$ is $a_n$
        Erase the last letter of $W$.
      If $W = \varepsilon$, terminate.
      Replace the last letter of $W$, say $a_i$, by $a_{i+1}$.
    Display $W$.
    $W \leftarrow W a_1$.
```

Table 1.1: Cassaigne's Backtracking Algorithm

2. $\text{ind}(f) \geq n$: there exist arbitrarily long words on n letters avoiding f.

Proving (1) can be difficult by hand, and often we resort to enumerating all words avoiding f. The backtracking algorithm of Cassaigne [8] described in Table 1.1 is a common way to do this. The algorithm will fail to end if f is n-avoidable.

Using backtracking, one can show, for example, that the only words over $\{ 0, 1 \}$ avoiding *aba.bab* are 0, 00, 001, 0010, 00100, 0011, 00110, 001100, 0011001, 00110010, 001101, 01, 010, 0100, 01001, 010011, 0100110, 01001100, 011, 0110, 01100, 011001, 0110010, 0110011, and 01101; and their inverses (replace 0 by 1 and 1 by 0). Hence, *aba.bab* is 2-unavoidable.

If a formula f is n-avoidable, then the set of infinite words avoiding f on an alphabet of size n is nonempty. Therefore, if we are interested in demonstrating that f is n-unavoidable but not in the actual number of words avoiding f on n letters, the backtracking can be improved by insisting that no suffix of any word be (after possibly relabeling) lexicographically smaller than the corresponding prefix of the same length. For example, 011 cannot be the prefix of some infinite word avoiding *aba.bab* unless 00 is.

Proving (2) generally involves the construction of an HD0L-system. (HD0L refers to a "homomorphic image of a deterministic and 0-context Lindenmayer-system". The "0" is in fact a zero.) An HD0L-system is a 5-tuple $(\Sigma, g, W, \Sigma', h)$, where W is a word over Σ, g is an endomorphism from Σ^* to Σ^*, and h is a homomorphism from Σ^* to $\Sigma^{'*}$. One then shows that $h(g^n(W))$ avoids a formula f for all $n > 0$. A special case of an HD0L-system is a D0L-system, where $\Sigma = \Sigma'$ and h is the identity map. Equivalently, we may identify a D0L-system by the triple (Σ, g, W).

We may write the images of g separated by "/" as an abbreviation for a D0L-system, with W implicitly understood to be 0. For example 012/02/1 (or $g = 012/02/1$ to be more specific) refers to the D0L-system $(\{0, 1, 2\}, g, 0)$, where $g(0) = 012, g(1) = 02,$ and $g(2) = 1$.

The words $g(x)$, $x \in \Sigma$, are called *building blocks* of g. Any word of the form $g(U)$, for some $U \in \Sigma^+$, is called simply a *block*.

If g is a D0L-system such that, for some $x \in \Sigma$, $g(x)$ begins with x, then g is said to be *prefix-preserving* with respect to x. The reason for the name is that one can show $g^n(x)$ is a prefix of $g^{n+1}(x)$ for all $n > 0$. If g is prefix-preserving with respect to x and $|g^n(x)| \to \infty$, define $g^\omega(x)$ to be the unique infinite word having $g^n(x)$ as a prefix for all $n > 0$.

A finite set of words $L \in \Sigma^*$ is a *prefix code* (resp. *suffix code*) if no word in L is a prefix (resp. suffix) of another. A set of words is a *biprefix code* if it is both a prefix and suffix code. A D0L-system g is a *prefix* (resp. *suffix*, *biprefix*) if the set $\{\, g(x) \mid x \in \Sigma \,\}$ is a prefix (resp. suffix, biprefix) code. If g is a prefix and $g(Y)$ is a prefix of $g(Z)$, with $Y, Z \in \Sigma^*$, then Y must be a prefix of Z. In particular, if $X = g(Z)$, then Z is the unique word producing X under g. An analogous statement holds if g is a suffix.

If L is a prefix code, R is an element of L^+, and W is a subword of R, we define a *cut* of W to be a pair (X, Y) such that (1) $W = XY$ and (2) for any words P, Q such that $R = PWQ$, we have $PX \in L^+$. For example, let $L = \{\, 01, 2, 031, 3 \,\}$ and $R \in L^+$. Suppose that $W = 301$ is a subword of R. By examination, we see that the letter 0 only occurs at the beginning of words from L. Therefore, $(3, 01)$ is a cut of W. The other cuts of W are $(\varepsilon, 301)$ and $(301, \varepsilon)$. We use the symbol $|$ to show a cut, as in $W = 3|01$ for the cut $(3, 01)$.

If L is a biprefix code, often a series of deductions allows us to determine where cuts go. For example, let $L = \{\, 01, 2, 031, 3 \,\}$ and suppose $120A$ and $0A30$ are both subwords of R for some word A. We have $120A = 1|2|0A$ and $0A30 = |0A3|0$, since 0 only begins the words of L and 1 only ends the words of L. The letter 3 occurs two different ways, either as the word 3 or as the middle letter of 031. In $|0A3|0$, the second possibility is excluded, so we have $|0A3|0 = |0A|3|0$. Since L is a biprefix code, the first word, $1|2|0A$, must be $1|2|0A|$.

If g is a prefix, n is a positive integer, and W is a subword of $g^n(x)$, we define cuts of W as we did for prefix codes, with $L = \{\, g(x) \mid x \in \Sigma \,\}$ and $R = g^n(x)$.

To show that a D0L-system g avoids a formula f, we argue by contradiction. We assume that $f \mid g^m(0)$, for m minimal, with φ showing the divisibility. We then modify φ, creating a new homomorphism φ' which shows $f \mid g^m(0)$ with the following property: if $P = x_1 x_2 \ldots x_n$ is a fragment of f, then $\varphi'(P) = \varphi'(x_1)\varphi'(x_2)\ldots\varphi'(x_n) = |\varphi'(x_1)|\varphi'(x_2)|\ldots|\varphi'(x_n)|$. That is, each $\varphi'(x)$ is a block of g and, in the context of $\varphi'(P)$, bordered by cuts. This gives a contradiction since $f \,|\, g^{m-1}(0)$ via $\varphi''(x) = g^{-1}(\varphi'(x))$, $x \in \mathrm{alph}(f)$. Proposition 1.6.3 below describes how φ' is typically constructed. We need the following definitions.

Definition 1.6.1. *A* slide setup *is a 4-tuple* (L, R, f, φ)*, where* L *is a finite set of words,* R *is an element of* L^+*, and* f *is a formula which divides* R *via* φ*. If, in addition,* L *is a prefix (resp. suffix, biprefix) code,* (L, R, f, φ) *is called a prefix (resp. suffix, biprefix) slide setup.*

Definition 1.6.2.

1. *A prefix slide setup* (L, R, f, φ) *satisfies the* prefix slide condition *if for every* $a \in \mathrm{alph}(f)$ *there exists a word* P'_a *which is a proper prefix of some element of* L *such that for every fragment* BaC *of* f *containing* a *and every way of writing* $R = X\varphi(B)\varphi(a)\varphi(C)Y$*, there exists* $Z \in L^*$ *such that* $X\varphi(B) = ZP'_a$*.*

2. *A suffix slide setup* (L, R, f, φ) *satisfies the* suffix slide condition *if for every* $a \in \mathrm{alph}(f)$ *there exists a word* S'_a *which is a proper suffix of some element of* L *such that for every fragment* BaC *of* f *containing* a *and every way of writing* $R = X\varphi(B)\varphi(a)\varphi(C)Y$*, there exists* $Z \in L^*$ *such that* $\varphi(C)Y = S'_aZ$*.*

Proposition 1.6.3. *Suppose the prefix slide setup* (L, R, f, φ) *satisfies the prefix slide condition. Then, there exists a homomorphism* φ' *such that if* Q *is a fragment of* f*, then* $\varphi'(Q)$ *is a subword of* R*. Moreover, if* $Q = a_1a_2 \dots a_n$*, then*

$$\varphi'(Q) = \varphi'(a_1)\varphi'(a_2)\dots\varphi'(a_n) = |\varphi'(a_1)|\varphi'(a_2)| \dots |\varphi'(a_n)|$$

Thus, if f' *represents the formula obtained by deleting those* $a \in \mathrm{alph}(f)$ *such that* $\varphi'(a) = \varepsilon$*, then* $f' \,|\, R$ *via* φ'*.*

Proof. Let $a \in \mathrm{alph}(f)$, P'_a be the word produced by the prefix slide condition, and BaC be a fragment of f containing a. Suppose $R = X\varphi(B)\varphi(a)\varphi(C)Y$. Since (L, R, f, φ) satisfies the prefix slide condition, there exists $Z \in L^*$ such that $ZP'_a = X\varphi(B)$. Substituting, we have $R = ZP'_a\varphi(a)\varphi(C)Y$. Since L is a prefix code, there exists a unique element W_a of L^* and a unique proper prefix U_a of some element of L such that $ZP'_a\varphi(a) = ZW_aU_a$. We define $\varphi'(a)$ to be W_a.

Let $Q = a_1a_2 \ldots a_n$ be a fragment of f. Since $f \,|\, R$ via φ, we have $R = X\varphi(Q)Y$ for some words X and Y. Viewing Q as a_1C_1, by the slide condition we have $X = Z_1P'_{a_1}$ for some $Z_1 \in L^*$. Thus, we can write

$$\begin{aligned} R &= ZP'_{a_1}\varphi(a_1)\varphi(a_2)\ldots\varphi(a_n)Y \\ &= Z\varphi'(a_1)U_{a_1}\varphi(a_2)\ldots\varphi(a_n)Y \end{aligned}$$

Viewing Q as $B_2a_2C_2$, by the slide condition we have $Z\varphi'(a_1)U_{a_1} = Z'P'_{a_2}$ for some $Z' \in L^*$. Since L is prefix, this implies that $U_{a_1} = P'_{a_2}$. Hence,

$$\begin{aligned} R &= Z\varphi'(a_1)U_{a_1}\varphi(a_2)\ldots\varphi(a_n)Y \\ &= Z\varphi'(a_1)P_{a_2}\varphi(a_2)\ldots\varphi(a_n)Y \\ &= Z\varphi'(a_1)\varphi'(a_2)U_{a_2}\ldots\varphi(a_n)Y \end{aligned}$$

Continuing in this fashion, we see that

$$R = Z\varphi'(a_1)\varphi'(a_2)\ldots\varphi'(a_n)U_{a_n}Y$$

Hence $\varphi'(Q)$ is a subword of R.

The remainder of the proof is clear. □

An analogous version of Proposition 1.6.3 holds for suffix slide setups satisfying the suffix slide condition.

1.7 Locked Formulas

A *locked formula* is a formula having no free sets. In their 1989 paper, Baker, McNulty, and Taylor [1] proved the following result.

Proposition 1.7.1. *Every locked formula is* 4*-avoidable.*

They actually showed something stronger: every locked formula f is avoided by the D0L-system $\Omega = 01/21/03/23$. Set $\mathbf{W} = \Omega^{\omega}(0)$. We reprove their result, isolating some lemmas that will be useful both here and in Chapter 2.

Lemma 1.7.2. *Let X be a subword of $\mathbf{W}$ having at least length 2. If X begins with* 1 *or* 3*, then every occurrence of X in $\mathbf{W}$ is preceded by the same letter.*

Proof. In $\mathbf{W}$, the odd index positions are either 0 or 2 and the even index positions are either 1 or 3. Moreover, since $g(\mathbf{W}) = \mathbf{W}$, we see that the odd index positions alternate between 0 and 2. Wherever X occurs in $\mathbf{W}$, it must be preceded by the alternate of its second letter. □

Say that letter $x \in \{0, 1, 2, 3\}$ is even (resp. odd) if x is 0 or 2 (resp. 1 or 3).

Lemma 1.7.3. *Suppose formula f divides $\Omega^m(0)$ via φ. If x_1x_2, x_3x_2, x_3x_4, x_5x_4, …, $x_{2n-1}x_{2n}$ is a sequence of transitions in f, then either*

1. *Every $\varphi(x_i)$, i odd, ends with an odd letter and every $\varphi(x_j)$, j even, begins with an even letter,* ***or***
2. *Every $\varphi(x_i)$, i odd, ends with an even letter and every $\varphi(x_j)$, j even, begins with an odd letter.*

Proof. This result is an immediate consequence of the alternation of even and odd letters in $\Omega^m(0)$. □

Corollary 1.7.4. *Let φ show $f \mid \Omega^m(0)$. If $a \in \operatorname{alph}(f)$ is not free, then $|\varphi(a)|$ is even.*

Proof. If a is not free in f, there exists a sequence of transitions ax_2, x_3x_2, x_3x_4, x_5x_4, $\dots$, $x_{2n-1}a$. Applying Lemma 1.7.3 with $x_1 = x_{2n} = a$, we see that $\varphi(a)$ begins and ends with letters of opposite parity. □

Lemma 1.7.5. *Suppose that $f \mid \Omega^n(0)$. Then, there exists a homomorphism φ which shows $f \mid \Omega^n(0)$ with $|\varphi(a)| = 1$ for some $a \in \operatorname{alph}(f)$.*

Proof. Suppose $f \mid \Omega^m(0)$, with m minimal, and let φ show the divisibility. By assumption, $m \leq n$. Suppose $|\varphi(a)| \geq 2$ for all $a \in \operatorname{alph}(f)$. We take L to be $\{ 01, 03, 21, 23 \}$ and set $R = \Omega^m(0)$, and we consider the prefix slide setup (L, R, f, φ). If $\varphi(a)$ begins with 1 or 3, set P'_a to be the letter determined by Lemma 1.7.2 for $X = \varphi(a)$; otherwise, set $P'_a = \varepsilon$. It is routine to check that the prefix slide condition holds. Hence, there exists φ' which shows $f \mid \Omega^m(0)$ with $\varphi'(a) \in L^*$ for all $a \in \operatorname{alph}(f)$. A simple check shows $\varphi'(a) \neq \varepsilon$ for all a. Thus, φ' induces a homomorphism which shows $f \mid \Omega^{m-1}(0)$, contradicting the minimality of m. Since $\Omega^m(0)$ is a prefix of $\Omega^n(0)$, the result is shown. □

We now prove Proposition 1.7.1.

Proof. Suppose φ shows $f \mid \Omega^m(0)$ for some m. On one hand, we may assume that $|\varphi(a)| = 1$ for some $a \in \operatorname{alph}(f)$. On the other hand, since f is locked, a is not free, and so $\varphi(a)$ must have even length. This, of course, is a contradiction. □

We close this chapter with the following result of Cassaigne [8].

Proposition 1.7.6. *Let f be a formula having only fragments of length 2. Then, f is either unavoidable or contains a locked subformula. Hence, either $\operatorname{ind}(f) \leq 4$ or $\operatorname{ind}(f) = \infty$.*

CHAPTER 2

Avoidance Bases

2.1 Introduction

In this chapter, we begin the search for an avoidable formula having index 5 or higher. Since $f \,|\, g$ implies $\mathrm{ind}(f) \geq \mathrm{ind}(g)$, the avoidable formulas with the highest indices should be those not divisible by any other avoidable formula. To this end, we define an "avoidance basis"—a collection of formulas having this property.

After proving some basic results, we construct an avoidance basis for avoidable formulas on three letters. We then show that every element in it is 4-avoidable. Thus, any index 5 formula must be on an alphabet of 4 or more letters.

2.2 Avoidance Bases

Let $\mathrm{Av}(n)$ be the set of avoidable formulas on the alphabet $\Delta_n = \{\, a_1, a_2, \ldots, a_n \,\}$.

Definition 2.2.1. *A subset S of* $\mathrm{Av}(n)$ *is an n-*avoidance basis *if:*

1. *For all $f \in \mathrm{Av}(n)$, there exists $g \in S$, not necessarily unique, such that $g \,|_e\, f$.*

2. *If $f, g \in S$, then $f \,|_e\, g$ implies $f = g$.*

If f is a member of an n-avoidance basis, then f is n-*minimal.* A formula f is *minimal* if f is n-minimal for some n.

Theorem 2.2.2. *n-avoidance bases exist for every n.*

Proof. Let $T = \{ f \in \mathrm{Av}(n) \mid \text{ every fragment of } f \text{ has length at most } 2^n \}$. Since T is a subset of the power set of $\{ W \in \Delta_n^* \mid |W| \leq 2^n \}$, T is finite.

T has property (1). To see this, choose $f \in \mathrm{Av}(n)$. If $f \in T$, we can take $g = f$. Otherwise, f has a fragment P of length greater than 2^n. Let Q be a subword of P having length 2^n. Regarding Q as a pattern, we see that $Q \in T$ by Theorem 1.5.3. Clearly, $Q \,|_e\, f$.

Let T' be a subset of T of minimal size satisfying property (1). We claim T' has property (2). Let f and g be distinct elements of T'. If $f \,|_e\, g$, then $T' \setminus \{ g \}$ has property (1), which contradicts the minimality of T'. □

Lemma 2.2.3. *Let f be minimal and let g be avoidable. Then, $g \,|_e\, f$ if and only if $g \sim_e f$. In particular, $g \,|_e\, f$ implies that* $\mathrm{alph}(g) = \mathrm{alph}(f)$.

Proof. Let f be in n-avoidance basis S and let g be avoidable. If $g \,|_e\, f$, then $|\,\mathrm{alph}(g)| \leq |\,\mathrm{alph}(f)| \leq n$. After relabeling the letters of g, we can assume that $g \in \mathrm{Av}(n)$. Since g is avoidable, there exists $f' \in S$ such that $f' \,|_e\, g$. By transitivity, we have $f' \,|_e\, f$. Since $f, f' \in S$, we must have $f = f'$. This implies that $f \sim_e g$. The converse is clear. □

The following proposition shows that n-avoidance bases are unique up to e-equivalence of their elements.

Proposition 2.2.4. *Let S be an n-avoidance basis and let $T \subset \mathrm{Av}(n)$. Then, T is an n-avoidance basis if and only if there exists a bijection $\Psi : S \to T$ such that $f \sim_e \Psi(f)$ for all $f \in S$.*

Proof. Suppose T is an n-avoidance basis. Let $f \in S$. Since f is avoidable, there exists $g \in T$ such that $g \,|_e\, f$. By Lemma 2.2.3, $f \sim_e g$. If $g' \in T$ also e-divides f, then $g' \sim_e f \sim_e g$. In particular, $g' \,|_e\, g$. Since T is an n-avoidance basis, $g' = g$.

Define $\Psi(f) = g$. If $\Psi(f) = \Psi(f') = g$, then $f' \sim_e g \sim_e f$, so $f = f'$. Thus, Ψ is injective. If $g \in T$, there exists $f \in S$ such that $f \,|_e\, g$. Since T is an n-avoidance basis, there exists $g' \in T$ such that $g' \,|_e\, f$. Transitivity implies that $g' = g$, and so $\Psi(f) = g$. Therefore, Ψ is surjective.

Now, suppose there exists a bijection $\Psi : S \to T$ such that $f \sim_e \Psi(f)$ for all f. If $h \in \mathrm{Av}(n)$, there exists $f \in S$ such that $f \,|_e\, h$. By assumption, $\Psi(f) \,|_e\, f$, so by transitivity, $\Psi(f) \,|_e\, h$ and property (1) of the definition of an n-avoidance basis is satisfied. If $g, g' \in T$ and $g \,|_e\, g'$, then $\Psi^{-1}(g) \,|_e\, \Psi^{-1}(g')$. This implies $\Psi^{-1}(g) = \Psi^{-1}(g')$, so $g = g'$. Property (2) is satisfied. Hence, T is an n-avoidance basis. □

Proposition 2.2.5. *If f is n-minimal, then f is $(n+1)$-minimal. Hence, we can construct a chain $S_1 \subset S_2 \subset S_3 \subset \ldots$, where each S_n is an n-avoidance basis.*

Proof. Suppose f is n-minimal. Let S be an $(n+1)$-avoidance basis. Since $f \in \mathrm{Av}(n+1)$, there exists $g \in S$ such that $g \,|_e\, f$. By the Lemma 2.2.3, this implies that $g \sim_e f$. By Lemma 2.2.4, we can substitute f for g in the avoidance basis S. This means that f is $(n+1)$-minimal. □

Proposition 2.2.5 does not hold if standard divisibility were used instead of e-divisibility in Definition 2.2.1. Indeed, Proposition 2.2.5 directly relies on Lemma 2.2.3, which uses the crucial fact that $f \,|_e\, g$ implies $|\,\mathrm{alph}(f)| \leq |\,\mathrm{alph}(g)|$.

The following lemma provides an intrinsic test for minimality.

Lemma 2.2.6. *Let f be an irredundant formula. Then, f is minimal if and only if all of the following hold:*

1. *The formula f is avoidable*

2. *Every fragment of f has at least length 2*

3. *Every simplification of f is unavoidable*

Proof. Let f be irredundant and set $n = |\operatorname{alph}(f)|$.

Suppose every fragment of the avoidable formula f has at least length 2 and every simplification of f is unavoidable. Let S be an n-avoidance basis. There exists $g \in S$ such that $g \,|_e\, f$. The formula g is e-equivalent to an irredundant formula g' which i-divides f. By Proposition 1.3.6, either $g' \sim_i f$ or $g' \,|_i\, \operatorname{Simp}(f, P)$ for some $P \in f$. The latter is impossible by assumption, so f is minimal.

If f is unavoidable, then f is not minimal.

If f is avoidable, but has a fragment of length 1, then $f = g.x$ for some $x \notin \operatorname{alph}(g)$, since f is irredundant. Deleting the free set x from f yields g. Since f is avoidable, g is avoidable. By Lemma 2.2.3, f cannot be minimal since $\operatorname{alph}(g) < \operatorname{alph}(f)$.

Finally, suppose $\operatorname{Simp}(f, P)$ is avoidable for some $P \in f$. If f were minimal, then by Lemma 2.2.3, $f \,|_e\, \operatorname{Simp}(f, P)$, say via φ. We show that in fact f i-divides $\operatorname{Simp}(f, P)$. Since $\operatorname{Simp}(f, P) \,|_i\, f$ and $f \,|_e\, \operatorname{Simp}(f, P)$, we have φ showing $\operatorname{Simp}(f, P) \,|_e\, \operatorname{Simp}(f, P)$. This means φ^{-1} shows $\operatorname{Simp}(f, P) \,|_e\, \operatorname{Simp}(f, P)$, too. Combining $f \,|_e\, \operatorname{Simp}(f, P)$ via φ and $\operatorname{Simp}(f, P) \,|_e\, \operatorname{Simp}(f, P)$ via φ^{-1}, we have $f \,|_i\, \operatorname{Simp}(f, P)$ as desired. This contradicts f being irredundant. □

2.3 Finding Avoidance Bases

We proved the existence of n-avoidance bases in Theorem 2.2.2. In this section, we describe a constructive way of generating them. Intuitively, the deletion of any free set from a minimal formula should result in a formula which is not only avoidable, but "close" to being minimal itself. Reversing this idea, if we have collection of minimal formulas on a small alphabet, we can try to insert letters into them hoping to obtain a minimal formula on a larger alphabet. This process, formalized below, forms "principal divisors", which are ideal candidates for minimality. We then apply Lemma 2.2.6 to decide which ones are minimal.

In Chapter 1, we defined the adjacency graph of a formula. We now generalize this idea. Let Σ be an alphabet, and let $\Delta_L = \{\, a_L \mid a \in \Sigma \,\}$ and $\Delta_R = \{\, a_R \mid a \in \Sigma \,\}$. An *adjacency graph* over Σ is a bipartite graph G with vertex set $\Delta_L \cup \Delta_R$ containing only edges between Δ_L and Δ_R. We sometimes write xy for an edge $\{\, x_L, y_R \,\}$ of G. If G is an adjacency graph over Σ, a subset $S \subset \Sigma$ is *free* if for all $a, b \in S$, a_L and b_R lie in different components of G. Clearly, if f is a formula, then $\mathrm{AG}(f)$ is an adjacency graph, and S is free in f if and only if S is free in $\mathrm{AG}(f)$. A graph G is an *avoidable adjacency graph* if $G = \mathrm{AG}(f)$ for some avoidable formula f. If G is an adjacency graph, let $\mathrm{form}(G)$ be the formula $\{\, xy \mid xy \text{ is an edge in } G \,\}$.

Definition 2.3.1. *Suppose G is an adjacency graph, and let S be a free set of G. A formula f is an S-*subformula *of G if:*

1. *The adjacency graph of f, $\mathrm{AG}(f)$, is a subgraph of G.*
2. *The formula $\sigma_S(f)$ obtained by deleting the letters of S from f is minimal.*
3. *No simplification of f satisfies* (1) *and* (2).

For example, consider $G = \{\, ab, bc, ca, cb \,\}$. The free sets are $\{\, b \,\}$ and $\{\, c \,\}$. The $\{\, b \,\}$-subformulas are *abca.cabc* and *cbc*. The $\{\, c \,\}$-subformulas are *abca.bcab* and *bcb*. As another example, let $f = abcabc$, and set $G = \mathrm{AG}(f) = \{\, ab, bc, ca \,\}$. The set $\{\, a \,\}$ is free in G, and the $\{\, a \,\}$-subformula is *bcab.cabc*.

If G is an avoidable adjacency graph, then for every free S in G, S-subformulas exist. Indeed, suppose $G = \mathrm{AG}(f)$. Since $\sigma_S(f)$ is avoidable, it must be i-divisible by some minimal formula g. This means some formula g'—which is g with elements of S inserted in its fragments—must i-divide f. This formula g' is an S-subformula.

If G is an adjacency graph over Δ_n and f is an S-subformula of G, then $\sigma_S(f)$ is a minimal formula over $n - |S|$ letters. Let $\sigma_S(f) = \Pi P_i$ and $f = \Pi Q_i$. By property (3), each Q_i can be written $a_1 B_1 a_2 B_2 \ldots B_{m-1} a_m$, with each B_j either ε or an element of S, and $P_i = a_1 a_2 \ldots a_m$. Thus, the set of S-subformulas of G can be constructed as follows:

1. Choose an avoidance basis T over $(n - |S|)$ letters.
2. For every $g \in T$, let $U(g)$ be the set of formulas over $\Delta_n \setminus S$ which are irredundant and e-equivalent to g.
3. For every $g' \in U(g)$, let $V(g')$ be the set of formulas g'' such that:
   (a) $\sigma_S(g'') = g'$. That is, g'' is g' with letters from S inserted in its fragments.
   (b) $\mathrm{AG}(g'')$ is a subgraph of G
4. The set of S-subformulas of G is the union of $V(g')$ over all $g' \in U(g)$ and $g \in T$.

Definition 2.3.2. *Let G be an avoidable adjacency graph. A* principal divisor *of G is any formula f of the form*

$$f = \mathrm{irr}((\Pi f_S).\,\mathrm{form}(G)),$$

where the product ranges over all free sets S in G and each f_S is an S-subformula of G. If G has no free sets, then form(G) *is its only principal divisor.*

A principal divisor of G is irredundant and avoidable.

Lemma 2.3.3. *The formula aa is minimal.*

Proof. Apply Lemma 2.2.6. □

Lemma 2.3.4. *Let f be a minimal formula with $|\operatorname{alph}(f)| > 1$. Then,* $\mathrm{AG}(f)$ *satisfies the following:*

1. *It has no isolated vertices.*

2. *If it has no free sets, then it must be a tree.*

3. *It does not contain the edge $\{ a_L, a_R \}$ for any $a \in \operatorname{alph}(f)$.*

Proof. If $\mathrm{AG}(f)$ had an isolated vertex, say a_L, then a would be free in f. Let $f' = \sigma_a(f)$. Since a_L is isolated, whenever a appeared in a fragment, it appeared in the rightmost position. This means that $f' \,|_i\, f$. Moreover, f' must be avoidable, contradicting Lemma 2.2.3.

Suppose $\mathrm{AG}(f)$ has no free sets. If $\mathrm{AG}(f)$ has more than one component, then f is divisible by an avoidable formula whose adjacency graph has only one component. If f has a fragment P having at least length 3, then we could simplify on P and f would still be avoidable. If f has only fragments of size 2 but is not a tree, then $\mathrm{AG}(f)$ contains a circuit. Simplifying on any edge in the circuit would still leave a tree, and so f would still be avoidable.

We can assume that aa is an element of any avoidance basis and, hence, no other minimal formula can be divisible by it. Thus, $\mathrm{AG}(f)$ does not contain the edge $\{ a_L, a_R \}$. □

Theorem 2.3.5. *Let T be an $(n-1)$-avoidance basis, and let A be the set of all principal divisors of all avoidable adjacency graphs on Δ_n. If $A' = \{ f \in A \mid f \text{ is minimal } \}$, then $A' \cup T$ contains an n-avoidance basis.*

Proof. Every n-minimal formula f is a principal divisor of $G = \mathrm{AG}(f)$. □

In the preceding theorem, we could restrict ourselves to those avoidable adjacency graphs having the properties listed in Lemma 2.3.4.

2.4 Avoidance Bases on Small Alphabets

In this section, we give an n-avoidance basis for $n = 1, 2, 3$. We prove directly that a 1-avoidance basis must be $\{ aa \}$, and then use Theorem 2.3.5 to generate 2- and 3-avoidance bases. For notational convenience, we write a for a_1, b for a_2, and c for a_3.

Theorem 2.4.1. *$\{ aa \}$ is a 1-avoidance basis.*

Proof. By Lemma 2.3.3, aa is minimal. If $f \in \mathrm{Av}(1)$, then the formula f must have a fragment of length at least 2, so f is e-divisible by aa. □

Theorem 2.4.2. *$\{ aa, aba.bab \}$ is a 2-avoidance basis.*

Proof. The only adjacency graph satisfying the conditions of Lemma 2.3.4 is $\{ ab, ba \}$. The free sets are $\{ a \}$ and $\{ b \}$. The only $\{ a \}$-subformula is bab, while the only $\{ b \}$-subformula is aba. Therefore, the only principal divisor is $aba.bab$. Simplifying on aba, we get $ab.ba.bab \sim bab$, which is clearly unavoidable. Similarly, simplifying on bab yields an unavoidable formula. By Lemma 2.2.6, $aba.bab$ is minimal. □

Now, we consider 3-letter formulas. For 3-minimal formulas which are not 2-minimal, there are four adjacency graphs (up to permutation of the letters) satisfying the conditions of Lemma 2.3.4:

1. $\{ ab, bc, ca \}$

2. $\{ ab, ba, bc, cb \}$

3. $\{ ab, bc, ca, cb \}$

4. $\{ ab, ac, ba, ca, cb \}$

$\{ ab, bc, ca \}$. The free sets are $\{ a \}$, $\{ b \}$, and $\{ c \}$. The only $\{ a \}$-subformula is *bcab.cabc*. Similarly, the only $\{ b \}$-subformula is *cabc.abca* and the only $\{ c \}$-subformula is *abca.bcab*. Hence, *abca.bcab.cabc* is the only principal divisor. By Lemma 2.2.6, it is minimal.

$\{ ab, ba, bc, cb \}$. The free sets are $\{ a \}$, $\{ b \}$, $\{ c \}$, and $\{ a, c \}$. There is one possibility for the $\{ a \}$-subformula, namely *bab*. Similarly, for the $\{ c \}$-subformula, we have *bcb*. There are several $\{ b \}$-subformulas: *aba*, *cbc*, and *abcba.cbabc*. Finally, there are two $\{ a, c \}$-subformulas, *bab* and *bcb*. Of the six possible principal divisors that can be formed, only *abcba.cbabc* is minimal.

$\{ ab, bc, ca, cb \}$. The free sets are $\{ b \}$ and $\{ c \}$. The $\{ b \}$-subformulas are *cbc* and *abca.cabc* and the $\{ c \}$-subformulas are *bcb* and *abca.bcab*. Thus, for the principal divisors, we have *bcb.cbc.ab.ca*, *abca.cabc.bcb*, *abca.bcab.cbc*, and *abca.bcab.cabc.cb*. The first and last of these are clearly not minimal. Using Lemma 2.2.6, we see that the middle two are minimal.

$\{ ab, ac, ba, ca, cb \}$. This graph has no free sets, and so *ab.ac.ba.ca.cb* is the only principal divisor. A quick check shows that it is minimal.

The result is summarized Table 2.1

2.5 Formulas on Three Letters Are 4-Avoidable

In this section, we show that every 3-minimal formula is 4-avoidable. Hence, any index 5 formula, if it exists, must have four letters.

aa
$aba.bab$
$abca.bcab.cabc$
$abcba.cbabc$
$abca.cabc.bcb$
$abca.bcab.cbc$
$ab.ac.ba.ca.cb$

Table 2.1: A 3-Avoidance Basis

Define the n-th circular formula C_n to be

$$C_n = \{\, a_k a_{k+1} \ldots a_n a_1 a_2 \ldots a_k \mid a_k \in \Delta_n \text{ and } 1 \le k \le n \,\}.$$

Using the convention $a = a_1$, $b = a_2$, $c = a_3$, and so forth, we see that the first few circular formulas are aa, $aba.bab$, and $abca.bcab.cabc$.

Proposition 2.5.1. *Every circular formula is* 4*-avoidable.*

Proof. We show that every circular formula is avoided by $\Omega = 01/21/03/23$. The proof is by induction on n.

If $n = 1$, we have $C_1 = a_1a_1$, which is locked. By Proposition 1.7.1, the result follows.

Suppose the result holds for $n < k$ but not for $n = k$. Suppose C_k divides $\Omega^m(0)$ via φ, with m minimal. This means $\varphi(a_j a_{j+1} \ldots a_k a_1 a_2 \ldots a_{j-1} a_j)$ is a subword of $\Omega^m(0)$ for $j = 1, 2, \ldots, k$. There are two cases to consider.

First, suppose that for some a_i, say a_1, we have $\varphi(a_i) = 0$ or $\varphi(a_i) = 2$. Since $\varphi(a_1 \ldots a_k a_1) \le \Omega^m(0)$, Lemma 1.7.2 implies that every occurrence of $X = \varphi(a_2 \ldots a_k a_1)$ is preceded by $\varphi(a_1)$. In particular, the image of the second fragment, $\varphi(a_2 \ldots a_k a_1 a_2)$, must be preceded by $\varphi(a_1)$. This fact contradicts the induction hypothesis since C_{k-1} divides $\Omega^m(0)$ via φ', where $\varphi'(a_1) = \varphi(a_1a_2)$ and $\varphi'(a_i) = \varphi(a_{i+1})$ for $i = 2, 3, \ldots, k-1$.

Now, suppose $\varphi(a_i) \neq 0$ and $\varphi(a_i) \neq 2$ for all $i > 0$. For convenience, let $a_0 = a_k$ and $a_{k+1} = a_1$. Set $L = \{\,01, 21, 03, 23\,\}$ and $R = \Omega^m(0)$, and consider the prefix slide setup (L, R, C_k, φ). Let $a_i \in \Delta_k$. If $\varphi(a_i)$ begins with 0 or 2, set $P'_{a_i} = \varepsilon$; otherwise, set P'_{a_i} to be the letter determined by Lemma 1.7.2 for the word $\varphi(a_i)\varphi(a_{i+1})$. Since $\varphi(a_{i-1}a_ia_{i+1})$ is a subword of $\Omega^m(0)$, $\varphi(a_{i-1})$ must end with P'_{a_i}. With this fact, one can verify the prefix slide condition. By Proposition 1.6.3, there exists φ' which shows $f \,|\, R$ such that $\varphi'(a) \in L^+$ for all $a \in \Delta_k$. (A simple argument shows $\varphi'(a) \neq \varepsilon$ for all a.) This implies $C_k \,|\, \Omega^{m-1}(0)$, which contradicts the choice for m. □

Corollary 2.5.2. *aa, aba.bab, and abca.bcab.cabc are* 4*-avoidable.*

The above result is not tight. For example, Thue [13] showed that *aa* is 3-avoidable and Cassaigne [8] later showed *aba.bab* is also 3-avoidable. It appears likely that *abca.bcab.cabc* is 3-avoidable as well.

Now we will show that the remaining three formulas are 4-avoidable. Both *abcba.cbabc* and *abca.cabc.bcb* are avoided by $g = 01/2/031/3$. Since *abca.bcab.cbc* is e-equivalent to the reversal of *abca.cabc.bcb*, they will have the same index by Corollary 1.4.3.

Let $\mathbf{G} = g^\omega(0)$ and observe that g is biprefix.

Lemma 2.5.3. *The only subwords of* $\mathbf{G}$ *of length 3 are 012, 013, 031, 101, 120, 123, 130, 132, 201, 203, 230, 301, 303, 310, 313, 320, and 323.*

Proof. $\mathbf{G}$ begins $012\ldots$. Let S be the set of subwords of $\mathbf{G}$ of length 3. This set is determined by two properties:

1. $012 \in S$
2. $S = \{\, U \mid U \leq g(V) \text{ for some } V \in S \text{ and } |U| = 3 \,\}$

The above list of words satisfies both these properties. □

Corollary 2.5.4. *Every subword of* $\mathbf{G}$ *except* 3 *contains a cut. Consequently, every subword other than* 3 *can be decomposed as* $S|X|P$*, where* S *is one of* ε*,* 1*, or* 31*;* $X \in \{\, 01, 2, 031, 3 \,\}^*$*; and* P *is one of* ε*,* 0*, or* 03*.*

Proof. If a word contains a 0, the cut must preceed the 0; if a word contains a 1, the cut must follow the 1; and if a word contains a 2, then there are cuts before and after the 2. Since 33 is not a subword of $\mathbf{G}$, any word having length greater than one contains a 0, 1, or 2, and so must have a cut. □

Lemma 2.5.5.

1. *If* XS *is a subword of* $\mathbf{G}$ *with* $|X| \geq 4$ *and* S *one of* ε*,* 1*, or* 31*, then every occurrence of* X *in* $\mathbf{G}$ *is followed by* S*.*
2. *If* PX *is a subword of* $\mathbf{G}$ *with* $|X| \geq 4$ *and* P *one of* ε*,* 0*, or* 03*, then every occurrence of* X *in* $\mathbf{G}$ *is preceded by* P*.*

Proof. We prove only the first statement, the proof of the second being similar. If $S = \varepsilon$, the result is trivially true. If X ends with 03, then X can only be followed by 1. So, we can assume that X ends with 0.

If X ends with 10, then X must be followed by 1 since 103 is not a subword of $\mathbf{G}$ by Lemma 2.5.3.

If X ends with 120, then X must be followed by 31. Otherwise, $1201 = 1|2|01|$ would be a subword of $\mathbf{G}$. Pulling back (that is, applying g^{-1}), this implies that either 010 or 210 is a subword of $\mathbf{G}$, which contradicts Lemma 2.5.3.

If X ends with 320, then X must be followed by 1. Otherwise, $32031 = |3|2|031|$ would be subword of $\mathbf{G}$. This is impossible since the pullback, 312, is not a subword of $\mathbf{G}$.

If X ends with 130, then X must be followed by 1. Otherwise, $13031 = 1|3|031|$ is a subword of $\mathbf{G}$. Pulling back, we get either 032 or 232, both of which are not subwords of $\mathbf{G}$.

If X ends with 1230, then X must be followed by 31. Otherwise, 12301 would be a subword of $\mathbf{G}$, and so would one of its pullbacks—0130 or 2130. The latter is impossible. Pulling back 0130 yields 030 or 032, both of which are impossible.

If X ends with 3230, then X must be followed by 1. Otherwise, 323031 is a subword of $\mathbf{G}$. Its pullback is $3132 = 31|3|2|$, which pulls back to 231, which is not a subword of $\mathbf{G}$. □

Lemma 2.5.6. $\mathbf{G}$ *avoids* aa.

Proof. Suppose aa divides $g^m(0)$ via φ, with m minimal. One can verify by exhaustion that $|\varphi(a)| < 4$ is impossible. Assuming $|\varphi(a)| \geq 4$, we can write $\varphi(a)$ as $S|A|P$ as in Corollary 2.5.4. Substituting, we get $S|A|PS|A|P$. By Lemma 2.5.5 with $X = S|A|P$, we can infer that $S|A|PS|A|P$ is preceded by P. Thus, we can conclude that $|PS|A|PS|A|$ is a nonempty subword of $\mathbf{G}$. This implies that aa divides $g^{m-1}(0)$, which contradicts the choice of m. □

Lemma 2.5.7. $\mathbf{G}$ *avoids* $aba.bab$.

Proof. Suppose $aba.bab$ divides $g^m(0)$ via φ, with m minimal. One can verify by exhaustion that $|\varphi(a)| + |\varphi(b)| < 4$ is impossible. Now, assume that $|\varphi(a)| + |\varphi(b)| \geq 4$. Since $\varphi(a) = \varphi(b) = 3$ is impossible, we may assume without loss of generality that $\varphi(a) \neq 3$, and hence it can be written as $\varphi(a) = S|A|P$ as in

Corollary 2.5.4. Substituting, we get $S|A|PBS|A|P$ and $BS|A|PB$ as subwords of $g^m(0)$. By Lemma 2.5.5, we can infer that $BS|A|PB$ must be preceded by P and followed by S, yielding $|PBS|A|PBS|$. If $A = \varepsilon$, then we have a square, which would contradict Lemma 2.5.6. If $A \neq \varepsilon$, then *aba.bab* divides $g^{m-1}(0)$ via $\varphi'(a) = g^{-1}(A)$ and $\varphi'(b) = g^{-1}(PBS)$. □

Lemma 2.5.8. *If $X \in \{0,1,2,3\}^+$, then* **G** *does not contain $X3X$.*

Proof. Suppose $g^m(0)$ contains $X3X$, with m minimal. Since **G** avoids *aa*, $X \neq 3$. Hence, we can write $X = S|Y|P$ as in Corollary 2.5.4. Substituting, we get $X3X = S|Y|P3S|Y|P$. If $P = \varepsilon$, then we have $S = \varepsilon$, so $Y \neq \varepsilon$. Pulling back, $g^{m-1}(0)$ would also contain a subword of the form $X3X$, which contradicts the minimality of m. So, we assume that $P \neq \varepsilon$, which implies P must be 0 and S must be 1. Since 103 is not a subword of $g^m(0)$, $Y \neq \varepsilon$. If $Y = 2$, then $1|2|031|2|0$ would be a subword of $g^m(0)$. This word, however, contains 203, which is impossible. If $Y = 3$, we would have $1|3|031|3|0$, which pulls back to 03230, 23230, 03232, or 23232. The first one contains 032 as a subword, which is impossible. The last three contain squares, which is impossible by Lemma 2.5.6. Hence, $|Y| > 1$, and so Lemma 2.5.5 applies. The word must be preceded by $P3$ and followed by $3S$. This gives $|P3S|Y|P3S|Y|P3S|$, which is impossible since it is divisible by *aa*. □

Lemma 2.5.9. *Suppose $X|Y|X$ is a subword of* **G** *for some words X and Y. Then, $X|Y|X = |X|Y|X|$.*

Proof. If $X = \varepsilon$, the result is clear. If $X = |X$, the result holds because **G** is prefix implies $|X|Y|X = |X|Y|X|$. This takes care of the cases where X begins with 0, 2, 30, or 32. If $X = 3$, then clearly $3|Y|3 = |3|Y|3|$. Finally, we note that X cannot begin with 1. □

Lemma 2.5.10. *If* $0X1$ *is a subword of* $\mathbf{G}$*, then* X *is* ε*,* 3*, or a word having at least length 3. In particular, if* $0X31$ *is subword of* $\mathbf{G}$*, then* X *is either* ε *or a word having at least length 2.*

Proof. Suppose $0X1 \leq \mathbf{G}$. If $|X| = 1$, then X must be 3. If $|X| = 2$, then X could only possibly be 10, but this implies $aa \,|\, \mathbf{G}$. □

Lemma 2.5.11. *For all words* $X, Y \in \{0, 1, 2, 3\}^*$ *and all letters* $z \in \{0, 1, 2\}$*,* $\mathbf{G}$ *does not contain both* $XzY3Xz$ *and* $zY3XzY$*.*

Proof. Suppose $g^m(0)$ contains $XzY3Xz$ and $zY3XzY$ for some choice of X, Y, and z. We assume that m is minimal with respect to this property. Neither X nor Y can be ε by Lemma 2.5.8. We examine the possibilities for z.

Case 1: $z = 0$. Substituting, we get $X0Y3X0 = X|0Y3X|0$ and $0Y3X0Y = |0Y3X|0Y$. Since Y cannot be 3, so we can write $Y = S|Y'|P$ as in Corollary 2.5.4. Substituting, we get $X|0S|Y'|P3X|0$ and $|0S|Y'|P3X|0S|Y'|P$. If $P = \varepsilon$, we have $|X|0S|Y'|3|X|0$ and $|0S|Y'|3|X|0S|Y'|$. Lemma 2.5.5 applies, so the first of these words must be followed by S. Pulling back, we get words of the form $XzY3Xz$ and $zY3XzY$, $z \neq 3$, which contradicts the choice for m. If $P \neq \varepsilon$, then $P = 0$. Again by Lemma 2.5.5, the words must be $|P3X|0S|Y'|P3X|0S|$ and $|0S|Y'|P3X|0S|Y'|P$. However, *aba.bab* divides this.

Case 2: $z = 1$. Plugging in, we get $X1Y3X1 = X1|Y3X1|$ and $1Y3X1Y = 1|Y3X1|Y$. We have that $X \neq 3$, so we can write it as $S|X'|P$. After replacing, we have $S|X'|P1|Y3S|X'|P1|$ and $1|Y3S|X'|P1|Y$. If $S = \varepsilon$, we get $|X'|P1|Y|3|X'|P1|$ and $1|Y|3|X'|P1|Y|$. Lemma 2.5.5 implies that the latter is preceded by P. Pulling back, we get words of the form $XzY3Xz$ and $zY3XzY$, with $z \neq 3$, which contradicts the choice for m. If $S \neq \varepsilon$, then $S = 1$. By Lemma

2.5.5, we get $S|X'|P1|Y3S|X'|P1|$ and $|P1|Y3S|X'|P1|Y3S|$, which is divisible by *aba.bab*.

Case 3: $z = 2$. Substituting, we get $X|2|Y3X|2|$ and $|2|Y3X|2|Y$. We have that $Y \neq 3$, so we can write Y as either $Y = |Y|$ or $Y = |Y'|0$. In the first case, we get $|X|2|Y|3|X|2|$ and $|2|Y|3|X|2|Y|$, which pulls back to $X1Y3X1$ and $1Y3X1Y$, violating the choice for m. In the second case, we get $X|2|Y'|03X|2|$ and $|2|Y'|03X|2|Y'|0$. Since 312 is not a subword of $\mathbf{G}$, $|X| > 1$. This allows us to conclude that $|03X|2|Y'|03X|2|$ and $|2|Y'|03X|2|Y'|0$ are subwords. However, this means *aba.bab* divides $\mathbf{G}$. □

Lemma 2.5.12. *Suppose XYX is a subword of $\mathbf{G}$. If the length of Y is small, there are only a small number of choices for X. Specifically, the results are summarized in Table 2.2.*

Proof. The proof is straightforward. We note that the result of Lemma 2.5.8 is included in the table. □

Theorem 2.5.13. **G** *avoids abcba.cbabc.*

Proof. Suppose *abcba.cbabc* divides $g^m(0)$ via φ, with m minimal. One can show that $|\varphi(a)| + |\varphi(b)| + |\varphi(c)| < 4$ is impossible. Therefore, we suppose that $|\varphi(abc)| \geq 4$. Let $A = \varphi(a)$ and $C = \varphi(c)$. There are two cases:

1. $\varphi(b) \neq 3$. Write $\varphi(b) = S|B|P$ as in Corollary 2.5.4. Thus, we have $AS|B|PCS|B|PA$ and $CS|B|PAS|B|PC$ as subwords of $\mathbf{G}$. By Lemma 2.5.5, we can conclude that each word must be preceded by P and followed by S. This gives $|PAS|B|PCS|B|PAS|$ and $|PCS|B|PAS|B|PCS|$. If $B = \varepsilon$, then *aba.bab* divides $g^m(0)$, contradicting Lemma 2.5.7. Otherwise, *abcba.cbabc* divides $g^{m-1}(0)$, which contradicts the choice for m.

Y	X
ε	ε
0	ε, 1, 3, 13
1	ε, 3, 20, 230
3	ε
03	ε
10	ε
12	ε, 0, 0310, 30, 03130
13	ε
20	ε, 3101, 31301
30	ε, 1
31	ε, 0, 30
130	ε

Table 2.2: Possibilities for X if $XYX \leq \mathbf{G}$

2. $\varphi(b) = 3$. Plugging in, we get $A3C3A$ and $C3A3C$ as subwords. Since A cannot be 3, we can write $A = S|A'|P$ as in Corollary 2.5.4. There are several subcases:

   (a) Both S and P are ε. The subwords are $|A'|3|C|3|A'|$ and $|C|3|A'|3|C|$. This implies *abcba.cbabc* divides $g^{m-1}(0)$, which violates the the minimality of m.

   (b) Both S and P are not ε. We get $S|A'|P3C3S|A'|P$ and $C3S|A'|P3C$. By Lemma 2.5.5, the latter must be preceded by $P3$ and followed by $3S$. This gives $S|A'|P3C3S|A'|P$ and $|P3C3S|A'|P3C3S|$. If $A' = \varepsilon$, then *aa* divides $g^m(0)$, which contradicts Lemma 2.5.6. If $A' \neq \varepsilon$, then *aba.bab* divides $g^m(0)$, which is impossible by Lemma 2.5.7.

   (c) S is ε but P is not. Here, we get $|A'|P3C|3|A'|P$ and $C|3|A'|P3C|$. We see that P must be 0 and that $C = 1|C'|$, yielding $|A'|031|C'|3|A'|0$ and $1|C'|3|A'|031|C'|$. Also, either $C' \neq \varepsilon$ or $A' \neq \varepsilon$. Otherwise, we would have $|031|3|0$ and $1|3|031|$. A simple proof shows that the latter is impossible. Hence Lemma 2.5.5 applies, and we can infer that $|A'|031|C'|3|A'|031|$ and $|031|C'|3|A'|031|C'|$ are subwords of $g^m(0)$. Pulling back, this implies that $X2Y3X2$ and $2Y3X2Y$ are subwords of $\mathbf{G}$, with $g(X) = A'$ and $g(Y) = C'$. This violates Lemma 2.5.11.

   (d) P is empty but S. This subcase is similar to (c).

□

Theorem 2.5.14. $\mathbf{G}$ *avoids abca.cabc.bcb.*

Proof. Assume that *abca.cabc.bcb* divides $g^m(0)$ via φ, where m is minimal.

First assume that $|\varphi(c)| \geq 3$. There are several cases:

1. $\varphi(a) \neq 3$ and $\varphi(b) \neq 3$. We can write $\varphi(a) = S_1|A|P_1$, $\varphi(b) = S_2|B|P_2$, and $\varphi(c) = S_3|C|P_3$ as in Corollary 2.5.4. Substituting, we get:

   $$S_1|A|P_1S_2|B|P_2S_3|C|P_3S_1|A|P_1,$$
   $$S_3|C|P_3S_1|A|P_1S_2|B|P_2S_3|C|P_3,$$
   $$\text{and} \quad S_2|B|P_2S_3|C|P_3S_2|B|P_2$$

   for the words. Since $|\varphi(c)| \geq 3$, we can apply Lemma 2.5.5 to conclude that the following are subwords of $g^m(0)$:

   $$|P_3S_1|A|P_1S_2|B|P_2S_3|C|P_3S_1|A|P_1S_2|,$$
   $$|P_2S_3|C|P_3S_1|A|P_1S_2|B|P_2S_3|C|P_3S_1|,$$
   $$\text{and} \quad |P_1S_2|B|P_2S_3|C|P_3S_2|B|P_2.$$

   Since $\varphi(bca)$ and $\varphi(bcb)$ are both subwords of $\mathbf{G}$, we can also infer that $S_1 = S_2$ from Lemma 2.5.5. If AP_1S_2B is empty, we have a square in the second word. If AP_1S_2 is empty but B is not, the second and third words contain *aba.bab*, with $\varphi'(a) = P_2S_3CP_3S_1$ and $\varphi'(b) = B$. If AP_1S_2 is not empty but B is, the first and second words contain *aba.bab*, with $\varphi'(a) = AP_1S_2$ and $\varphi'(b) = P_2S_3CP_3S_1$. Finally, if neither AP_1S_2 nor B are empty, we have that *abca.cabc.bcb* divides $g^{m-1}(0)$ via $\varphi'(a) = g^{-1}(AP_1S_2)$, $\varphi'(b) = g^{-1}(B)$, and $\varphi'(c) = g^{-1}(P_2S_3CP_3S_1)$.

2. $\varphi(a) = 3$ and $\varphi(b) \neq 3$. We write $\varphi(b) = S_2|B|P_2$ and $\varphi(c) = S_3|C|P_3$, substitute, and apply Lemma 2.5.5 to get the following subwords of $g^m(0)$:

   $$|P_33S_2|B|P_2S_3|C|P_33S_2|,$$
   $$|P_2S_3|C|P_33S_2|B|P_2S_3|C|P_3,$$
   $$\text{and} \quad S_2|B|P_2S_3|C|P_3S_2|B|P_2.$$

(a) If $P_3 = \varepsilon$, then $S_2 = \varepsilon$, so the words simplify to:

$$|3|B|P_2S_3|C|3|, \quad |P_2S_3|C|3|B|P_2S_3|C|, \quad \text{and} \quad |B|P_2S_3|C|B|P_2.$$

If $B = \varepsilon$, the second word contains a block of the form $X3X$, which contradicts Lemma 2.5.8. If $B \neq \varepsilon$, we have *abca.cabc.bcb* dividing $g^{m-1}(0)$ via $\varphi'(a) = g^{-1}(3)$, $\varphi'(b) = g^{-1}(B)$, and $\varphi'(c) = g^{-1}(P_2S_3C)$.

(b) If $P_3 \neq \varepsilon$, we have a contradiction to Lemma 2.5.5: $S_3|C|P_3$ is followed by $3S_2$ in the second word, but is followed by just S_2 in the third.

3. $\varphi(a) \neq 3$ and $\varphi(b) = 3$. Write $\varphi(a) = S_1|A|P_1$ and $\varphi(c) = S_3|C|P_3$. Substituting and using Lemma 2.5.5, we get:

$$|P_3S_1|A|P_13S_3|C|P_3S_1|A|P_1,$$
$$S_3|C|P_3S_1|A|P_13S_3|C|P_3S_1|,$$
$$\text{and} \quad 3S_3|C|P_33.$$

We consider four subcases:

(a) $S_3 = P_3 = \varepsilon$. This implies that $P_1 = S_1 = \varepsilon$, so we have $|A|3|C|A|$, $|C|3|A|3|C|$, and $|3|C|3|$. Thus, *abca.cabc.bcb* divides $g^{m-1}(0)$, contradicting the minimality of m.

(b) $S_3 = \varepsilon$ and $P_3 \neq \varepsilon$. This implies $P_1 = \varepsilon$. We have:

$$|P_3S_1|A|3|C|P_3S_1|A|, \quad |C|P_3S_1|A|3|C|P_3S_1|, \quad \text{and} \quad |3|C|P_33.$$

The first two words pull back to $zA3CzA$ and $CzA3Cz$, with $z = g^{-1}(P_3S_1)$. This is impossible by Lemma 2.5.11.

(c) $S_3 \neq \varepsilon$ and $P_3 = \varepsilon$. This case also leads to a contradiction of Lemma 2.5.11.

(d) $S_3 \neq \varepsilon$ and $P_3 \neq \varepsilon$. Applying Lemma 2.5.5, we get:

$$|P_3S_1|A|P_13S_3|C|P_3S_1|A|P_13S_1|,$$
$$P_13S_3|C|P_3S_1|A|P_13S_3|C|P_3S_1|,$$
$$\text{and} \quad P_13S_3|C|P_33.$$

The first two words are divisible by *aba.bab* via $\varphi'(a) = |P_3S_1|A|$ and $\varphi'(b) = |P_13S_3|C|$. This contradicts Lemma 2.5.7.

This concludes Case 3, and so proves the result when $|\varphi(c)| \geq 3$.

If $|\varphi(c)| < 3$, then $\varphi(c)$ must be one of the following: 0, 1, 2, 3, 01, 03, 10, 12, 13, 20, 23, 30, 31, or 32.

Suppose that $\varphi(c) = |C|$. Letting $A = \varphi(a)$ and $B = \varphi(b)$, we have $AB|C|A$, $|C|AB|C|$, and $B|C|B$ for the subwords. By Lemma 2.5.9, $B|C|B = |B|C|B|$, and so the words must be $|A|B|C|A|$, $|C|A|B|C|$, and $|B|C|B|$. This means that *abca.cabc.bcb* divides $g^{m-1}(0)$, which contradicts the choice for m. Hence, we can assume that $\varphi(c)$ is not 2, 01, 23, or 32.

Using the fact that $\varphi(b)\varphi(c)\varphi(b)$ is a subword of $g^m(0)$, the remaining cases lean heavily on Lemma 2.5.12. We see immediately that 3, 03, 10, and 13 are impossible choices for $\varphi(c)$. Let $A = \varphi(a)$.

1. $\varphi(c) = 0$ and $\varphi(b) = 1$. We get:

$$A1|0A, \quad |0A1|0, \quad \text{and} \quad 101.$$

By Lemma 2.5.10, $A = 3$ or $|A| > 2$. If $A = 3$, $A1|0A$ implies a violation of Lemma 2.5.6. If $|A| > 2$, then Lemma 2.5.5 implies $A1|0A$ is preceded by 0 and followed by 1, which again produces a square.

2. $\varphi(c) = 0$ and $\varphi(b) = 3$. We get:

$$A|3|0A|, \quad |0A|3|0, \quad \text{and} \quad |3|03.$$

   By Lemma 2.5.10, $A = 1$. However, 0130 is not a subword of $\mathbf{G}$.

3. $\varphi(c) = 0$ and $\varphi(b) = 13$. We get:

$$A1|3|0A, \quad |0A1|3|0, \quad \text{and} \quad 1|3|01|3|.$$

   The word A is either 3 or has length greater than 2. In either case, the first word extends to $|0A1|3|0A1|$, which violates Lemma 2.5.10.

4. $\varphi(c) = 1$ and $\varphi(b) = 3$. We get:

$$|A31|A, \quad 1|A31|, \quad \text{and} \quad 31|3.$$

   The first word implies A is either 0 or 30. Both choices are illegal in $1|A31|$.

5. $\varphi(c) = 1$ and $\varphi(b) = 20$. We get:

$$|A|2|01|A|, \quad 1|A|2|01|, \quad \text{and} \quad |2|01|2|0.$$

   Pulling back $|A|2|01|A|$ gives a word of the form $X10X$, which is impossible by Lemma 2.5.12.

6. $\varphi(c) = 1$ and $\varphi(b) = 230$. We get:

$$|A|2|3|01|A|, \quad 1|A|2|3|01|, \quad \text{and} \quad |2|3|01|2|3|0.$$

   Pulling back the first word gives a word of the form $X130X$, which is impossible by Lemma 2.5.12.

7. $\varphi(c) = 12$ and $\varphi(b) = 0$. We get:

$$|A|01|2|A|, \quad 1|2|A|01|2|, \quad \text{and} \quad |01|2|0.$$

Pulling back $|A|01|2|A|$ twice gives a word of the form $X0X$. Using Lemma 2.5.12, we see that A must be 031, 3, or 0313. The choice $A = 3$ is impossible in $1|2|A|01|2|$. Hence, the length of A is at least 3. By Lemma 2.5.5, $1|2|A|01|2|$ must be preceded by 0. This implies $aba.bab$ divides $\mathbf{G}$.

8. $\varphi(c) = 12$ and $\varphi(b) = 0310$. We have $|A|031|01|2|A|$ for the first word. Pulling back twice yields a word of the form $X10X$, which is impossible.

9. $\varphi(c) = 12$ and $\varphi(b) = 30$. We have:

$$|A|3|01|2|A|, \quad 1|2|A|3|01|2|, \quad \text{and} \quad |3|01|2|3|0.$$

   Pulling back $|A|3|01|2|A|$ twice, we get a word of the form $X30X$. Using Lemma 2.5.12, we see that $A = 031$. By Lemma 2.5.5, the second word is preceded by 0. This implies $abca.cabc.bcb$ divides $g^{m-1}(0)$.

10. $\varphi(c) = 12$ and $\varphi(b) = 03130$. We have $|A|031|3|01|2|A|$ for the first word. Pulling back twice yields a word of the form $X130X$, which violates Lemma 2.5.12.

11. $\varphi(c) = 20$ and $\varphi(b) = 3101$. We have:

$$A31|01|2|0A, \quad |2|0A31|01|2|0, \quad \text{and} \quad 31|01|2|031|01|.$$

   Lemma 2.5.10 implies that A has length at least two. Hence, we can infer that $A31|01|2|0A$ is preceded by 0 and followed by 31, which means $g^{m-1}(0)$ contains $abca.cabc.bcb$.

12. $\varphi(c) = 20$ and $\varphi(b) = 31301$. Here, we get:

$$A31|3|01|2|0A, \quad |2|0A31|3|01|2|0, \quad \text{and} \quad 31|3|01|2|031|3|01|.$$

   By Lemma 2.5.10, $|A| \geq 2$. After applying Lemma 2.5.5, we again pullback to an encounter of $abca.cabc.bcb$.

13. $\varphi(c) = 30$ and $\varphi(b) = 1$. We get $A1|3|0A$ for the first word, which violates Lemma 2.5.12.

14. $\varphi(c) = 31$ and $\varphi(b) = 0$. We get $|A|031|A|$ and $31|A|031|$ for the first two words, the second of which must be preceded by 0. This implies $aba.bab$ divides $\mathbf{G}$.

15. $\varphi(c) = 31$ and $\varphi(b) = 30$. Here, we have:

$$|A|3|031|A|, \quad 31|A|3|031|, \quad \text{and} \quad |3|031|3|0.$$

The second word must be preceded by 0, so $abca.cabc.bcb$ divides $g^{m-1}(0)$.

This completes the proof. □

CHAPTER 3

Minimally Locked Formulas

3.1 Introduction

An unsettled question in the study of avoidable patterns is whether there exist patterns of arbitrarily high index. Heretofore, the highest known index of any avoidable pattern was 4, an example of which was given in a paper by Baker, McNulty, and Taylor [1]. At one point, it had been conjectured that avoidable patterns having index 5 or higher do not exist. (We show that this conjecture is false in Chapter 4.)

The pattern that Baker et al. used was *abwbcxcaybazac*. Baker was one of the first to notice the relationship between patterns and formulas described by Theorem 1.4.4, which states that the index of a pattern P is the same as the index of the formula f obtained by replacing all isolated variables in P by the symbol ".", discarding any empty fragments. Hence, $\mathrm{ind}(abwbcxcaybazac) = \mathrm{ind}(ab.bc.ca.ba.ac)$. Below, we will write *ab.bc.ca.ba.ac* as *ab.ba.ac.ca.bc*.

The formula *ab.ba.ac.ca.bc* is locked. It was shown in the paper by Baker et al. [1] that every locked pattern is avoided by Ω. (See Proposition 1.7.1.) Cassaigne [8] commented that a pattern is locked if and only if its associated formula is locked. Combining these two results, we have $\mathrm{ind}(f) \leq 4$ for every locked formula f.

Moreover, it was shown that *ab.ba.ac.ca.bc* was 3-unavoidable, and hence, it has index equal to 4. It is natural to ask whether one can find more examples of index 4 formulas among the set of locked formulas. Every locked formula is e-divisible by some minimal and locked formula. To see this, let f be any locked formula, and let f' be the formula having as fragments the transitions of f. If f' is minimal, we are done. Otherwise, f' is e-divisible by some minimal formula, which clearly must be locked. A formula is a *minimally locked formula* (MLF) if it is minimal and locked. If a locked formula has index 4, then any minimally locked formula dividing it must also have index 4. We conjecture that all minimally locked formulas, except *aa*, have index 4.

Every minimally locked formula has only fragments of length 2 and must not be divisible by any other locked formula. Hence, *ab.ba.ac.ca.bc* is minimally locked, but not *ab.ba.ac.ca.bc.ad.db* (it is divisible by the locked formula *ab.ba.ac.ca.bc*) and not *aba.ac.ca.bc* (it contains a length 3 fragment). As a consequence of the fragment sizes, every minimally locked formula f must be 3-avoidable since $f \mid aa$.

3.2 Indices of MLF's on Small Alphabets

In this section, we show that *ab.ba.ac.ca.bc* has index 4. We then show that every MLF on four letters is also index 4. Since every MLF is 4-avoidable, it is sufficient to demonstrate 3-unavoidability. For the simplest cases, a hand proof of this fact is available.

There is only one minimally locked formula (up to e-equivalence) on three letters, namely *ab.ba.ac.ca.bc*. Its 3-unavoidability is a direct consequence of the following lemma.

Lemma 3.2.1. *Every squarefree infinite word on three letters contains all six transitions.*

Proof. Let the alphabet be $\{a, b, c\}$, and suppose that the transition ab is missing from the squarefree infinite word $\mathbf{W}$. Some suffix of $\mathbf{W}$ must begin with a; otherwise, $\mathbf{W}$ contains only b and c, and hence, could not be squarefree. So without loss of generality, we can assume that $\mathbf{W}$ begins with ac. If $\mathbf{W}$ begins with aca, then adding a or c produces a square, while adding b yields ab. Thus, we can assume that aca does not appear as a subword of $\mathbf{W}$ and that $\mathbf{W}$ begins with acb. There are two possibilities: $\mathbf{W}$ begins with $acba$ or with $acbc$. If $\mathbf{W}$ begins with $acba$, then $\mathbf{W}$ must begin with $acbac$. At this point, we cannot continue. Adding a violates the aca exclusion, while adding b or c yields a square. Thus, we may also assume that $acba$ does not appear as a subword of $\mathbf{W}$. The second possibility, $acbc$, can only be extended to $acbcacb$, but this word allows no extension: adding a gives $acba$ as a subword, whereas adding b or c yields a square. □

To see that $ab.ba.ac.ca.bc$ is 3-unavoidable, we note that any infinite word $\mathbf{W}$ on three letters that avoids it must be squarefree, and so must have all six transitions. Using the notation of Lemma 3.2.1, this means $ab.ba.ac.ca.bc$ i-divides $\mathbf{W}$, which is impossible. As a remark, something stronger, in fact, is true: the nonminimally locked formula $ab.ba.ac.ca.bc.cb$ is index 4 as well.

Now we consider minimally locked formulas on four letters. They are (up to e-equivalence): $ab.cb.ca.da.dc.bc.bd$, $ab.cb.cd.bd.ba.da.dc$, $ab.cb.cd.ca.da.dc.bc$, $ab.cb.cd.ca.ba.da.dc$, and $ab.cb.ca.da.cd.bd.bc$; and their reversals.

Lemma 3.2.2. *Let* $\mathbf{W}$ *be an infinite squarefree word over* $\{a, b, c\}$*. For some permutation* x, y, z *of the letters* a, b, c*, there exists:*

1. *A word U such that xU, yU, and Uz are all subwords of* $\mathbf{W}$.

2. *A word V such that xV, yV, and Vx are all subwords of* $\mathbf{W}$.

Proof. The proof is by construction. Let $\mathbf{W}$ be an infinite squarefree word over $\{a, b, c\}$. There are thirty squarefree words of length 5 over $\{a, b, c\}$. One of these words must occur twice in $\mathbf{W}$, say with starting indices of m and n, such that $m, n > 30$ and $n - m \leq 30$. Set $L = 5$.

If the letters in positions $m - 1$ and $n - 1$ are the same, reset m to be $m - 1$, n to be $n - 1$, L to be $L + 1$, and repeat this step. This process must stop before $m + L = n$; otherwise, $\mathbf{W}$ contains a square.

Hence, there is a word, call it W', of length L which occurs twice, beginning at positions m and n, with each occurrence preceded by different letters. We set x and y to be those two letters, and we set z to be the third letter of the alphabet. Since every squarefree word over three letters of length greater than 3 must contain x, y, and z, we can write $W' = XzY$, where Y is a word not containing z. We set $U = X$. Similarly, we can write $W' = X'xY'$, where Y' is a word not containing x. We set $V = X'$. The words U and V satisfy the lemma. □

Lemma 3.2.3. *Let $\mathbf{W}$ be an infinite squarefree word over $\{a, b, c\}$. For some permutation x, y, z of the letters a, b, c, there exists:*

1. *A word U such that Ux, Uy, and zU are all subwords of* $\mathbf{W}$.

2. *A word V such that Vx, Vy, and xV are all subwords of* $\mathbf{W}$.

Proof. The proof is similar to that of the previous lemma. □

Theorem 3.2.4. *Every minimally locked formula on four letters has index 4.*

Proof. We need to show every MLF on four letters is 3-unavoidable.

We begin with $ab.cb.ca.da.dc.bc.bd$. Suppose it is 3-avoidable, and let $\mathbf{W}$ be an infinite word on $\{a, b, c\}$ avoiding it. This infinite word $\mathbf{W}$ must be squarefree. By Lemma 3.2.2, there exists a word U such that xU, yU, and Uz are all subwords of $\mathbf{W}$, where x, y, and z are the letters a, b, and c in some order. We define φ by $\varphi(a) = U$, $\varphi(b) = z$, $\varphi(c) = x$, and $\varphi(d) = y$. One can check that $\varphi(ab)$, $\varphi(ca)$, and $\varphi(da)$ are all subwords. Since $\mathbf{W}$ contains all six transitions, it contains $\varphi(cb)$, $\varphi(dc)$, $\varphi(bc)$, and $\varphi(bd)$, too.

To prove the result for the remaining MLF's, it is sufficient to find a letter e in each formula such that e_L has one neighbor and e_R has two neighbors or such that e_L has two neighbors and e_R has one neighbor. In the former case, we apply Lemma 3.2.2; in the latter case, we apply 3.2.3. We then proceed as in the above paragraph.

For $ab.cb.cd.bd.ba.da.dc$, $ab.cb.cd.ca.da.dc.bc$, and $ab.cb.ca.da.cd.bd.bc$, a has this property. For $ab.cb.cd.ca.ba.da.dc$, b does. □

3.3 Growth Rates

In the previous section, we showed that every minimally locked formula on four letters has index 4. In this section, we measure to what extent they are 4-avoidable.

Let f be a formula, and let $T_n(f, m)$ be the number of words having length n over an alphabet of size m avoiding f. If f and m are understood, we simply write T_n. If $m < \operatorname{ind}(f)$, then $T_n = 0$ for sufficiently large n. For many avoidable formulas, if $m \geq \operatorname{ind}(f)$, T_n grows exponentially, meaning that for all n we have $C\alpha^n < T_n$ for some $C > 0$ and $\alpha > 1$. In this case, we think of f as *easily* m-avoidable. On the other hand, again assuming $m \geq \text{ind}(f)$, if T_n grows polynomially, meaning for all $T_n < Cn^\alpha$ for some $C > 0$ and $\alpha > 0$, we think of f as *barely* m-avoidable. In the 1989 paper by Baker, McNulty, and Taylor [1], it was shown that *ab.ba.ac.ca.bc* is barely 4-avoidable. (Actually, it was shown that *abwbcxcaybazac* is barely 4-avoidable, but this result implies that *ab.ba.ac.ca.bc* is barely 4-avoidable.) We prove this result, and then show that some of the MLF's on four letters are barely 4-avoidable. We conjecture that every MLF on four letters is barely 4-avoidable

Proposition 3.3.1. *ab.ba.ac.ca.bc is barely 4-avoidable.*

The proof of the proposition relies on several simple ideas and lemmas. Let $\mathbf{W}$ be an infinite word on $\{a, b, c, d\}$. Say that $\mathbf{W}$ is "red" if it contains the transitions *ab*, *ba*, *cd*, and *dc*. Say that $\mathbf{W}$ is "yellow" if it contains the transitions *ac*, *ca*, *bd*, and *db*. Finally, say that $\mathbf{W}$ is "blue" if it contains the transitions *ad*, *da*, *bc*, and *cb*.

Lemma 3.3.2. *An infinite word on* $\{a, b, c, d\}$ *avoiding* $f = ab.ba.ac.ca.bc$ *must be exactly two of the three colors.*

Proof. If an infinite word has all three colors, then it does not avoid f. Indeed, *ab*, *ba*, *ac*, *ca*, and *bc* are all subwords of it. Hence, it is sufficient to show that the word must contain two colors. This fact is easily verified by computer analysis. □

Lemma 3.3.3. *An infinite word* $\mathbf{W}$ *on* $\{a, b, c, d\}$ *avoiding* $f = ab.ba.ac.ca.bc$ *must have only the transitions of the two colors.*

Proof. If $\mathbf{W}$ contains an extra transition other than those of the two colors, it is easy to find an encounter of f. For example, suppose $\mathbf{W}$ is red and blue and also contains *ac*. One can see that $f \mid \mathbf{W}$ via $\varphi(a) = b$, $\varphi(b) = a$, and $\varphi(c) = c$. □

Lemma 3.3.4. *Let* $\mathbf{W}$ *be an infinite word over* $\{\, a, b, c, d \,\}$ *having only transitions from two of the three colors. Then* $\mathbf{W} = h(\mathbf{W_0})$ *for some infinite word* $\mathbf{W_0}$ *over* $\{\, a, b, c, d \,\}$ *and some endomorphism* h *from the following six:*

1. $ac/ad/bc/bd$

2. $ca/cb/da/db$

3. $ab/ad/cb/cd$

4. $ba/bc/da/dc$

5. $ab/ac/db/dc$

6. $ba/bd/ca/cd$

Proof. Suppose $\mathbf{W}$ is red and yellow. The transitions of $\mathbf{W}$ are ab, ba, cd, dc, ac, ca, bd, and db. If the first letter of $\mathbf{W}$ is a or d, then $\mathbf{W}$ is an infinite product of the words ab, ac, db, and dc. This means that $\mathbf{W} = h(\mathbf{W_0})$ for some $\mathbf{W_0}$, where h is the fifth endomorphism on the list. If the first letter of $\mathbf{W}$ is b or c, then $\mathbf{W}$ is an infinite product of the words ba, bd, ca, and cd, which means that $\mathbf{W} = h(\mathbf{W_0})$ for some W_0, where this time h is the sixth endomorphism on the list. The cases where $\mathbf{W}$ is red and blue or blue and yellow are similar. □

We now prove Proposition 3.3.1.

Proof. By the Lemma 3.3.3, there exists a k such that every word over $\{\, a, b, c, d \,\}$ with length k avoiding $ab.ba.ac.ca.bc$ has only the transitions of exactly two colors. Let $M = \max\{\, T_1, T_2, \ldots, T_k \,\}$. By induction on n, we show that if $2^{n-1}k + 1 \leq j \leq 2^n k$, then $T_j \leq 12^n M$.

First, suppose that $k+1 \leq j \leq 2k$, and let W be a word of length j. By Lemma 3.3.4, W can be written as $h(W')Y$, where h is one the six endomorphisms, W' is some word on $\{a, b, c, d\}$ having at most length k avoiding f, and $|Y| \leq 1$. There are at most M possible words W', at most six choices for h, and at most two choices for Y, yielding at most $12M$ possibilities for W.

Assume that if $2^{n-1}k+1 \leq j \leq 2^n k$, then $T_j \leq 12^n M$, and suppose $2^n k+1 \leq j \leq 2^{n+1}k$. If W is a word of length j, we can write $W = h(W')Y$, with W' and Y as in the previous paragraph. Counting, we have that there are at most $12^n M$ choices for W', at most six choices for h, and at most two choices for Y, implying at most $12^{n+1}M$ choices for W. The induction step is proven.

We conclude the proof. Suppose j satisfies $2^{n-1}k+1 \leq j \leq 2^n k$. Then,

$$T_j \leq 12^n M \leq 2^{4n} M < (\frac{16M}{k^4}) j^4.$$

Thus, T_j has a quartic bound. □

If f is barely m-avoidable, $g \,|\, f$, and g is m-avoidable, then g is barely m-avoidable since $T_n(g, m) \leq T_n(f, m)$. If f is barely m-avoidable and $\tilde{f}$ is the reversal of f, then $\tilde{f}$ is barely m-avoidable, since $T_n(\tilde{f}, m) = T_n(f, m)$.

Corollary 3.3.5. *The minimally locked formulas*

$$ab.cb.cd.bd.ba.da.dc, \quad ab.cb.cd.ca.da.dc.bc, \quad \textit{and} \quad ab.cb.cd.ca.ba.da.dc,$$

and their reversals, are barely 4-avoidable.

Proof. Each of the above formulas divides $ab.ba.ac.ca.bc$:

1. $ab.cb.cd.bd.ba.da.dc \,|\, ab.ba.ac.ca.bc$ via $\varphi(a) = a$, $\varphi(b) = b$, $\varphi(c) = a$, and $\varphi(d) = c$.

2. $ab.cb.cd.ca.da.dc.bc \,|\, ab.ba.ac.ca.bc$ via $\varphi(a) = c$, $\varphi(b) = a$, $\varphi(c) = b$, and $\varphi(d) = a$.

3. $ab.cb.cd.ca.ba.da.dc \,|\, ab.ba.ac.ca.bc$ via $\varphi(a) = c$, $\varphi(b) = a$, $\varphi(c) = b$, and $\varphi(d) = a$.

□

It seems likely that the other minimally locked formulas on four letters—$ab.cb.ca.da.dc.bc.bd$ and $ab.cb.ca.da.cd.bd.bc$ and their reversals—are also barely 4-avoidable, but this conjecture has yet to be proven.

CHAPTER 4

Index 5 Formulas

4.1 Introduction

An open problem of avoidance theory has been the existence of avoidable patterns of index 5 or higher. We show that the formula $\rho = ab.ba.ac.bc.cda.dcd$ (and hence the pattern $abebafacgbchcdaidcd$) has index 5. Afterward, we give other examples of index 5 formulas and offer a conjecture on a class of formulas we believe are index 5.

In discovering an index 5 formula, we made several assumptions. First, we assumed that it would divide some fairly small prefix of $\mathbf{W}$, the fixed point of $\Omega = 01/21/03/23$. (Any index 5 formula must divide $\mathbf{W}$; otherwise it would be 4-avoidable.) As it turns out, ρ divides the prefix of $\mathbf{W}$ of length 25. The second assumption that we made was that an index 5 formula would contain only small fragments, since a formula with large fragments typically has low index. As it can be seen, the largest fragment of ρ has length 3. The third assumption that we made was that a homomorphism showing the division of an index 5 formula into $\mathbf{W}$ would be simple—i.e., if φ shows this division, then $|\varphi(x)|$ would be small for all variables x in the formula. For ρ, one division of it into $\mathbf{W}$ has $|\varphi(x)| \leq 3$ for all $x \in \{ a, b, c, d \}$. Finally, we assumed that an index 5 formula would have only four letters. It is fairly simple to construct an avoidable formula satisfying these assumptions. Choose a prefix P of $\mathbf{W}$, a maximum fragment size m, and a maximum image length n for φ. Choose the number of letters for the formula and let Δ be an alphabet of that size. Let φ be a nonerasing homomorphism from Δ^* to $\{0,1,2,3\}^*$ such that $|\varphi(x)| \leq n$ for all $x \in \Delta$. Let f be the formula $\{U \in \Delta^+ \mid |U| \leq m \text{ and } \varphi(U) \leq P\}$, which clearly satisfies the assumptions. If f is unavoidable, choose a different P, m, n, or φ. If f is avoidable, then it is likely not minimal. However, we can generate the minimal formulas dividing f by taking repeated simplifications. In this fashion, ρ was discovered.

4.2 The Formula ρ Is 4-Unavoidable

The question of the 4-unavoidability of ρ is settled by backtracking. We note one improvement that can be made.

We observe that ρ divides both $xyxy$ (via $\varphi(a) = \varphi(b) = xy$, $\varphi(c) = x$, and $\varphi(d) = y$) and $xxyxx$ (via $\varphi(a) = \varphi(b) = \varphi(d) = x$ and $\varphi(c) = y$). If $\mathbf{W}$ is an infinite word over four letters, say $\{a,b,c,d\}$, which avoids ρ, then $\mathbf{W}$ must also avoid $xyxy$ and $xxyxx$. Avoiding $xyxy$ means that $\mathbf{W}$ cannot have any squares other aa, bb, cc, or dd. Avoiding $xxyxx$ means that $\mathbf{W}$ can have at most one occurrence of any square. Hence, some suffix of $\mathbf{W}$ must not have any squares at all. Therefore, if ρ is 4-avoidable, then there exists an infinite *squarefree* word avoiding it. This observation greatly reduces the number of cases backtracking must go through to prove 4-unavoidability.

4.3 5-Special Formulas

In this section, we look for other 4-unavoidable formulas by mimicking the pattern of ρ. Specifically, there are three observations that can be made about it. First, ρ is not divisible by any locked formula; otherwise, it would be 4-avoidable.

Second, in ρ, if c precedes anything, it precedes d, and if d follows anything, it follows c. Hence, the edge $\{\, c_L, d_R \,\}$ is one component of the adjacency graph. In fact, further inspection of $\mathrm{AG}(\rho)$ shows that $\{\, c \,\}$ and $\{\, d \,\}$ are the only free sets. Finally, the fragments cda and dcd stand out, not only by their length, but also by the role they play in the avoidability of ρ: dcd ensures that deletion of c yields a square, whereas the deletion of d from cda yields ca, which, when combined with the other fragments, produces the locked formula $ab.ba.ac.bc.ca$. We codify these observations in the following definition.

Definition 4.3.1. *Let f be a formula, and let x, y, and z be distinct letters of* $\mathrm{alph}(f)$*. We say that f is* 5-special *if it satisfies the following conditions:*

1. *No locked formula divides f.*
2. *The formula $yxy.xyz$ i-divides f.*
3. *The adjacency graph of f has exactly two components, the first of which consists of only the edge $\{\, x_L, y_R \,\}$ and the second of which is not disconnected by the removal of y_L.*

Examples of 5-special formulas include $eaeb.bcb.da.bdca$ (with $x = a$, $y = e$, and $z = b$), $yxy.wvxyzx.vwx.vz$ (with $x = x$, $y = y$, and $z = z$), and $\rho = ab.ba.ac.bc.cda.dcd$ (with $x = c$, $y = d$, and $z = a$).

A 5-special formula is *minimally 5-special* if no simplification of it is also 5-special. Amongst all 5-special formulas, these will have the highest indices. Like ρ, any minimally 5-special formula has only two fragments of length 3, yxy and xyz; every other fragment has length 2.

Conjecture 4.3.2. *Every minimally* 5*-special formula has index 5.*

$ab.ba.ac.bc.cda.dcd$
$cb.ca.da.dc.bc.bd.aeb.eae$
$cb.cd.bd.ba.da.dc.aeb.eae$
$cb.cd.ca.da.dc.bc.aeb.eae$
$ab.cb.cd.ca.da.dc.bec.ebe$
$cb.cd.ca.ba.da.dc.aeb.eae$
$ab.cb.cd.ca.da.dc.bea.ebe$
$cb.ca.cd.bd.da.dc.aeb.eae$
$ab.cb.ca.cd.da.dc.bed.ebe$
$cb.ca.ba.bd.ed.ec.dc.de.afb.faf$
$cb.ca.da.dc.ec.ed.bd.be.afb.faf$

Table 4.1: Known Index 5 Formulas

In the next section, we will show that any 5-special formula is 5-avoidable. Assuming this result for now, the conjecture follows if we could prove that every minimally 5-special formula is 4-unavoidable. Like ρ, any minimally 5-special formula divides both $xyxy$ and $xxyxx$. Hence, to show the 4-unavoidability, we can insist that any infinite word over four letters avoiding it be squarefree.

Table 4.1 lists confirmed index 5 formulas.

4.4 5-Special Formulas Are 5-Avoidable

In this section, we show that every 5-special formula is avoided by the D0L-system $\Psi = 01/02/3204/31/3234$, and hence every 5-special formula is 5-avoidable. In what follows, we will mention some common characteristics that Ψ has with the D0L-system $\Omega = 01/21/03/23$ discussed in Chapter 1.

Let $\mathbf{R} = \Psi^\omega(0) = 01020132040102\ldots.$

Theorem 4.4.1. $\mathbf{R}$ *avoids every* 5*-special formula.*

Corollary 4.4.2. *The formula* ρ *is 5-avoidable.*

The proof of Theorem 4.4.1 relies on several lemmas.

Lemma 4.4.3. 0*'s and* 3*'s occupy the odd index positions and* 1*'s,* 2*'s, and* 4*'s occupy the even index positions of* $\mathbf{R}$.

This is the first similarity between Ψ and Ω. In $\Omega^\omega(0)$, 0's and 2's are in the odd index positions and 1's and 3's are in the even positions. In $\Psi^\omega(0)$, there is a similar partitioning of the letters. Because of this partitioning, it will be convenient to define $C_1 = \{\,0, 3\,\}$ and $C_2 = \{\,1, 2, 4\,\}$.

Lemma 4.4.4. *Suppose* W_1W_2*,* W_3W_2*,* W_3W_4*,* $\ldots$*,* $W_{2n-1}W_{2n}$ *are all subwords of* $\mathbf{R}$*. Then, for some permutation* π *of* $\{\,1, 2\,\}$*,* W_k *ends with an element of* $C_{\pi(1)}$ *for all odd* k *and* W_k *begins with an element of* $C_{\pi(2)}$ *for all even* k*. In particular, if* $W_{2n} = W_1$*, then* $|W_1|$ *is even.*

Proof. This follows from Lemma 4.4.3. One should compare this result to Corollary 1.7.4. □

Definition 4.4.5. *Let* W *be a subword of* $\mathbf{R}$*. Suppose* $W \neq 0, 3$*.*

1. *The* left signature *(resp.* right signature*) of* W*, written as* $\mathrm{LSig}(W)$ *(resp.* $\mathrm{RSig}(W)$*), is the furthest left (resp. right) letter of* W *which is an element of* C_2*.*

2. *The* signature *of* W*, denoted by* $\mathrm{Sig}(W)$*, is the pair* $(\mathrm{LSig}(W), \mathrm{RSig}(W))$*.*

If W *is* 0 *or* 3*, the signature of* W *is not defined.*

For example, $\mathrm{LSig}(0132) = 1$, $\mathrm{RSig}(0132) = 2$, and $\mathrm{Sig}(0132) = (1, 2)$. We have $\mathrm{Sig}(4) = (4, 4)$, but $\mathrm{Sig}(3)$ is not defined.

The following lemma is the analog of the alternation of 0's and 2's in the odd index positions of $\Omega^\omega(0)$.

Lemma 4.4.6. *Let XYZ be a subword of $\mathbf{R}$, and suppose that both $\mathrm{Sig}(X)$ and $\mathrm{Sig}(Z)$ are defined and that Y is one of ε, 0, or 3.*

1. *If $\mathrm{RSig}(X) = 1$, then $\mathrm{LSig}(Z) = 2$.*

2. *If $\mathrm{RSig}(X) = 4$, then $\mathrm{LSig}(Z) = 1$.*

3. *If $\mathrm{RSig}(X) = 2$, then $\mathrm{LSig}(Z)$ is either 1 or 4. Specifically, if X ends 02, 020, or 023, then $\mathrm{LSig}(Z) = 1$, and if X ends 32, 320, or 323, then $\mathrm{LSig}(Z) = 4$. If X is 2, 20, or 23, then $\mathrm{LSig}(Z)$ could be either 1 or 4.*

In other words, in the even index positions, 1's are followed by 2's, 2's are followed by either 1's or 4's, depending on what element of C_1 immediately precedes the 2, and 4's are followed by 1's.

Proof. We can write $\mathbf{R} = W_1W_2\dots$, where each $W_i \in \{01, 02, 04, 31, 32, 34\}$. It is sufficient to show that 01 and 31 can only be followed by 02 or 32; that 02 can only be followed by 01 or 31; that 32 can only be followed by 04 or 34; and that 04 or 34 can only be followed by 01 or 31.

01 (resp. 31) pulls back to 0 (resp. 3), which can only be followed by 1, 2, or 4. But $\Psi(1) = 02$, $\Psi(2) = 3204$, and $\Psi(4) = 3234$, so 01 (resp. 31) can only be followed by 02 or 32.

02 pulls back to 1, which can only be followed by 0 or 3. Since $\Psi(0) = 01$ and $\Psi(3) = 31$, 02 is only followed by 01 or 31.

Clearly, 32 must be followed by 04 or 34.

Finally, consider 04 (resp. 34). It must be preceded by 32. Pulling back, we get 2 (resp. 4). This letter must be followed by 0 or 3, and so 04 (resp. 34) must be followed by 01 or 31. □

Reversing the previous lemma yields this corollary.

Corollary 4.4.7. *Suppose XYZ is a subword of $\mathbf{R}$, that $\mathrm{Sig}(X)$ and $\mathrm{Sig}(Z)$ are both defined, and that Y is ε, 0, or 3.*

1. *If $\mathrm{LSig}(Z) = 2$, then $\mathrm{RSig}(X) = 1$.*

2. *If $\mathrm{LSig}(Z) = 4$, then $\mathrm{RSig}(X) = 2$.*

3. *If $\mathrm{LSig}(Z) = 1$, then $\mathrm{RSig}(X)$ is either 2 or 4.*

Lemma 4.4.8. *Suppose $W_1X_1W_2$, $W_3X_2W_2$, $W_3X_3W_4$, $\ldots$, $W_{2n-1}X_{2n-2}W_{2n-2}$, $W_{2n-1}X_{2n-1}W_1$ are all subwords of $\mathbf{R}$; that each X_i is ε, 0, or 3; and that $\mathrm{Sig}(W_j)$ is defined for all j.*

1. *$\mathrm{Sig}(W_1)$ must be one of the following: $(1,2)$, $(1,4)$, $(2,1)$, $(4,2)$, or $(4,4)$.*

2. *If, in addition to the above, no W_i is 2, 20, or 23 for i odd, then $\mathrm{Sig}(W_1)$ cannot be $(4,4)$.*

Proof. Set $W_{2n} = W_1$. We use Lemma 4.4.6 and Corollary 4.4.7.

(1) Suppose $\mathrm{RSig}(W_1) = 1$. Then $\mathrm{LSig}(W_2) = 2$, which implies $\mathrm{RSig}(W_3) = 1$. Continuing in this way, we have $\mathrm{LSig}(W_k) = 2$ for k even and $\mathrm{RSig}(W_k) = 1$ for k odd. In particular, $\mathrm{LSig}(W_{2n}) = 2$. Hence, $\mathrm{Sig}(W_1) = (2,1)$.

If $\mathrm{RSig}(W_1) = 2$, then $\mathrm{LSig}(W_2)$ is either 1 or 4. This implies that $\mathrm{RSig}(W_3)$ is either 2 or 4, which implies $\mathrm{LSig}(W_4)$ is either 1 or 4. Continuing, we can eventually conclude that $\mathrm{LSig}(W_{2n})$ is either 1 or 4. Hence, $\mathrm{Sig}(W_1)$ is either $(1, 2)$ or $(4, 2)$.

Finally, suppose $\mathrm{RSig}(W_1) = 4$. Then, $\mathrm{LSig}(W_2) = 1$, and so $\mathrm{RSig}(W_3)$ is either 2 or 4, which implies $\mathrm{LSig}(W_4)$ is either 1 or 4. Continuing, we have that $\mathrm{LSig}(W_{2n})$ is either 1 or 4. Hence, $\mathrm{Sig}(W_1)$ is either $(1, 4)$ or $(4, 4)$.

These three cases prove (1).

(2) Suppose $\mathrm{RSig}(W_1) = 4$, and so $\mathrm{LSig}(W_2) = 1$. Consider W_3. By Lemma 4.4.6, W_3 determines the left signature of any word that follows it, if that word has a signature at all. Since $\mathrm{LSig}(W_2) = 1$, it must be that $\mathrm{LSig}(W_4) = 1$ as well. By induction, we can show $\mathrm{LSig}(W_k) = 1$ for all even k. In particular, $\mathrm{LSig}(W_{2n}) = 1$. □

Corollary 4.4.9. *Suppose that $W_3X_2W_2$, $W_3X_3W_4$, $\ldots$, $W_{2n-1}X_{2n-2}W_{2n-2}$, $W_{2n-1}X_{2n-1}W_{2n}$ are all subwords of* **R***, that each X_i is ε, 0, or 3, and that* $\mathrm{Sig}(W_j)$ *is defined for all j. Moreover, suppose no W_i is* 2, 20, *or* 32 *for i odd. Then the left signatures of W_i, i even, are all the same.*

Proof. This follows from the proof of Lemma 4.4.8 (2). Note that word $W_1X_1W_2$ is omitted from the sequence of words since its presence is not need to prove the result. □

Proposition 4.4.10. **R** *avoids every locked formula.*

Proof. It is sufficient to prove that **R** avoids every *minimally* locked formula f. Suppose $f \mid \Psi^m(0)$, with m minimal, and let φ show the divisibility. By Lemma 4.4.4, there are two possibilities: either $\varphi(a)$ begins with an element of C_1 and ends with an element of C_2 for every $a \in \mathrm{alph}(f)$, or vice versa. In either case, $|\varphi(a)| \geq 2$ for all $a \in \mathrm{alph}(f)$; thus, the signature of $\varphi(a)$ is defined for all

$a \in \mathrm{alph}(f)$. Since f is locked, the hypotheses of Lemma 4.4.8 are satisfied, so $\varphi(a)$ cannot be 32, 20, nor 23 for any $a \in \mathrm{alph}(f)$. (Otherwise, $\varphi(a)$ would have $(2, 2)$ as a signature.)

First, consider the case where $\varphi(a)$ begins with an element of C_1 and ends with and element of C_2 for all $a \in \mathrm{alph}(f)$. Let $L = \{\, 01, 02, 3204, 31, 3234 \,\}$, and consider the prefix slide setup $(L, \Psi^m(0), f, \varphi)$. If $\varphi(a)$ begins with 04 or 34, set $P'_a = 32$; otherwise, set $P'_a = \epsilon$. It is clear that prefix slide condition holds. By Proposition 1.6.3, we have a homomorphism φ' such that $\varphi'(Q)$ is a subword of $\Psi^\omega(0)$ for all $Q \in f$ and $\varphi'(a) \in L^*$ for all $a \in \mathrm{alph}(f)$. Since $\varphi(a) \neq 32$, one can check that $\varphi'(a) \neq \varepsilon$ for any $a \in \mathrm{alph}(f)$. This implies that $f \mid \Psi^{m-1}(0)$.

Now suppose $\varphi(a)$ begins with an element of C_2 and ends with an element of C_1 for all $a \in \mathrm{alph}(f)$. Let $L' = \{\, 01, 02, 04, 31, 32, 34 \,\}$. This time, we use the suffix version of Proposition 1.6.3. Since $\varphi(a)$ cannot be 20 or 23, by Lemma 4.4.6 there exists a unique element of C_2, say S'_a, such that $\varphi(a)S'_a$ is a subword of $\Psi^m(0)$. The suffix slide condition holds for the suffix slide setup $(L', \Psi^m(0), f, \varphi)$. In this case, $\varphi'(a)$ is the word $\varphi(a)S'_a$ minus its initial letter. Since $|\varphi'(a)| = |\varphi(a)| \neq \varepsilon$, φ' shows $f \mid \Psi^m(0)$. We can now apply the first case. □

Lemma 4.4.11. *If UVU is a subword of $\mathbf{R}$, where $U, V \in \{\, 0, 1, 2, 3, 4 \,\}^+$ and $|V| = 1$, then V is one of* 1, 2, *or* 4 *and U is either* 0 *or* 3.

Proof. Suppose that $UVU \leq \mathbf{R}$ with $|V| = 1$. We observe that $|UV|$ is even by Lemma 4.4.3. Hence, $|U|$ is odd.

Suppose that V is 1, 2, or 4 and that $|U| \geq 3$. By Lemma 4.4.6, U determines the left signature of what follows it, which is V. This means that UVU is followed by V. However, $\mathbf{R}$ is squarefree by Proposition 4.4.10. This is a contradiction.

Suppose $V = 0$ and that UVU is subword of $\Psi^m(0)$, with m minimal. By

Lemma 4.4.8, the signature of U must be $(1,2)$, $(1,4)$, $(2,1)$, $(4,2)$, or $(4,4)$. However, $(4,4)$ is impossible by Lemma 4.4.6. Hence, $|U| \geq 3$. There are three cases. First, suppose $\mathrm{LSig}(U) = 1$. We can write $U = 1U'$ for some word U'. We have $UVU = 1|U'01|U' = 1|U'|01|U'|$. Pulling back, we get $\Psi^{-1}(U')0\Psi^{-1}(U')$, which contradicts the choice for m. Second, suppose $\mathrm{Sig}(U) = (2,1)$. We can write $U = 2U'1$ for some word U'. Here, we have $UVU = 2U'102U'1 = 2|U'1|02|U'1|$. We see that UVU must be preceded by 0. However, this contradicts Proposition 4.4.10. Finally, suppose $\mathrm{Sig}(U) = (4,2)$. We can write $U = 4U'2$ for some word U'. Substituting, we get $UVU = 4U'204U'2 = 4|U'204|U'2$. We see that U' can be write $U' = U''3$, yielding $4|U''|3204|U''|32$. Ignoring the ending 32, we see that this pulls back to either $2\Psi^{-1}(U'')2\Psi^{-1}(U'')$ or $4\Psi^{-1}(U'')2\Psi^{-1}(U'')$. The former is impossible by Proposition 4.4.10. The latter can only happen if $|\Psi^{-1}(U'')| = 1$. This contradicts Lemma 4.4.6: 4 is followed by 2.

The case of $V = 3$ is similar to that of $V = 0$. □

We now prove Theorem 4.4.1.

Proof. Suppose the 5-special formula f divides $\Psi^m(0)$, with m minimal. Let φ show the divisibility. We begin with some general observations.

If $u \neq x, y$, then u_L and u_R are connected in $AG(f)$. By Lemma 4.4.4, $|\varphi(u)|$ is even. If $u = x$, then $\varphi(u)$ cannot be 0 or 3. Otherwise, $\varphi(yxy)$ would be a subword of $\mathbf{R}$, which violates Lemma 4.4.11. Therefore, $\mathrm{Sig}(\varphi(u))$ must be defined for all $u \neq y$. The signature of $\varphi(y)$ may or may not be defined.

Let $u \neq x, y$. In $AG(f)$, u_L and u_R are connected by a path which does not pass though y_L. By Lemma 4.4.8, we can conclude that $\mathrm{Sig}(\varphi(u)) \neq (2,2)$. In particular, $\varphi(u)$ cannot be 2, 20, 23, or 32.

A 5-special formula has only two free sets, $\{x\}$ and $\{y\}$. Deleting $\{x\}$ yields a square, while deleting $\{ y \}$ yields a locked formula. To prove this last part, let $u \neq y$. If $u \neq x$, there is a path from u_L to u_R in $\mathrm{AG}(f)$ which avoids y_L. Since the same path exists in $\mathrm{AG}(\sigma_y(f))$, u is not free in $\sigma_y(f)$. If $u = x$, then x_L is connected to z_R in $\sigma_y(f)$ since $xyz \,|_i\, f$. Since z_R is connected to x_R by a path which avoids y_L in $\mathrm{AG}(f)$, z_R is connected to x_R in $\mathrm{AG}(\sigma_y(f))$. By joining these two paths, we have that x is not free in $\sigma_y(f)$.

We now are ready to proceed. There are two cases.

Case 1: Suppose $\varphi(y)$ has a defined signature.

We show that $\varphi(x)$ cannot be 2, 20, 23, or 32. If $\varphi(x) = 2$, then $\varphi(yxy) = \varphi(y)2\varphi(y)$. By Lemma 4.4.11, this implies that $\varphi(y)$ is 0 or 3, which contradicts it having a defined signature. If $\varphi(x) = 20$, then $\varphi(yxy) = \varphi(y)20\varphi(y)$. This means $\mathrm{RSig}(\varphi(y)) = 1$, and so $\varphi(y)20\varphi(y)$ must be followed by 2. This violates Lemma 4.4.11. The case with $\varphi(x) = 23$ is similar. Finally, if $\varphi(x) = 32$, then $\mathrm{LSig}(\varphi(y)) = 4$. This means that $\varphi(yxy)$ is preceded by 32, which produces a square. This violates Proposition 4.4.10.

Let L be the suffix code $\{ 01, 02, 04, 31, 32, 34 \}$, and consider the suffix slide setup $(L, \Psi^m(0), f, \varphi)$. For $a \in \mathrm{alph}(f)$, if $\varphi(a)$ ends with an element of C_2, set $S'_a = \varepsilon$. If $\varphi(a)$ ends with an element of C_1 and $a \neq y$, set S'_a to be the unique element of C_2 such that $\varphi(a)S'_a$ is a subword of $\Psi^m(0)$ as determined by Lemma 4.4.6. Finally, if $\varphi(y)$ ends with an element of C_1, set S'_y to be the unique element of C_2 such that $\varphi(x)\varphi(y)S'_y$ is a subword of $\Psi^m(0)$ as determined by Lemma 4.4.6.

We need to verify the suffix slide condition. It can only fail for y and only if y is the first letter of some fragment. Suppose $\varphi(y)$ ends with an element of C_1. Since $\varphi(xyz)$ is a subword, we have S'_y is the first letter of $\varphi(z)$. We show that if yu is a transition of f, for some $u \in \mathrm{alph}(f)$, then $\varphi(u)$ also begins with S'_y. The vertices z_R and u_R lie in the same component in the adjacency graph of f.

Moreover, there exists a path between z_R and u_R which avoids y_L. By Corollary 4.4.9, $\varphi(z_R)$ and $\varphi(u_R)$ begin with the same letter. The suffix slide condition is verified.

By Proposition 1.6.3, we have $\varphi'(Q)$ is a subword of $\Psi^m(0)$ for all $Q \in f$. If $\varphi'(u) = \varepsilon$, then $\varphi(u)$ must have been 1, 2, or 4. This could have only occurred if $u = y$. However, $\varphi'(y) = \varepsilon$ means $\sigma_y(f) \mid \Psi^m(0)$. This violates Proposition 4.4.10. Therefore, φ' shows $f \mid \Psi^m(0)$. Moreover, $\varphi'(a)$ begins with an element of C_1 and ends with an element of C_2.

Now, we take $L' = \{ 01, 02, 3204, 31, 3234 \}$, and consider the prefix slide setup $(L', \Psi^m(0), f, \varphi')$. If $\varphi'(a)$ begins with 04 or 34, set $P'_a = 32$. Otherwise, set $P'_a = \varepsilon$. One checks that the prefix slide condition holds. Hence, by Proposition 1.6.3 we have φ'' such that $\varphi''(Q)$ is a subword of $\Psi^m(0)$ for all $Q \in f$. If $\varphi''(u) = \varepsilon$, then $\varphi'(u) = 32$, which can only happen if $u = y$. If this is the case, $\sigma_y(f) \mid \Psi^m(0)$, a violation of Proposition 4.4.10. Therefore, $f \mid \Psi^{m-1}(0)$. This contradicts the minimality of m.

Case 2: $\varphi(y)$ is 0 or 3.

Subcase 1: $|\varphi(x)| = 1$. Since $\sigma_y(f)$ is locked, there is a path which connects x_L to x_R in $AG(\sigma_y(f))$. This corresponds to a sequence xu_1, u_2u_1, u_2u_3, $\ldots$, $u_{2n}x$ of transitions of $\sigma_y(f)$, with each $u_i \neq x$. This means xV_1u_1, $u_2V_2u_1$, $u_2V_3u_3$, $\ldots$, $u_{2n}V_{2n+1}x$ is a sequence of subwords of fragments of f, where each V_i is either ε or y. If we apply φ to each subword, we have a sequence of subwords of $\Psi^m(0)$. By Lemma 4.4.8, we see that the signature of x can only be $(1, 2)$, $(1, 4)$, $(2, 1)$, or $(4, 2)$. This contradicts the assumption that $\varphi(x)$ has length 1.

Subcase 2: $|\varphi(x)| > 1$. In fact, since $\varphi(yxy)$ is a subword of $\mathbf{R}$, Lemma 4.4.3 implies that $|\varphi(yx)|$ is even. Therefore $|\varphi(x)| \geq 3$. In particular, $\varphi(x)$ cannot be 2, 20, 23, or 32. Now proceed as in Case 1, skipping the first paragraph. □

4.5 Index 6?

Having found an index 5 formula, we are tempted to conjecture that formulas for all indices exist. However, despite our best efforts, we have yet to find a formula having an index higher than 5.

It seems there are two courses one can follow to find an index 6 formula. One plan, if you are optimistic, is to assume that such a formula can be found over five letters. If this is the case, one can construct a 5-avoidance basis. After removing the formulas that are locked or 5-special, one should get a reasonably sized list (less than a million?). Since it seems likely that an index 6 formula has fragments of small lengths and that its adjacency graph has few components (but at least two), one can pare this list of candidates down further.

Another plan would be to mimic the discovery process of ρ. If a formula is index 6, it must divide $\mathbf{R}$. For small values of m, n and k, one could easily generate words W over n letters having length at most m which divide $\mathbf{R}$, or at least some fixed prefix of it, via φ, with $|\varphi(x)| \leq k$ for all letters x. Construct the set S of all such ordered pairs (W, φ). Then, for every possible φ, form the formula $\{\, W \mid (W, \varphi) \in S \,\}$. This formula is probably not minimal, but by considering simplifications, one can eventually generate a list of candidates.

The best option, of course, would be a combination of both these two plans. However, lurking in the background is the combinatorial explosion encountered in verifying a formula over five letters is 5-unavoidable. In the absence of hand proofs of unavoidability, the discovery of index 6, if it exists, may be computationally out of reach.

APPENDIX A

Cubefree Words on Two Letters

A.1 Introduction

One of the oldest known results in the study of combinatorics on words is the avoidability of aaa over two letters. Thue [14] showed, for example, that aaa was avoided by $\mu = 01/10$, the endomorphism which generates the Thue-Morse sequence. In this chapter, we study two areas. First, what can be said about the density of the letters in any cubefree infinite word? The Thue-Morse sequence is an example of a word having half 0's and half 1's. Second, what can be said about the growth rate of cubefree words? Using the vocabulary of Chapter 3, is aaa easily or barely 2-avoidable?

For any word U over alphabet Σ, let the *count* of letter i in U, denoted $\mathrm{ct}_i(U)$, be the number of i's in U. For example, if $U = 0100110$, $\mathrm{ct}_0(U) = 4$ and $\mathrm{ct}_1(U) = 3$. Clearly, $|U| = \Sigma_{i\in\Sigma} \mathrm{ct}_i(U)$.

Let CF be the set of all infinite cubefree words over $\{0, 1\}$. For $\mathbf{W} \in CF$, let $\mathbf{W}[n]$ be the prefix of $\mathbf{W}$ of length n.

Definition A.1.1. *Let* $\mathbf{W} \in CF$*.*

1. *Let* $i \in \{0, 1\}$*. The* i-density *of* $\mathbf{W}$*, denoted by* $\mathrm{den}_i(\mathbf{W})$*, is*

$$\mathrm{den}_i(\mathbf{W}) = \liminf_{n\to\infty} \frac{\mathrm{ct}_i(\mathbf{W}[n])}{n}$$

2. The full density *of* $\mathbf{W}$*, denoted* $\mathrm{fullden}(\mathbf{W})$*, is the ordered pair*

$$\mathrm{fullden}(\mathbf{W}) = (\mathrm{den}_0(\mathbf{W}), \mathrm{den}_1(\mathbf{W})).$$

Clearly, for any $\mathbf{W} \in CF$, we have $\mathrm{den}_0(\mathbf{W}) + \mathrm{den}_1(\mathbf{W}) \leq 1$. Let CF_Δ be the set of $\mathbf{W} \in CF$ for which equality holds.

Proposition A.1.2. *Let* $\mathbf{W} \in CF$*. The following are equivalent:*

(1) $\mathbf{W} \in CF_\Delta$.

(2) $\lim \frac{\mathrm{ct}_0(\mathbf{W}[n])}{n}$ *exists*

(3) $\lim \frac{\mathrm{ct}_1(\mathbf{W}[n])}{n}$ *exists*

For example, let $\mathbf{W} = \mu^\omega(0) = 01101001\ldots$. For all n, $\mathrm{ct}_0(\mathbf{W}[n]) = \lfloor \frac{n}{2} \rfloor + m_n$, where m_n is either 0 or 1. Hence, $\mathrm{den}_0(\mathbf{W}) = \frac{1}{2}$. Similarly, $\mathrm{den}_1(\mathbf{W}) = \frac{1}{2}$. Thus, $\mathbf{W} \in CF_\Delta$.

A.2 Fixed Points of Endomorphisms

A homomorphism is said to be *cubefree* if the image of any cubefree word is itself cubefree. In this section, we consider those $\mathbf{W} \in CF$ that arise as fixed points of some nonerasing cubefree endomorphism. We show that any such $\mathbf{W}$ lie in CF_Δ.

The following theorem, due to Bean, Ehrenfeucht, and McNulty [2], gives a sufficient condition for a nonerasing homomorphism to be cubefree.

Theorem A.2.1. *Let* Σ *be an alphabet. If* h *is a nonerasing homomorphism from* Σ^* *to* $\{0, 1\}^*$ *such that:*

1. $h(U)$ *is cubefree whenever* U *is a cubefree word of length at most 4.*

2. *If $a, b \in \Sigma$ and $h(a) \leq h(b)$, then $a = b$.*

3. *If $a, b, c \in \Sigma$ and $Xh(a)Y = h(b)h(c)$, where $X, Y \in \{0,1\}^*$, then either $X = \varepsilon$ and $a = b$ or $Y = \varepsilon$ and $a = c$.*

Then, h is cubefree.

The two permutations of the alphabet $\{0,1\}$ yield trivial cubefree endomorphisms. One can verify that 001/011 yields a homomorphism that satisfies the hypotheses of Theorem A.2.1. We note that 01/10, the Thue-Morse endomorphism, does not, although it is cubefree.

Let h be any endomophism on $\{0,1\}^*$. The *content matrix* of h, denoted C_h, is the matrix

$$C_h = \begin{pmatrix} \mathrm{ct}_0(h(0)) & \mathrm{ct}_0(h(1)) \\ \mathrm{ct}_1(h(0)) & \mathrm{ct}_1(h(1)) \end{pmatrix}$$

For any word U, $\mathrm{ct}_i(h(U)) = \mathrm{ct}_i(h(0))ct_0(U) + \mathrm{ct}_i(h(1))ct_1(U)$. It is easily shown by induction that $C_h^k \left(\begin{smallmatrix} \mathrm{ct}_0(U) \\ \mathrm{ct}_1(U) \end{smallmatrix} \right) = \left(\begin{smallmatrix} \mathrm{ct}_0(h^k(U)) \\ \mathrm{ct}_1(h^k(U)) \end{smallmatrix} \right)$.

Definition A.2.2. *An endomorpism h of $\{0,1\}^*$ is* growing *provided:*

1. *$h(0) = 0X$, for some nonempty word X.*

2. *$|h(1)| \geq 1$.*

If h is growing, then h is prefix-preserving with respect to 0 and $|h^n(0)| \to \infty$. Thus, h generates an infinite word $\mathbf{W_h} = h^\omega(0)$. Karhumäki [11] refers to growing endomorphisms as pp-morphisms.

For the remainder of this section, h will be a growing endomorphism with $C_h = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix} \right)$. The eigenvalues of C_h are $\lambda_1 = \frac{a+d+\sqrt{(a-d)^2+4bc}}{2}$ and $\lambda_2 = \frac{a+d-\sqrt{(a-d)^2+4bc}}{2}$.

Lemma A.2.3. *If $\mathbf{W_h}$ is cubefree, then $\lambda_1 \neq \lambda_2$.*

Proof. It is sufficient to show that $(a-d)^2+4bc>0$. If the result fails, we must have $a=d$ and either $b=0$ or $c=0$.

If $c=0$, then $h(0)=00\ldots0$, which implies $h^2(0)$ contains 000, and so $\mathbf{W_h}$ is not cubefree. Thus, $h(0)$ contains a 1.

If $b=0$ and $d>1$, then $h^2(0)$ must contain 11, and so $h^3(0)$ must contain 111, which violates cubefreeness.

Therefore, we must have $b=0$ and $a=d=1$. This means $h(0)$ begins 01, $h^2(0)$ begins 011, and $h^3(0)$ begins 0111. Again, this produces a cube. □

Lemma A.2.4. *If* $\mathbf{W_h}$ *is cubefree, then* $\lim_{k\to\infty}\frac{\mathrm{ct}_0(h^k(0))}{|h^k(0))|}=\lim_{k\to\infty}\frac{\mathrm{ct}_0(h^k(1))}{|h^k(1))|}=\frac{c}{c-a+\lambda_1}$.

Proof. We have $|\lambda_2|<\lambda_1$ since $a+d>0$. For each λ_i, associate the eigenvector $\left(\begin{smallmatrix} c \\ \lambda_i-a\end{smallmatrix}\right)$. By Lemma A.2.3, C_h is diagonalizable. Therefore,

$$\begin{aligned} C_h^k &= \begin{pmatrix} c & c \\ \lambda_1-a & \lambda_2-a \end{pmatrix}\begin{pmatrix} \lambda_1^k & 0 \\ 0 & \lambda_2^k \end{pmatrix}\begin{pmatrix} \frac{\lambda_2-a}{c(\lambda_2-\lambda_1)} & \frac{-1}{\lambda_2-\lambda_1} \\ \frac{a-\lambda_1}{c(\lambda_2-\lambda_1)} & \frac{1}{\lambda_2-\lambda_1} \end{pmatrix} \\ &= \begin{pmatrix} \frac{\lambda_1^k(\lambda_2-a)-\lambda_2^k(\lambda_1-a)}{\lambda_2-\lambda_1} & \frac{-c\lambda_1^k+c\lambda_2^k}{\lambda_2-\lambda_1} \\ \frac{(\lambda_1-a)(\lambda_2-a)(\lambda_1^k-\lambda_2^k)}{\lambda_2-\lambda_1} & \frac{-(\lambda_1-a)\lambda_1^k+(\lambda_2-a)\lambda_2^k}{\lambda_2-\lambda_1} \end{pmatrix} \end{aligned}$$

This matrix equation gives us these explicit formulas:

$$C_h^k\begin{pmatrix}1\\0\end{pmatrix}=\begin{pmatrix}\mathrm{ct}_0(h^k(0))\\ \mathrm{ct}_1(h^k(0))\end{pmatrix}=\begin{pmatrix}\frac{\lambda_1^k(\lambda_2-a)-\lambda_2^k(\lambda_1-a)}{\lambda_2-\lambda_1}\\ \frac{(\lambda_1-a)(\lambda_2-a)(\lambda_1^k-\lambda_2^k)}{\lambda_2-\lambda_1}\end{pmatrix}$$

$$C_h^k\begin{pmatrix}0\\1\end{pmatrix}=\begin{pmatrix}\mathrm{ct}_0(h^k(1))\\ \mathrm{ct}_1(h^k(1))\end{pmatrix}=\begin{pmatrix}\frac{-c\lambda_1^k+c\lambda_2^k}{\lambda_2-\lambda_1}\\ \frac{-(\lambda_1-a)\lambda_1^k+(\lambda_2-a)\lambda_2^k}{\lambda_2-\lambda_1}\end{pmatrix}$$

We set $\eta = \frac{\lambda_2}{\lambda_1}$ and recall that $|h^k(i)| = \mathrm{ct}_0(h^k(i)) + \mathrm{ct}_1(h^k(i))$ for $i = 0, 1$. We have

$$\begin{aligned}\frac{\mathrm{ct}_0(h^k(0))}{|h^k(0)|} &= \frac{c(\lambda_2 - a) - c(\lambda_1 - a)\eta^k}{c(\lambda_2 - a) - c(\lambda_1 - a)\eta^k + (\lambda_1 - a)(\lambda_2 - a) - (\lambda_2 - a)(\lambda_1 - a)\eta^k} \\ &\to \frac{c}{c + \lambda_1 - a}\end{aligned}$$

and

$$\begin{aligned}\frac{\mathrm{ct}_0(h^k(1))}{|h^k(1)|} &= \frac{c^2\eta^k - c^2}{c^2\eta^k - c^2 + c(\lambda_2 - a)\eta^k - c(\lambda_1 - a)} \to \frac{-c^2}{-c^2 - c(\lambda_1 - a)} \\ &= \frac{c}{c + \lambda_1 - a}\end{aligned}$$

□

Theorem A.2.5. *If* $\mathbf{W_h}$ *is cubefree, then* $\lim \frac{\mathrm{ct}_0(\mathbf{W_h}[n])}{n} = \frac{c}{c-a+\lambda_1}$. *Hence,* $\mathbf{W_h} \in CF_\Delta$.

Proof. Choose $\epsilon > 0$. By Lemma A.2.4, there exists k such that both

$$\left| \frac{\mathrm{ct}_0(h^k(0))}{|h^k(0)|} - \frac{c}{c - a + \lambda_1} \right| < \epsilon$$

and

$$\left| \frac{\mathrm{ct}_0(h^k(1))}{|h^k(1)|} - \frac{c}{c - a + \lambda_1} \right| < \epsilon.$$

For convenience, let $g = h^k$, and set $\alpha = \max(|g(0)|, |g(1)|)$.

We note that $g(\mathbf{W_h}) = \mathbf{W_h}$. For all n, there exists j such that $g(\mathbf{W_h}[j])$ is a prefix of $\mathbf{W_h}[n]$ and $\mathbf{W_h}[n]$ is a prefix of $g(\mathbf{W_h}[j+1])$. The difference between the lengths of $g(\mathbf{W_h}[j+1])$ and $g(\mathbf{W_h}[j])$ is at most α. Therefore, $\frac{|g(\mathbf{W_h}[j])|}{n} \to 1$ and $\frac{|g(\mathbf{W_h}[j+1])|}{n} \to 1$ as $n \to \infty$.

Let $n > 0$. We have

$$\frac{\mathrm{ct}_0(g(\mathbf{W_h}[j]))}{n} \leq \frac{\mathrm{ct}_0(\mathbf{W_h}[n])}{n} \leq \frac{\mathrm{ct}_0(g(\mathbf{W_h}[j+1]))}{n}$$

$$\frac{|g(\mathbf{W_h}[j])|}{n} \cdot \frac{\mathrm{ct}_0(g(\mathbf{W_h}[j]))}{|g(\mathbf{W_h}[j])|} \leq \frac{\mathrm{ct}_0(\mathbf{W_h}[n])}{n} \leq \frac{|g(\mathbf{W_h}[j+1])|}{n} \cdot \frac{\mathrm{ct}_0(g(\mathbf{W_h}[j+1]))}{|g(\mathbf{W_h}[j+1])|} \tag{A.1}$$

and so we have

$$\frac{|g(\mathbf{W_h}[j])|}{n} \cdot (\frac{c}{c-a+\lambda_1} - \epsilon) \leq \frac{\mathrm{ct}_0(\mathbf{W_h}[n])}{n} \leq \frac{|g(\mathbf{W_h}[j+1])|}{n} \cdot (\frac{c}{c-a+\lambda_1} + \epsilon)$$

Taking the lim inf and lim sup yields

$$\frac{c}{c-a+\lambda_1} - \epsilon \leq \liminf \frac{\mathrm{ct}_0(\mathbf{W_h}[n])}{n} \leq \frac{c}{c-a+\lambda_1} + \epsilon$$

and

$$\frac{c}{c-a+\lambda_1} - \epsilon \leq \limsup \frac{\mathrm{ct}_0(\mathbf{W_h}[n])}{n} \leq \frac{c}{c-a+\lambda_1} + \epsilon$$

Since ϵ is arbitrary, the result is shown. □

A.3 A Bound for the Density Function

Let $\delta = \inf_{\mathbf{W} \in CF} \mathrm{den}_0(\mathbf{W})$ $(= \inf_{\mathbf{W} \in CF} \mathrm{den}_1(\mathbf{W}))$. By symmetry, we have $\delta \leq \mathrm{den}_0(\mathbf{W}) \leq 1 - \delta$ for all $\mathbf{W} \in CF$. Unfortunately, δ is difficult to compute directly. In this section, we bound δ from above and below, yielding an estimate $\delta \approx 0.406$.

Definition A.3.1. *Let CF_n be the set of cubefree words over $\{0, 1\}$ of length n.*

Definition A.3.2. *Let $S(n) = \min\{ \mathrm{ct}_0(U) \mid U \in CF_n \}$. Define $\lambda = \sup_{n \geq 1} \frac{S(n)}{n}$.*

Clearly, $S(n) + S(m) \leq S(n+m)$ for all n.

Lemma A.3.3. *$\frac{S(n)}{n} \to \lambda$.*

Proof. Choose $\epsilon > 0$. There exists N such that $0 \leq \lambda - \frac{S(N)}{N} < \epsilon$. For any n, write $n = Nq + r$, where $0 \leq r < N$. We have

$$\frac{q}{q+1}\frac{S(N)}{N} + \frac{S(r)}{n} \leq \frac{qS(N) + S(r)}{n} \leq \frac{S(n)}{n} \leq \lambda.$$

As $n \to \infty$, we have $\frac{q}{q+1} \to 1$ and $\frac{S(r)}{n} \to 0$. Taking the lim inf, we have

$$\lambda - \epsilon < \frac{S(N)}{N} < \liminf \frac{S(n)}{n} \leq \lambda.$$

Since ϵ was arbitrary, we have $\liminf \frac{S(n)}{n} = \lambda$. This proves the result. □

The next result shows that the supremum of $\frac{S(n)}{n}$ is never attained.

Theorem A.3.4. *$\frac{S(n)}{n} < \lambda$ for all n.*

Proof. Suppose, for some m, $\lambda = \frac{S(m)}{m}$. If $k > 0$, then $S(km) \geq kS(m)$. Dividing by km, we have $\frac{S(km)}{km} \geq \frac{S(m)}{m} = \lambda$. Since $\lambda \geq \frac{S(km)}{km}$, we can conclude that $\frac{S(km)}{km} = \frac{S(m)}{m}$, so $S(km) = kS(m)$.

Let $T = \{\, W \in CF_n \mid \mathrm{ct}_0(W) = S(m) \,\}$. Let $k > 2|T|$, and let U be a cubefree word over $\{\, 0, 1 \,\}$ of length km having $S(km)$ 0's. If we write $U = U_1U_2 \ldots U_k$, with $|U_i| = m$ for all $i \leq k$, we must have $\mathrm{ct}_0(U_i) = S(m)$. Thus, U is a product of words from T. We order the elements of T lexicographically based on $0 \prec 1$. We claim that $U_i \succeq U_{i+1}$ for $i < k$.

Suppose not. Then, for some i, we have $U_i \prec U_{i+1}$. There exist X, Y, and Y' such that $U_i = X0Y$ and $U_{i+1} = X1Y'$. Consider the subword $YX1$ of U_iU_{i+1}, which has length m. We have $\mathrm{ct}_0(YX1) = \mathrm{ct}_0(X0Y) - 1 = S(m) - 1$. This contradicts the definition of $S(m)$, and the claim is proven.

Since U is cubefree, it is not the case that $U_i = U_{i+1} = U_{i+2}$ for any i. In other words, at most two consecutive U_i are the same. This implies that the set $\{\, U_i \mid i \leq k \,\}$ has at least $|T| + 1$ distinct words, which is impossible. □

The next result allows us to bound δ from below.

Theorem A.3.5. *$\lambda \leq \delta$.*

Proof. Choose $\epsilon > 0$. There exists a $\mathbf{W} \in CF$ such that

$$\delta + \epsilon > \mathrm{den}_0(\mathbf{W}).$$

Since

$$\mathrm{den}_0(\mathbf{W}) = \liminf \frac{\mathrm{ct}_0(\mathbf{W}[n])}{n} \geq \lim \frac{S(n)}{n} = \lambda,$$

the result follows. □

With computer analysis, it can be shown that $S(2888) = 1173$. This gives us the lower bound of

$$.4061634 < \frac{1173}{2888} \leq \lambda \leq \delta.$$

On the other hand, the endomorphism h generated by

$$\begin{cases} 0 \mapsto AAFACAFAC \\ 1 \mapsto AAFACAAFC \end{cases} \tag{A.2}$$

where $A = 001101101011011$, $C = 001101101011$, and $F = 001101011011$ is cubefree by Theorem A.2.1. By Theorem A.2.5, $\mathrm{den}_0(\mathbf{W_h}) = \frac{50}{123} < .4065041$. This shows $0.4061634 < \delta < 0.4065041$.

A.4 Full Density

The full density of $\mathbf{W} \in CF$ is the ordered pair $(\mathrm{den}_0(\mathbf{W}), \mathrm{den}_1(\mathbf{W}))$. In this section, we attempt to describe $I = \{\, \mathrm{fullden}(\mathbf{W}) \mid \mathbf{W} \in CF \,\} \subset [0,1]^2$. The results of the previous section show that I is a subset of the triangle bounded by $x = \delta$, $y = \delta$, and $y = 1 - x$.

Lemma A.4.1. *Let Σ_0 and Σ_1 be two disjoint alphabets. Let $\mathbf{W} = a_1a_2a_3\cdots \in CF$. Suppose $\mathbf{W}' = b_1b_2b_3\cdots \in (\Sigma_0 \cup \Sigma_1)^\omega$, with $b_i \in \Sigma_0$ if $a_i = 0$ and $b_i \in \Sigma_1$ if $a_i = 1$. Then, $\mathbf{W}'$ is cubefree.*

Proof. If $\mathbf{W}'$ contains a cube, say

$$b_ib_{i+1}\ldots b_{i+n-1} = b_{i+n}b_{i+n+1}\ldots b_{i+2n-1} = b_{i+2n}b_{i+2n+1}\ldots b_{i+3n-1},$$

then $\mathbf{W}$ contains the cube

$$a_ia_{i+1}\ldots a_{i+n-1} = a_{i+n}a_{i+n+1}\ldots a_{i+2n-1} = a_{i+2n}a_{i+2n+1}\ldots a_{i+3n-1}.$$

□

Theorem A.4.2. *Let $n > 0$. Suppose h is a uniform cubefree homomorphism from $\{0,1,2\}^*$ to $\{0,1\}^*$ with $|h(x)| = n$ for all $x \in \{0,1,2\}$ and*

$$\mathrm{ct}_0(h(0)) \le \mathrm{ct}_0(h(1)) \le \mathrm{ct}_0(h(2)).$$

Set $L = \mathrm{ct}_0(h(0)) + \mathrm{ct}_0(h(1))$ and $H = \mathrm{ct}_0(h(1)) + \mathrm{ct}_0(h(2))$. For all a and b such that

1. $a + b < 1$
2. $a > \frac{L}{2n}$
3. $b < \frac{H}{2n}$

there exists $\mathbf{W} \in CF$ such that $\mathrm{fullden}(\mathbf{W}) = (a, 1-b)$.

Proof. Let a and b satisfy (1), (2), and (3).

Let $\mathbf{M} = \mu^\omega(0) = 01101001\ldots$ be the Thue-Morse sequence. We construct an infinite word $\mathbf{V} = a_1a_2\ldots$ by the following algorithm.

1. We set $F_1 = 0$ and $k = 1$.

2. We define a_k as follows:

   (a) If the kth letter of $\mathbf{M} = 0$, set $a_k = 1$.

   (b) If the kth letter of $\mathbf{M} = 1$ and $F_k = 0$, set $a_k = 0$.

   (c) If the kth letter of $\mathbf{M} = 1$ and $F_k = 1$, set $a_k = 2$.

3. Let $U_k = a_1 a_2 \dots a_k$. Set

$$Q_k = \frac{\mathrm{ct}_0(h(0)) \cdot \mathrm{ct}_0(U_k) + \mathrm{ct}_0(h(1)) \cdot \mathrm{ct}_1(U_k) + \mathrm{ct}_0(h(2)) \cdot \mathrm{ct}_2(U_k)}{nk}.$$

   If $Q_k \geq b$, set $F_{k+1} = 0$. If $Q_k \leq a$, set $F_{k+1} = 1$.

4. Increase k by 1. Go to step 2.

By Lemma A.4.1, $\mathbf{V}$ is cubefree. Since h is a cubefree map, $\mathbf{W} = h(\mathbf{V})$ is cubefree. We claim $\mathrm{fullden}(\mathbf{W}) = (a, 1-b)$

For any $\mathbf{U} \in CF$, we have $\liminf \frac{\mathrm{ct}_1(\mathbf{U}[k])}{k} = 1 - \limsup \frac{\mathrm{ct}_0(\mathbf{U}[k])}{k}$. Therefore, to prove the result, it sufficient to show $\liminf \frac{\mathrm{ct}_0(\mathbf{W}[k])}{k} = a$ and $\limsup \frac{\mathrm{ct}_0(\mathbf{W}[k])}{k} = b$.

Since $\mathbf{M} \in \{01, 10\}^\omega$, we have $\mathbf{V} \in \{01, 10, 12, 21\}^\omega$. This means $\mathbf{W} \in \{h(01), h(10), h(12), h(21)\}^\omega$. We observe that both $h(01)$ and $h(10)$ have L 0's, while both $h(12)$ and $h(21)$ have H 0's. We say that $h(01)$ and $h(10)$ are "low" words and that $h(12)$ and $h(21)$ are "high" words.

We write $\mathbf{W} = W_1 W_2 W_3 \dots$, where each W_i is a low or high word. A simple calculation shows that $Q_{2k} = \frac{\mathrm{ct}_0(W_1 W_2 \dots W_k)}{2kn}$.

The sequence $\{F_k\}$ does not stabilize. Indeed, if $F_k = F_{k+1} = F_{k+2} \cdots = 0$, then W_k, W_{k+1}, W_{k+2}, $\dots$, are all low words, which would imply that $Q_k \to \frac{L}{2n}$. Hence, for some $j > k$, we have $Q_j < a$ and so $F_{j+1} = 1$. A similar argument holds for a sequence of 1's.

We say that position k is a switching point if $F_k = 0$ but $F_{k+1} = 1$. Clearly, if k is a switching point, then $Q_{k-1} > Q_k$ and $Q_k < Q_{k-1}$. We consider the sequence $Q_{k_1}, Q_{k_2}, \ldots,$ where $k_1 < k_2 < \cdots$ are the switching points. Using simple estimates, we can show $Q_{k_i} \to a$. This implies that $\liminf Q_k = a$, and hence $\liminf \frac{\mathrm{ct}_0(\mathbf{W}[k])}{k} = a$.

A similar argument shows that $\limsup \frac{\mathrm{ct}_0(\mathbf{W}[k])}{k} = b$. □

There are several uniform cubefree homomorphisms from $\{0,1,2\}^*$ to $\{0,1\}^*$ of length 12. One of them,

$$h(0) = 001101101011, \quad h(1) = 001101011011, \quad \text{and} \quad h(2) = 001001010011,$$

yields $L = 10$ and $H = 12$, which shows every $(a,b) > (\frac{5}{12}, \frac{1}{2})$ is the image of some $\mathbf{W} \in CF$. Another,

$$h(0) = 001101011011, \quad h(1) = 001010010011, \quad \text{and} \quad h(2) = 001001010011,$$

yields $L = 12$ and $H = 14$, which shows every $(a,b) > (\frac{1}{2}, \frac{5}{12})$ is achievable. A third homomorphism,

$$h(0) = 001101011011, \quad h(1) = 001001101011, \quad \text{and} \quad h(2) = 001001010011,$$

yields $L = 11$ and $H = 13$, which shows every $(a,b) > (\frac{11}{24}, \frac{11}{24})$ is achievable.

Corollary A.4.3. *The cardinality of CF is 2^ω. In fact, there are 2^ω words in CF, no two of which have a common suffix.*

Proof. The cardinality of fullden(CF) is 2^ω. If two infinite words have a common suffix, then they have the same image under fullden. This implies the result. □

A.5 Growth Rates

As a digression, we show aaa is easily 2-avoidable. Let $T_n = |CF_n|$ be the number of cubefree words on $\{0,1\}$ of length n, and let $\zeta = \inf_{n\to\infty}(T_n)^{\frac{1}{n}}$.

Lemma A.5.1. $\lim(T_n)^{\frac{1}{n}} = \zeta$

Proof. The proof is similar to Lemma A.3.3. □

Theorem A.5.2. *Let $\Sigma = \{1,2,\ldots,2n\}$. Suppose there exists a uniform cubefree homomorphism h from Σ^* to $\{0,1\}^*$ of length k. Then, $\zeta \geq n^{\frac{1}{k-1}}$.*

Proof. Let $W = a_1a_2\ldots a_r$ be a cubefree word over $\{0,1\}$. Let S be the set of words $U = b_1b_2\ldots b_r$ over $\{1,2,\ldots,2n\}$ with $b_i \in \{1,2,\ldots n\}$ if $a_i = 0$ and $b_i \in \{n+1, n+2, \ldots, 2n\}$ if $a_i = 1$. Every word in S is cubefree by Lemma A.4.1. Now, we consider $S' = \{h(U) \mid U \in S\}$. Every word in S' is cubefree and has length rk. Since $|S'| = |S| = n^r$, we have $T_{rk} \geq n^r T_r$. Taking r-th roots, we have

$$((T_{rk})^{\frac{1}{rk}})^k \geq n(T_r)^{\frac{1}{r}}$$

Letting $r \to \infty$, we have $\zeta^k \geq n\zeta$. The result follows. □

There is a uniform cubefree homomorphism from $\{1,2,3,4\}^*$ to $\{0,1\}^*$ of length 12, namely

$$h(0) = 001001101101, \quad h(1) = 001010010011, \quad h(2) = 001010011011, \quad \text{and} \quad h(3) = 001011001101.$$

This implies $\zeta \geq 2^{\frac{1}{11}} > 1.06504$. With some computer analysis, one can show $T_{40} = 9992596$, and so $\zeta \leq 9992596^{\frac{1}{40}}$. Hence $1.06504 \leq \zeta \leq 1.49621$.

References

[1] K. A. Baker, G. F. McNulty, and W. Taylor, *Growth problems for avoidable words*, Theoret. Comput. Sci **69** (1989), 319-345

[2] D. R. Bean, A. Ehrenfeucht, and G. F. McNulty, *Avoidable patterns in strings of symbols*, Pac. J. of Math. **85** (1979), 261-294

[3] J. Berstel, *Axel Thue's work on repetitions in words*, L.I.T.P., Institut Blaise Pascal, Université Pierre et Marie Curie, Paris, France (1992)

[4] J. Berstel and P. Séébold, *A characterization of overlap-free morphisms*, Disc. App. Math. **46** (1993), 275-281

[5] F.-J. Brandenburg, *Uniformly growing k-th power-free homomorphisms*, Theoret. Comput. Sci. **23** (1983), 69-82

[6] J. Brinkhuis, *Non-repetitive sequences on three symbols*, Quart. J. Math. Oxford **34** (1983), 145-149

[7] J. Cassaigne, *Unavoidable binary patterns*, Acta Inf. **30** (1993), 385-395

[8] J. Cassaigne, *Motifs évitables et régularité dans les mots*, Thèse de Doctorat, Université Paris VI, 1994

[9] J. D. Currie, *Open problems in pattern avoidance*, Amer. Math. Monthly **100** (1993), 790-793

[10] J. D. Currie, *On the structure and extendibility of k-power free words*, Eur. J. Comb. **16** (1995), 111-124

[11] J. Karhumäki, *On cube-free ω-words generated by binary morphisms*, Disc. Applied Math. **5** (1983), 279-297

[12] M. Lothaire, *Combinatorics on Words, Encyclopedia of Mathematics and its Applications* vol. 17, Addison-Wesley, Reading, Mass., 1983

[13] A. Thue, *Über unendliche Zeichenreihen*, Norske Vid. Selsk. Skr., I. Mat. Nat. Kl., Christiana **7** (1906), 1-22

[14] A. Thue, *Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen*, Norske Vid. Selsk. Skr., I. Mat. Nat. Kl., Christiana **1** (1912), 1-67

[15] A. I. Zimin, *Blocking sets of terms*, Math. USSR Sbornik **47**, 2 (1984), 353-364. English translation.